IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 15 April 2025, accepted 14 May 2025, date of publication 20 May 2025, date of current version 6 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3571735

SURVEY

Transformers for Vision: A Survey on Innovative Methods for Computer Vision

BALAMURUGAN PALANISAMY[1], VIKAS HASSIJA[2], ARPITA CHATTERJEE[1], ARPITA MANDAL[1], DEBANSHI CHAKRABORTY[1], AMIT PANDEY[3], G. S. S. CHALAPATHI[1], (Senior Member, IEEE), AND DHRUV KUMAR[4]

[1]Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, Pilani Campus, Vidya Vihar, Pilani, Rajasthan 333031, India
[2]School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, Odisha 751024, India
[3]School of CSET, Bennett University, Gautam Buddha Nagar, Greater Noida 201310, India
[4]Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, Pilani Campus, Vidya Vihar, Pilani, Rajasthan 333031, India

Corresponding author: Dhruv Kumar (dhruv.kumar@pilani.bits-pilani.ac.in)

This work was supported by the Birla Institute of Technology and Science, Pilani, India.

ABSTRACT Transformers have emerged as a groundbreaking architecture in the field of computer vision, offering a compelling alternative to traditional convolutional neural networks (CNNs) by enabling the modeling of long-range dependencies and global context through self-attention mechanisms. Originally developed for natural language processing, transformers have now been successfully adapted for a wide range of vision tasks, leading to significant improvements in performance and generalization. This survey provides a comprehensive overview of the fundamental principles of transformer architectures, highlighting the core mechanisms such as self-attention, multi-head attention, and positional encoding that distinguish them from CNNs. We delve into the theoretical adaptations required to apply transformers to visual data, including image tokenization and the integration of positional embeddings. A detailed analysis of key transformer-based vision architectures such as ViT, DeiT, Swin Transformer, PVT, Twins, and CrossViT are presented, alongside their practical applications in image classification, object detection, video understanding, medical imaging, and cross-modal tasks. The paper further compares the performance of vision transformers with CNNs, examining their respective strengths, limitations, and the emergence of hybrid models. Finally, current challenges in deploying ViTs, such as computational cost, data efficiency, and interpretability, and explore recent advancements and future research directions including efficient architectures, self-supervised learning, and multimodal integration are discussed.

INDEX TERMS Action recognition, computer vision, convolutional neural networks (CNNs), efficient transformers, image classification, object detection, segmentation, video processing, vision transformers, visual reasoning.

I. INTRODUCTION

The field of Computer Vision has been fundamentally shaped by Deep Learning, with Convolutional Neural Networks (CNNs) dominating the landscape for nearly a decade. Since AlexNet's breakthrough in 2012, CNNs have become the de facto standard, powering advances in image classification, object detection, and segmentation through architectures like VGG, ResNet, and EfficientNet. The efficacy of these models in capturing hierarchical visual features can be attributed to their inherent inductive biases, specifically translation equivariance, and locality, which confer an ability to learn efficiently from limited data, thereby enhancing their performance in diverse visual recognition tasks.

However, CNNs face inherent limitations. Their local receptive fields struggle with long-range dependencies, and their sequential processing of features creates bottlenecks in modeling global context. These constraints become apparent in tasks requiring holistic understanding, such as scene interpretation or cross-modal reasoning. Meanwhile, the

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

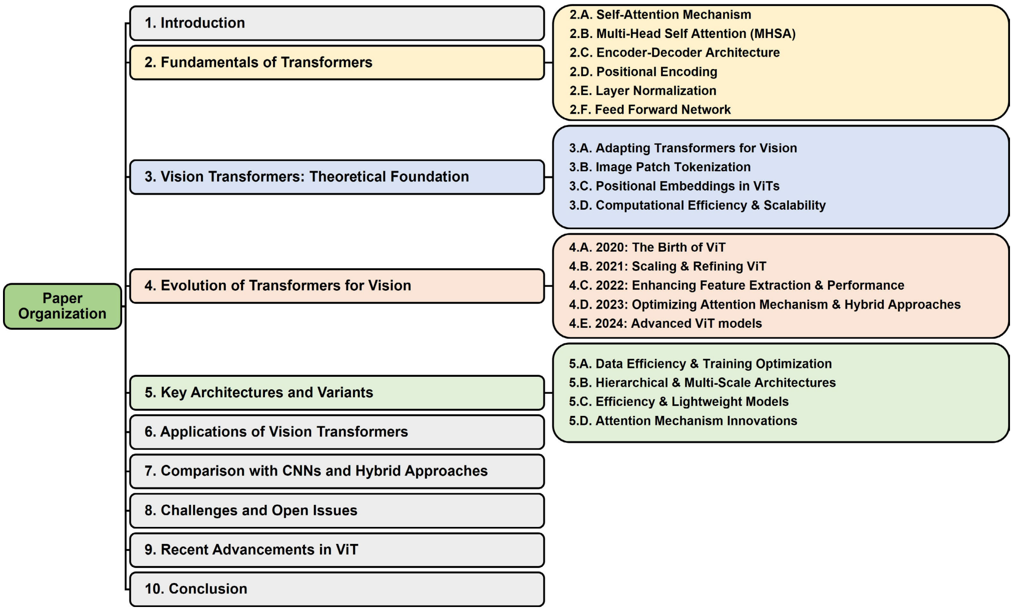
Paper Organization
1. Introduction
2. Fundamentals of Transformers
2.A. Self-Attention Mechanism
2.B. Multi-Head Self Attention (MHSA)
2.C. Encoder-Decoder Architecture
2.D. Positional Encoding
2.E. Layer Normalization
2.F. Feed Forward Network
3. Vision Transformers: Theoretical Foundation
3.A. Adapting Transformers for Vision
3.B. Image Patch Tokenization
3.C. Positional Embeddings in ViTs
3.D. Computational Efficiency & Scalability
4. Evolution of Transformers for Vision
4.A. 2020: The Birth of ViT
4.B. 2021: Scaling & Refining ViT
4.C. 2022: Enhancing Feature Extraction & Performance
4.D. 2023: Optimizing Attention Mechanism & Hybrid Approaches
4.E. 2024: Advanced ViT models
5. Key Architectures and Variants
5.A. Data Efficiency & Training Optimization
5.B. Hierarchical & Multi-Scale Architectures
5.C. Efficiency & Lightweight Models
5.D. Attention Mechanism Innovations
6. Applications of Vision Transformers
7. Comparison with CNNs and Hybrid Approaches
8. Challenges and Open Issues
9. Recent Advancements in ViT
10. Conclusion

FIGURE 1. Organization of the paper.

TABLE 1. Comparison of this survey paper with other literature survey papers.

Work	Model Specs & Evaluations	Evolution of ViT	Key Architectures	ViT Applications	Hybrid Models Overview	Challenges & Open Issues	CNN vs. ViT Comparison	Recent Advancements in ViT
[1]	✓	✗	Limited	✗	✗	✓	✗	✗
[2]	✗	✗	✓	✓	✓	✓	✗	✗
[3]	✗	✗	Limited	✓	✗	✓	✗	✓
[4]	✓	✗	✓	✗	✗	✗	✗	✓
[5]	✗	✓	Limited	✓	✗	✗	✗	✓
This Survey	✓	✓	✓	✓	✓	✓	✓	✓

Transformer architecture revolutionized natural language processing (NLP) through self-attention mechanisms, offering parallel processing and dynamic feature weighting. The success of models like BERT and GPT demonstrated Transformers' ability to capture complex relationships in sequential data.

The emergence of Vision Transformers (ViTs) was precipitated by the innovative application of the Transformer's self-attention mechanism to visual data, where images are discretized into sequences of patches. Notwithstanding earlier attempts to combine CNNs with attention mechanisms, as exemplified by Non-Local Networks, the Vision Transformer (ViT) provided conclusive evidence in 2020 that Transformer-based architectures, unadulterated by convolutional components, could outperform CNNs when scaled to a sufficient extent. Moreover, recent innovations, including incorporating hierarchical attention in Swin Transformers [6] and masked autoencoding in Masked AutoEncoder (MAE) [7], have substantially bridged the gaps in data efficiency and computational performance, ushering in a novel era in visual information processing.

This review aims to provide a comprehensive overview of Vision Transformers and their hybrid variants, comparing them with traditional CNNs. We explore their architectural foundations, key innovations, performance benchmarks, and ongoing challenges. By synthesizing recent research, we aim to highlight the strengths and limitations of ViTs and outline future directions for advancing transformer-based vision models. A comparison of this survey paper with other survey papers available in the literature is summarized in Table 1. The primary contributions of this paper are described as below:

1) In-depth examination of the key architectures and variants of ViTs thereby offering a comprehensive understanding of the evolutionary landscape of ViT models.
2) A comparative analysis of ViTs with traditional CNNs and hybrid approaches, highlighting each paradigm's relative strengths and weaknesses and shedding light on the trade-offs between them.
3) Explores the diverse range of ViT applications, highlighting their potential and versatility in various domains and demonstrating their significant impact on real-world problems.
4) Presents various challenges and open problems inherent to the current ViT architectures, underscoring the existing limitations and identifying key areas that necessitate additional research and development.

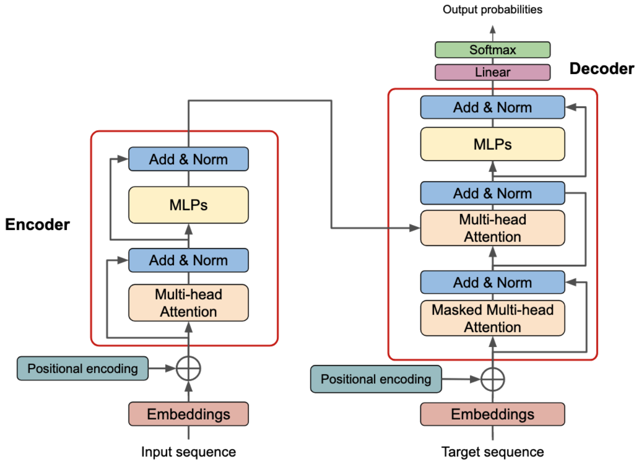

FIGURE 2. Transformer architecture [8].

5) Provides a concise overview of the recent advancements in the field of ViTs, offering a glimpse into the current state-of-the-art and highlighting the prevailing trends.

The organization of the survey is depicted in Figure 1. This survey is organized to comprehensively explore ViTs, beginning with foundational concepts and progressively advancing to cutting-edge developments. Section I is the introduction that establishes the historical context and motivations for transitioning from CNNs to transformer-based vision models, followed by an explanation of core transformer mechanisms in Section II. Section III lays the theoretical groundwork for adapting transformers to visual data, while Section IV traces their evolutionary trajectory in computer vision. A detailed analysis of key architectures and variants is presented in Section V, followed by their diverse applications in Section VI. The paper then compares ViTs and CNNs critically in Section VII, highlighting hybrid approaches and trade-offs. Sections VIII and IX address current challenges and emerging solutions, concluding with future directions and a synthesis of key insights in Section X. This structure facilitates a logical progression from fundamental principles to advanced research frontiers, enabling readers to understand both the theoretical underpinnings and practical implications of ViTs in computer vision.

II. FUNDAMENTALS OF TRANSFORMERS

The fundamentals of Transformers are built upon the Self-Attention Mechanism, which enables the model to weigh the importance of different input elements relative to each other. The Transformer architecture is shown in Figure 2. Additionally, key components include:

1) Multi-Head Self Attention
2) Encoder-Decoder Architecture
3) Positional Encoding
4) Feed Forward Network
5) Layer Normalization

These components work in tandem to facilitate the processing of sequential data, such as text or images, and have been instrumental in achieving state-of-the-art results in various Natural Language Processing (NLP) and Computer Vision tasks. The following subsections provide the necessary fundamental details to understand the transformer architecture:

A. SELF-ATTENTION MECHANISM

The self-attention mechanism is the core computational building block of Transformers, including Vision Transformers (ViTs). It enables a model to weigh and process relationships between different parts of an input sequence, allowing it to capture both local and global dependencies efficiently. Unlike convolutional neural networks (CNNs) that process information using local receptive fields, self-attention [9] computes pairwise interactions between all elements, providing a more flexible and scalable approach to feature extraction.

Given an input sequence (e.g., text tokens or image patches), the self-attention mechanism maps each element in the sequence to a new representation by computing attention scores relative to all other elements. This involves three key steps:

1) **Input Representation:** For an input sequence $X \in \mathbb{R}^{n \times d}$ (where n is the sequence length and d is the feature dimension), three trainable linear projections are applied to obtain the following:
   - *Query (Q):* Represents the current element being processed.
   - *Key (K):* Represents other elements in the sequence for comparison.
   - *Value (V):* Contains the content to be aggregated based on attention.

   These are computed as:

$$Q = XW_Q, K = XW_K, V = XW_V \qquad (1)$$

   where W_Q, W_K and W_V are learnable weight matrices.

2) **Attention Calculation:** The core operation is computing the similarity between the query and key vectors using the scaled dot product. This measures how much attention each element should pay to others:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

   The dot product measures the similarity between the query and keys. The scaling factor $\sqrt{d_k}$, normalizes the dot product to stabilize gradients and prevent exploding attention values. The softmax activation converts the scores into a probability distribution, ensuring attention weights sum to 1. The weighted sum ensures each value vector is aggregated based on its attention weight.

3) **Output Representation:** The attention-weighted sum produces an output sequence with the same length as the input but enriched with global contextual information. This process is repeated in multiple layers to capture hierarchical relationships across the entire input.

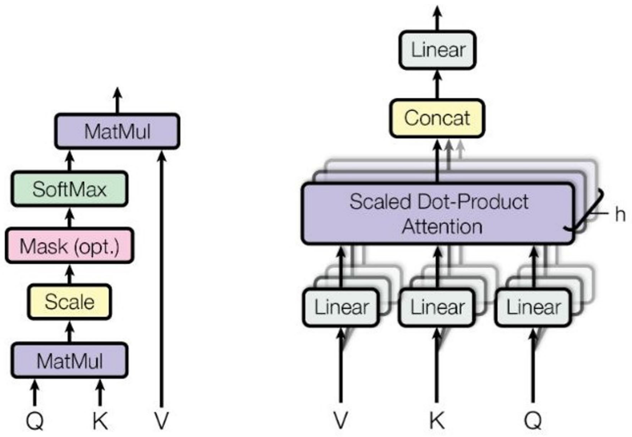

FIGURE 3. Self-Attention mechanism (Left). Multi-Head Attention consists of several attention layers running in parallel (Right) [11].

B. MULTI-HEAD SELF-ATTENTION (MHSA)

To enhance the model's ability to capture different types of relationship, Transformers use multi-head self-attention (MHSA) [10]. Instead of performing attention once, the input is projected into multiple subspaces (or ''heads''), and attention is computed independently in each:

$$MHSA(X) = Concat(head_1, \ldots, head_h)W_O \tag{3}$$

Each *head* corresponds to a different set of learned projections, and h is the number of attention heads. W_O is a final linear projection combining all heads' outputs. MHSA allows the model to focus on diverse aspects of the input simultaneously i.e., such as local texture patterns in one head and global object structures in another. Figure 3 shows the block diagram of the working of self-attention and Multi-Head attention.

C. ENCODER-DECODER ARCHITECTURE

The encoder-decoder architecture is a fundamental design pattern used in many deep learning models, including Transformers. Originally introduced for sequence-to-sequence (Seq2Seq) tasks like machine translation, this architecture has been adapted for vision tasks in models like Vision Transformers (ViTs) and image-to-image translation models [12]. The encoder captures essential features from the input by processing the input (e.g., an image or image patches) to create a compact latent representation. The decoder uses the encoded representation to generate a structured output, such as an image, class label, or sequence. The following is the mathematical representation of Encoder and Decoder functions for the given X:

$$Z = Encoder(X); Y = Decoder(Z) \tag{4}$$

D. POSITIONAL ENCODING

Unlike Convolutional Neural Networks (CNNs), which inherently capture spatial hierarchies through their local receptive fields, Vision Transformers (ViTs) lack an innate understanding of image structure [13]. Since the Transformer architecture was initially designed for sequential data in Natural Language Processing (NLP), where token order is crucial, ViTs must explicitly encode spatial information to preserve the relative positions of image patches. This is achieved through positional encoding (PE), which helps the model distinguish the position of each patch in the input sequence. In the ViT architecture, images are divided into non-overlapping patches and then flattened into 1D sequences of embeddings. While the self-attention mechanism can capture global dependencies, it is position-agnostic which means it does not retain any information about the original spatial arrangement of patches.

E. LAYER NORMALIZATION

Layer Normalization (LayerNorm) is a crucial component of Vision Transformers (ViTs) that stabilizes training and improves model convergence [14]. Originally introduced for sequence-based tasks in natural language processing, LayerNorm is adapted in ViTs to handle high-dimensional patch embeddings. It normalizes activations within a layer, ensuring stable gradients and preventing internal covariate shifts during training. Layer Normalization normalizes the inputs across the feature dimension for each data point independently. Without normalization, the distribution of intermediate representations can shift significantly during training, causing unstable gradients and poor convergence. LayerNorm mitigates these issues by ensuring consistent activation magnitudes. Given an input $x \in \mathbb{R}^d$ (where d is the feature dimension), the normalized output is computed as:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}.\gamma + \beta \tag{5}$$

where,

- x_i represents the input features.
- μ is the mean of the input across the feature dimension computed as $\mu = \frac{1}{d}\sum_{i=1}^{d} x_i$
- σ is the standard deviation computed using $\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2 + \epsilon}$ with a small constant ϵ for numerical stability.
- γ and β are learnable parameters that scale and shift the normalized output.

F. FEED FORWARD NETWORK

The Feed Forward Network (FFN) [15] is a crucial component of the Vision Transformer (ViT) architecture, situated after the multi-head self-attention (MHSA) mechanism in each Transformer block. Its primary function is to apply point-wise transformations to the token representations, allowing the model to capture non-linear dependencies and complex feature interactions. This section elaborates on the structure, function, and significance of FFNs within the ViT framework. Each FFN in a Transformer block consists of two main layers:

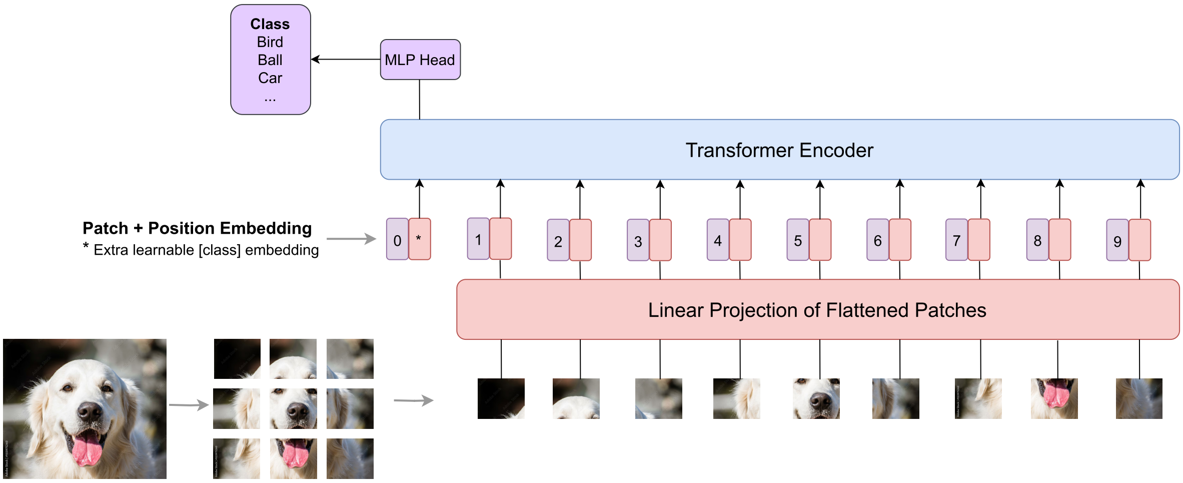
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

FIGURE 4. ViT Model overview. The image split into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard transformer encoder [8].

1) *Linear Projection Layers:* These are fully connected layers that transform the input into a higher-dimensional space and then project it back to the original dimension.
2) *Activation Function:* Introduces non-linearity, allowing the model to capture complex patterns. ViTs typically use the GELU (Gaussian Error Linear Unit) activation for improved performance.

Given an input token representation, $X \in \mathbb{R}^d$, the FFN applies the following operations:

$$FFN(X) = W_2\sigma(W_1X + b_1) + b_2 \tag{6}$$

where,

- W_1 and W_2 are weight matrices.
- b_1 and b_2 are bias vectors.
- $\sigma(.)$ represents a non-linear activation function (commonly GELU).

In each Transformer block, the FFN follows the multi-head self-attention (MHSA) layer and is sandwiched between two Layer Normalization (LN) steps, as shown:

1) *Input Processing:* Token embeddings enter the MHSA module.
2) *Self-Attention:* Captures global relationships between patches.
3) *Add & Norm:* Residual connection and layer normalization.
4) *Feed Forward Network (FFN):* Applies point-wise transformation to each token.
5) *Add & Norm:* Another residual connection and layer normalization.

The overall output of a Transformer block is computed as:

$$\text{Output} = LN(FFN(LN(MHSA(X) + X)) + X) \tag{7}$$

III. VISION TRANSFORMERS: THEORETICAL FOUNDATION

Vision Transformers (ViTs) adapt the Transformer architecture, which was originally designed for sequential data in natural language processing (NLP), to 2D image data, introducing several unique theoretical considerations. A key adaptation involves image tokenization, where an input image is divided into fixed-size patches that are flattened and projected to form a sequence of tokens. This transformation allows images to be processed in a structured format compatible with the Transformer framework. Another essential modification is the use of 2D positional encodings to preserve spatial information, ensuring the model captures the relative position of each patch. These adaptations enable ViTs to model long-range dependencies across an image effectively. Understanding these foundational changes provides insight into how ViTs extend the Transformer's capabilities to visual data. Figure 4 shows the overview of ViT Model. The following subsections briefly explain about various theoretical foundations necessary to understand ViT better:

A. ADAPTING TRANSFORMERS FOR VISION TASKS

The Transformer architecture, originally designed for natural language processing (NLP), operates on one-dimensional sequences of tokens, making it inherently different from the structured nature of images, which exist as two-dimensional grids with complex spatial hierarchies. Adapting Transformers to computer vision introduces several fundamental challenges that must be addressed to achieve effective learning and inference [16].

1) IMAGE-TO-SEQUENCE CONVERSION (PATCH EMBEDDINGS)

Unlike NLP models that process discrete word tokens, images lack a natural tokenization process. Vision Transformers (ViTs) overcome this by partitioning an image into small, non-overlapping patches, which are then flattened and projected into a latent embedding space using a trainable linear layer. This transformation converts the structured 2D image into a 1D sequence of patch embeddings, making it compatible with the Transformer framework [17].

However, this approach discards pixel-level details and imposes constraints on capturing fine-grained local patterns, necessitating additional mechanisms to compensate for lost spatial information.

2) PRESERVING SPATIAL RELATIONSHIPS (POSITIONAL ENCODING)

Transformers lack an intrinsic sense of order, unlike convolutional neural networks (CNNs), which inherently model spatial hierarchies through localized receptive fields. In NLP, positional encodings are used to incorporate word order into the model, but extending this concept to images requires adapting positional embeddings to a 2D domain [18]. ViTs employ 2D positional encodings, which are either learnable or fixed sinusoidal functions, to retain spatial information about patch locations within the image. The effectiveness of these encodings directly impacts the model's ability to understand structural relationships between different regions of the image.

One of the major challenges in scaling Transformers to vision tasks is the computational cost of self-attention. Standard self-attention mechanisms exhibit $O(N^2)$ complexity [19], where N represents the number of tokens. In Vision Transformers (ViTs), since an image is tokenized into numerous patches, self-attention computations become prohibitively expensive, especially for high-resolution images. This quadratic scaling in memory and computation restricts the feasibility of training and deploying ViTs on resource-constrained hardware.

To overcome these challenges, researchers have introduced various efficiency-driven innovations that allow Transformers to achieve state-of-the-art performance in vision tasks. These innovations primarily focus on reducing computational overhead while preserving performance, ensuring that ViTs remain practical for real-world applications. Key advancements include:

1) *Patch-Based Tokenization:* Instead of processing raw pixel values or feature maps extracted from convolutional layers, ViTs divide images into fixed-size patches, treating them as input tokens. This significantly reduces sequence length compared to a per-pixel representation, making computations more feasible. Additionally, hybrid models integrate convolutional feature extractors before applying self-attention, enhancing local feature representation while maintaining the global reasoning capability of Transformers.
2) *2D-Aware Positional Encoding:* Unlike NLP Transformers, which use 1D positional embeddings to capture word order, ViTs require 2D positional encodings to retain spatial relationships between image patches. These encodings can be absolute (fixed embeddings based on patch positions) or relative (dynamically adjusted based on inter-patch relationships). Some models further employ learnable embeddings optimized during training, improving adaptability to different image resolutions.
3) *Hierarchical Attention Mechanisms:* To enhance efficiency and feature extraction, recent advancements such as those seen in Swin Transformers introduce hierarchical attention. Unlike standard ViTs that treat all patches equally, hierarchical models organize patches into progressively coarser levels, resembling the feature pyramids in CNNs. The shifted window attention mechanism, a core technique in Swin Transformers, restricts self-attention computations to local windows while allowing cross-window interactions. This approach significantly reduces computational overhead while preserving long-range dependencies.

By incorporating these innovations, Vision Transformers have demonstrated strong performance across various computer vision tasks, including image classification, object detection, and segmentation. However, ongoing research continues to refine their efficiency and adaptability, paving the way for broader real-world deployment, even in resource-constrained environments.

B. IMAGE PATCH TOKENIZATION

In Vision Transformers (ViTs), image patch tokenization is the process of converting a 2D image into a sequence of token embeddings. This transformation allows images to be processed by the self-attention mechanism of the Transformer model, which traditionally handles 1D sequences [20]. The patch tokenization method is central to how ViTs "see" and interpret visual data. The following are the steps involved:

1) *Splitting the Image into Patches:* Given an input image $X \in \mathbb{R}^{H\times W\times C}$, where where H is the image height, W is the image width, and C is the number of color channels, the image is divided into N non-overlapping square patches of size $P \times P$ (where P is the path size). The number of patches is calculated as:
$$N = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right) \tag{8}$$
Each patch $x_p \in \mathbb{R}^{P\times P\times C}$ captures a local region of the image, providing a manageable unit of information for subsequent processing.
2) *Flattening and Projecting Patches:* Once the image is divided into patches, each patch x_p is flattened into a 1D vector of size $P^2 \times C$, representing the raw pixel information. To prepare these vectors for the Transformer model, they are linearly projected into a latent space of dimension d using a learnable projection matrix:
$$z_p = Linear(x_p), z_p \in \mathbb{R}^d \tag{9}$$
This operation produces a collection of patch embeddings $Z \in \mathbb{R}^{N\times d}$, where each row represents the embedding of a specific patch. These patch embeddings serve as input tokens to the Transformer model, analogous to word embeddings in NLP.
3) *Incorporating the [CLS] Token for Classification:* Inspired by BERT in NLP, ViTs introduce a learnable

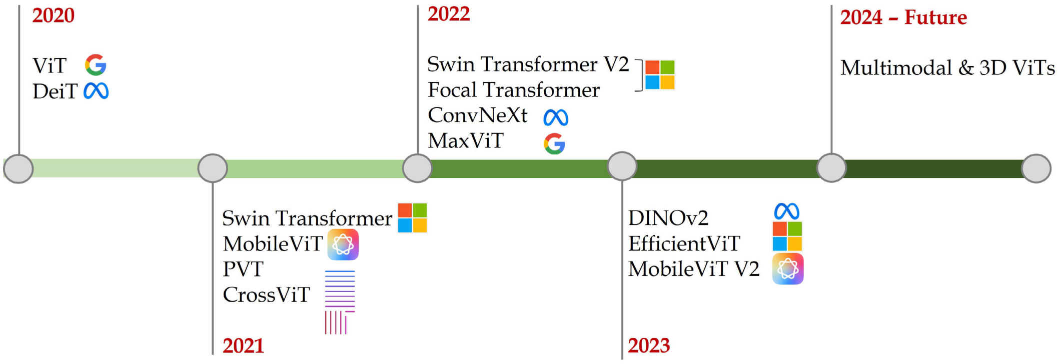

FIGURE 5. Evolution of vision transformers (ViTs): A Timeline of Key Models from 2020 to 2024 and beyond.

class token z_{cls} to the sequence of patch embeddings. This token is prepended to the patch sequence and serves as a global representation of the entire image. During training, the model learns to store information relevant to classification within this token. The final input sequence to the Transformer is:

$$Z_{final} = [z_{cls}; z_1 \ldots z_N] \in \mathbb{R}^{(N+1)\times d} \tag{10}$$

At the output stage, the model uses the representation of z_{cls} for downstream classification tasks.

C. POSITIONAL EMBEDDINGS IN ViTs

Transformers lack an inherent understanding of the spatial structure of their input because the self-attention mechanism is permutation-invariant. While this is suitable for natural language processing (NLP) sequences where token order is explicitly encoded, images possess a 2D grid structure that must be preserved for accurate visual interpretation. Positional embeddings (PEs) address this limitation by encoding spatial information and adding it to the patch embeddings, allowing the model to maintain spatial awareness during training and inference [21].

In ViTs, image patches are tokenized and flattened into a 1D sequence. However, the spatial arrangement of these patches is lost during this process. Without positional embeddings, the model would treat all patches as an unordered collection, failing to capture relationships between nearby and distant regions of the image. By incorporating positional information, ViTs can preserve spatial relationships between image patches, distinguish patch locations, enable the model to learn spatial hierarchies, and enhance performance on vision tasks by understanding object structure and context.

Given a sequence of patch embeddings $Z \in \mathbb{R}^{N\times d}$ (where N is the number of patches and d is the embedding dimension), a positional embedding matrix $P \in \mathbb{R}^{N\times d}$ is added element-wise:

$$Z_{input} = Z + P \tag{11}$$

This enriched representation, combining both patch content and spatial location, is then fed into the Transformer layers. Additionally, the [CLS] token (used for classification) receives a dedicated positional embedding.

D. COMPUTATIONAL EFFICIENCY AND SCALABILITY CONSIDERATIONS

Vision Transformers (ViTs) face significant computational challenges due to the quadratic complexity of the self-attention mechanism, which scales poorly with increasing image resolution. To improve efficiency and scalability, several strategies have been developed. Hierarchical models like Swin Transformers use local attention windows to reduce computational cost, while linear and sparse attention techniques (e.g., Linformer, Performer) approximate attention calculations to achieve linear complexity. Token reduction methods, such as patch merging and dynamic pruning, further decrease the sequence length during processing. Hybrid models combining CNNs and Transformers leverage the efficiency of CNNs for local feature extraction while maintaining the global modeling capabilities of Transformers. Efficient scaling requires balancing patch size, model depth, and training data volume, with lightweight variants like MobileViT optimized for resource-constrained devices. These innovations make ViTs more practical for large-scale vision tasks and real-world applications while retaining their strong performance [22].

IV. EVOLUTION OF TRANSFORMERS FOR VISION

The evolution of Transformers for vision has undergone significant transformations, from the introduction of Vision Transformers (ViTs) to the development of variants like Swin Transformers and MAE. These advancements have enabled efficient and effective image recognition, processing, and generation capabilities. The incorporation of self-attention mechanisms and hierarchical architectures has further enhanced the performance of ViTs [23]. As a result, ViTs have become a cornerstone of computer vision research, driving innovation and progress in the field. Figure 5 shows the milestones in ViT development from the 2020 to 2024. The following subsections provide a brief study of the evolution:

A. 2020: THE BIRTH OF VISION TRANSFORMERS

The introduction of Vision Transformer (ViT) by Dosovitskiy et al. marked a significant milestone in computer vision. ViT treated an image as a sequence of patches and applied the transformer architecture, previously dominant in NLP, to visual tasks. Unlike traditional convolutional neural networks (CNNs), ViT leveraged self-attention mechanisms to capture global dependencies in an image, significantly improving classification performance when trained on large-scale datasets like ImageNet. However, ViT required massive computational resources, making it impractical for smaller datasets and real-world applications [1].

To address this limitation, Data-efficient Image Transformer (DeiT) was introduced by Facebook AI in 2020. DeiT improved ViT's training efficiency by incorporating knowledge distillation, which allowed smaller models to learn from larger teacher models. This breakthrough made ViTs accessible to a broader audience, enabling their deployment in scenarios with limited data.

B. 2021: SCALING AND REFINING VISION TRANSFORMERS

With the success of ViT, researchers focused on enhancing efficiency and adaptability. The Swin Transformer, developed by Microsoft, introduced a hierarchical feature representation and shifted windows, enabling the model to process images at varying scales. This hierarchical approach allowed Swin Transformer to outperform ViT in segmentation and object detection tasks [1], making it a strong competitor against CNNs. Swin Transformer's ability to handle variable image sizes efficiently made it a leading architecture for real-world vision applications.

Another key development was the Pyramid Vision Transformer (PVT), which integrated a pyramid structure into the transformer design, mirroring the multi-scale feature extraction process of CNNs. PVT enabled ViTs to handle dense prediction tasks like object detection and semantic segmentation more effectively, making it more efficient for vision applications beyond classification [24].

Meanwhile, BoTNet (Bottleneck Transformers) combined CNNs with transformers, leveraging self-attention in the later stages of a ResNet-like architecture. This hybrid approach demonstrated that transformers could complement, rather than replace, CNNs in visual tasks, providing both efficiency and superior feature extraction capabilities.

C. 2022: ENHANCING FEATURE EXTRACTION AND PERFORMANCE

As transformers continued to gain traction, researchers sought ways to improve their computational efficiency and feature extraction capabilities. T2TViT (Tokens-to-Token ViT) refined the patch embedding process in ViTs by introducing a progressive tokenization mechanism. This method improved feature representation, allowing for better performance in image classification and recognition tasks [25].

CrossViT, another innovative model, introduced a dual-branch architecture that processed both small- and large-scale image patches simultaneously. By combining information from different resolutions, CrossViT enhanced the model's ability to capture both local and global image features, improving classification accuracy.

Meanwhile, Swin Transformer V2 was introduced as an upgraded version of the Swin Transformer, optimizing its efficiency for large-scale datasets and high-resolution images. Swin V2 incorporated novel normalization techniques and scaling strategies to improve robustness and adaptability across diverse vision tasks.

D. 2023: OPTIMIZING ATTENTION MECHANISMS AND HYBRID APPROACHES

In 2023, researchers focused on improving attention mechanisms and making transformers more computationally efficient. Focal Transformer introduced focal self-attention, a mechanism that prioritizes the most relevant parts of an image while reducing computational overhead. This optimization made it more effective for tasks requiring fine-grained attention, such as object detection and segmentation.

Hybrid architectures also gained momentum, with models like ConvNext blending CNNs and transformers to achieve state-of-the-art performance in image classification and object detection. ConvNext demonstrated that CNN-based inductive biases could still play a crucial role in visual tasks, even in the era of transformers [26].

MaxViT emerged as a powerful transformer model, introducing a multi-axis attention mechanism that captured both local and global dependencies simultaneously. This innovation significantly improved ViTs' ability to process complex scenes and high-resolution images while maintaining efficiency.

E. 2024: ADVANCEMENTS IN EFFICIENCY, EDGE AI, AND SELF-SUPERVISED LEARNING

In 2024, Vision Transformers continued evolving toward efficiency, edge computing, and self-supervised learning. DINOV2 (Meta AI) was introduced as an advanced self-supervised vision model, allowing transformers to achieve remarkable performance in image classification and object detection without relying on large labeled datasets. DINOV2 leveraged self-supervised learning techniques to extract meaningful visual representations, making it highly effective in data-scarce environments [27].

EfficientViT addressed the growing need for computationally lightweight transformers by reducing the number of parameters and optimizing hardware efficiency, making it suitable for real-time applications. EfficientViT was particularly impactful for embedded systems, robotics, and real-time video analysis.

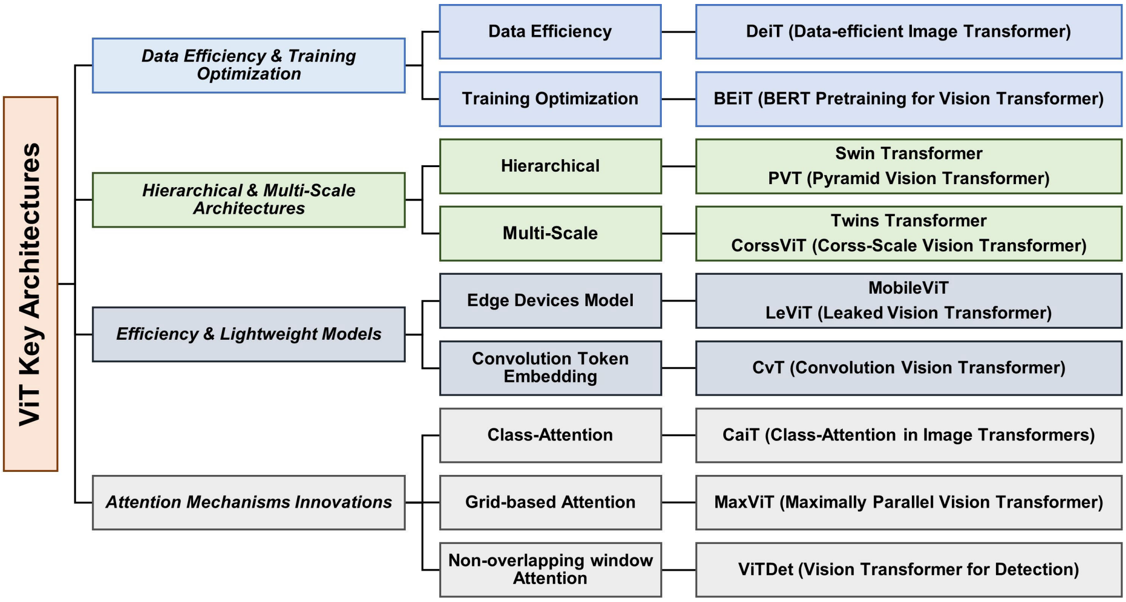
ViT Key Architectures
Data Efficiency & Training Optimization
Hierarchical & Multi-Scale Architectures
Efficiency & Lightweight Models
Attention Mechanisms Innovations
Data Efficiency
Training Optimization
Hierarchical
Multi-Scale
Edge Devices Model
Convolution Token Embedding
Class-Attention
Grid-based Attention
Non-overlapping window Attention
DeiT (Data-efficient Image Transformer)
BEiT (BERT Pretraining for Vision Transformer)
Swin Transformer
PVT (Pyramid Vision Transformer)
Twins Transformer
CorssViT (Corss-Scale Vision Transformer)
MobileViT
LeViT (Leaked Vision Transformer)
CvT (Convolution Vision Transformer)
CaiT (Class-Attention in Image Transformers)
MaxViT (Maximally Parallel Vision Transformer)
ViTDet (Vision Transformer for Detection)

FIGURE 6. ViT key architectures and its variants.

For mobile and edge devices, MobileViT V2 introduced an optimized transformer architecture that maintained high accuracy while operating within low-power constraints. By improving energy efficiency, MobileViT V2 enabled the deployment of transformer models on smartphones, IoT devices, and augmented reality (AR) applications.

As transformers continue to evolve, their adaptability across multiple domains solidifies their position as the future of computer vision and AI-driven perception.

V. KEY ARCHITECTURES AND VARIANTS

Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs for various computer vision tasks. Unlike CNNs, which rely on local feature extraction, ViTs leverage self-attention mechanisms to model global relationships across an image. They take advantage of the transformer architecture, initially designed for natural language processing (NLP), to process image data. Since the introduction of the original ViT, numerous variants have been developed to improve efficiency, scalability, and performance. In the following, we outline the key architectures and their improvements also, Figure 6 provides the key architectures of ViT and its variants.

A. DATA EFFICIENCY AND TRAINING OPTIMIZATION

These architectures follow the standard ViT design but introduce modifications to improve training efficiency [28], scalability, or robustness. Some of the standard ViT variants are briefly described below:

1) DATA-EFFICIENT IMAGE TRANSFORMER (DeiT)

The Data-Efficient Image Transformer (DeiT), developed by Meta AI, is a variant of the Vision Transformer (ViT) designed to address ViT's heavy reliance on large-scale datasets for effective training. Unlike the original ViT, which required massive pretraining on datasets like JFT-300M or ImageNet-21k, DeiT is capable of achieving competitive performance when trained from scratch on ImageNet-1k, a significantly smaller dataset. A key innovation of DeiT [29] is the introduction of a distillation token, which enables the model to benefit from knowledge distillation by learning from a CNN-based teacher model. Figure 7 shows the distallation procedure which includes distallation tocken along with class tocken to reproduce predicted label by the teacher instead of true label. This approach improves training efficiency, generalization, and convergence speed without requiring additional labeled data.

DeiT retains the self-attention-based transformer architecture but incorporates several optimization techniques, including RandAugment, Mixup, CutMix, and stochastic depth, to improve robustness and data efficiency. It comes in different variants, such as DeiT-Ti [32] (Tiny), DeiT-S (Small), and DeiT-B (Base), balancing accuracy and computational cost for various applications. DeiT outperforms traditional CNNs like ResNet-50 in image classification while maintaining a comparable model size and efficiency, making it suitable for edge AI and real-world deployment. By significantly reducing the data requirements for ViTs, DeiT paves the way for wider adoption of transformer-based architectures in computer vision, bridging the gap between CNN efficiency and transformer scalability.

2) BEiT (BERT PRETRAINING FOR VISION TRANSFORMERS)

The BERT Pretraining for Vision Transformers (BEiT) [33]is a self-supervised learning framework that adapts BERT-style masked pretraining for Vision Transformers

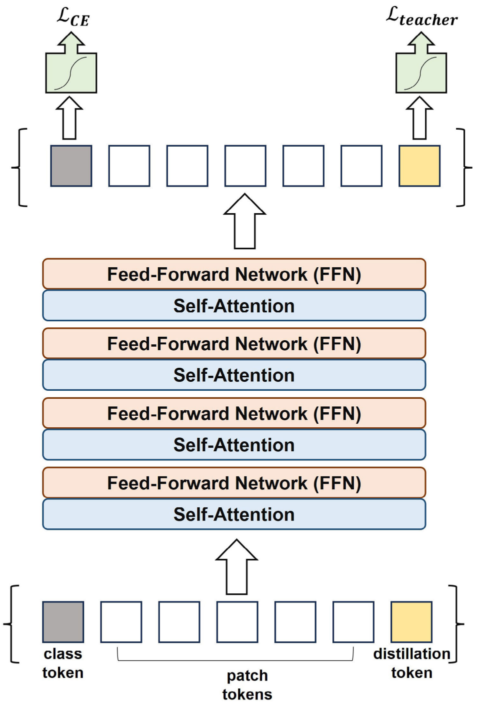

FIGURE 7. Distillation procedure including distillation token. The objective is to reproduce the (hard) label predicted by the teacher, instead of true label (based on [30]).

(ViTs). Introduced by Microsoft Research, BEiT follows the concept of Masked Image Modeling (MIM), where random patches of an input image are masked and the model is trained to predict their corresponding visual representations. Unlike traditional supervised learning, BEiT learns rich, transferable visual representations without requiring labeled data, making it particularly effective for pretraining ViTs on large-scale image datasets.

BEiT consists of two main stages: pretraining and fine-tuning. During pretraining, a ViT model learns to reconstruct masked patches by predicting discrete visual tokens derived from a pre-trained tokenizer (e.g., dVAE from DALL·E) rather than raw pixel values. This forces the model to develop semantic understanding of objects and textures. In the fine-tuning phase, the pretrained BEiT model is adapted for downstream tasks like image classification, object detection, and segmentation. By leveraging self-supervised masked pretraining, BEiT significantly improves the efficiency and generalization of ViTs, making them more data-efficient and competitive with CNNs in vision tasks.

B. HIERARCHICAL AND MULTI-SCALE ARCHITECTURES

Hierarchical and Multi-Scale Architectures in Vision Transformers (ViTs) are designed to capture features at multiple scales, making them particularly effective for tasks like object detection, segmentation, and other dense prediction tasks [34]. These architectures often mimic the pyramidal structure of Convolutional Neural Networks (CNNs), where features are extracted at progressively higher levels of abstraction. Below is a detailed explanation of the key architectures in this category:

1) SWIN TRANSFORMER

The Swin Transformer (Shifted Window Transformer) [31] is a hierarchical Vision Transformer (ViT) designed to improve efficiency, scalability, and performance for computer vision tasks such as image classification, object detection, and segmentation. Introduced by Microsoft Research, it addresses the computational inefficiency of standard ViTs, which use global self-attention and struggle with high-resolution images. Swin Transformer introduces a hierarchical feature extraction mechanism, similar to CNNs, where feature maps are progressively reduced in size across layers, making it highly efficient for dense prediction tasks. Figure 8 shows the architecture of Swin Transformer.

A key innovation of Swin Transformer is the shifted window attention mechanism, where self-attention is computed within non-overlapping local windows to reduce complexity, followed by a shifted window approach that allows information exchange between neighboring windows. This enables Swin Transformer to capture both local and global dependencies while significantly reducing computational cost compared to standard ViTs. With its hierarchical structure and efficient self-attention, Swin Transformer has become a foundational architecture for vision tasks, outperforming CNN-based models like ResNet and enabling state-of-the-art results in image classification, object detection (Faster R-CNN, Cascade R-CNN), and segmentation (Swin-UNet, Mask R-CNN).

2) PYRAMID VISION TRANSFORMER (PVT)

The Pyramid Vision Transformer (PVT) [36] is a hierarchical Vision Transformer designed to improve efficiency and scalability for dense vision tasks such as object detection and segmentation. Unlike the standard Vision Transformer (ViT), which maintains a fixed sequence length throughout its layers, PVT progressively reduces the number of tokens as the network deepens, similar to the feature pyramid structure used in CNNs. This hierarchical tokenization enables multi-scale feature representation, making PVT more suitable for vision tasks that require fine-grained spatial details.

A key innovation of PVT is its spatial-reduction attention (SRA), which reduces the number of tokens in deeper layers while still capturing global information. This significantly lowers memory and computational costs compared to traditional ViTs, making PVT more efficient for high-resolution images. Due to its strong feature representation and scalability, PVT has been widely adopted for tasks such as semantic segmentation (e.g., PVT + UPerNet), object detection (e.g., PVT + RetinaNet), and dense prediction applications, outperforming CNN-based architectures in these domains. Its hierarchical structure allows it to bridge the gap between ViTs and CNNs, offering the benefits of

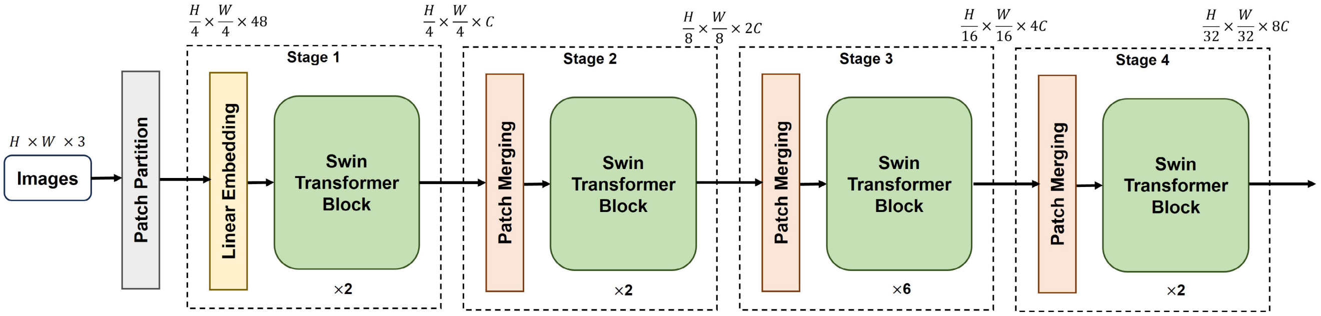

FIGURE 8. The architecture of a Swin Transformer (Swin-T) (based on [31]).

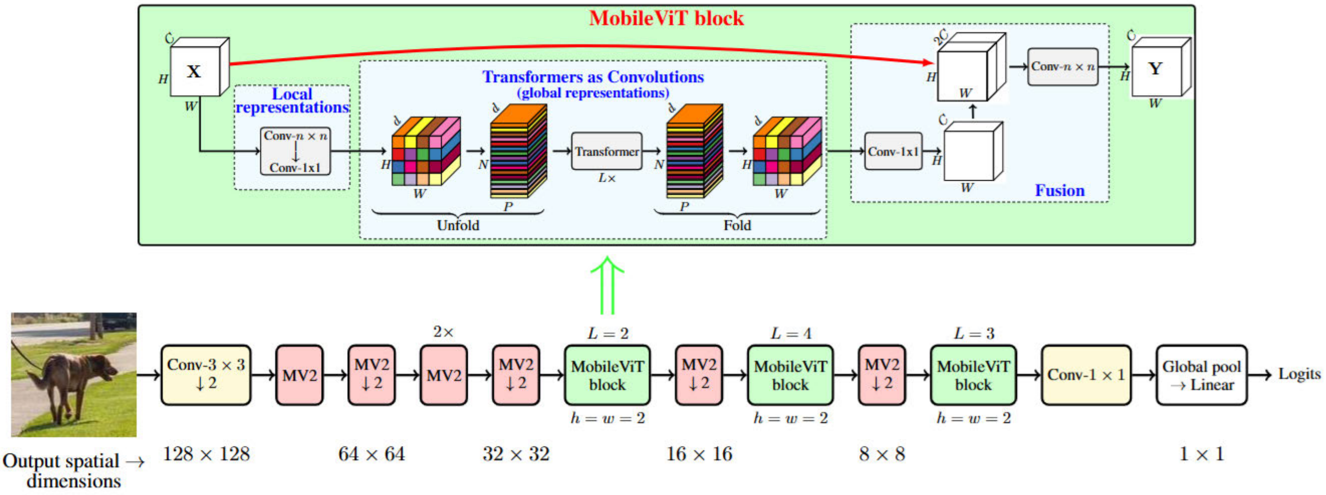

FIGURE 9. Architecture of MobileViT (based on [35]).

global self-attention while maintaining efficiency for real-world applications.

3) TWINS TRANSFORMER

The Twins Transformer [37] is a hierarchical Vision Transformer (ViT) that improves the efficiency and scalability of vision models by integrating both local and global self-attention mechanisms. Unlike standard ViTs, which rely on global self-attention throughout, Twins Transformer introduces a dual-branch architecture that balances efficiency and expressiveness by combining locally-grouped self-attention (LCSA) for fine-grained spatial feature extraction and global sub-sampled attention (GSA) for capturing long-range dependencies.

A key advantage of the Twins Transformer is its hierarchical multi-scale structure, similar to CNNs, which enables better dense prediction tasks such as object detection and segmentation. By limiting self-attention to localized regions while incorporating periodic global attention, Twins achieves lower computational complexity than standard ViTs while maintaining strong feature representation. This makes it particularly effective in vision tasks such as image classification, object detection (Twins + Faster R-CNN), and segmentation (Twins + Mask R-CNN, UPerNet). By striking a balance between local feature extraction and global context understanding, Twins Transformer serves as a highly efficient alternative to traditional ViTs and CNNs for large-scale vision applications.

4) CrossViT (CROSS-SCALE VISION TRANSFORMER)

The Cross-Scale Vision Transformer (CrossViT) is a multi-scale Vision Transformer [38] designed to improve representation learning by processing images at multiple resolutions simultaneously. Unlike standard ViTs, which operate on a single fixed-size patch embedding, CrossViT introduces a dual-branch architecture, where one branch processes larger patches (coarse features) for global context, while the other processes smaller patches (fine details) for local feature extraction. This cross-scale fusion allows the model to capture both fine-grained and high-level semantic information, leading to improved performance on various vision tasks.

A key innovation in CrossViT is the Cross-Attention Module (CAM), which enables efficient information exchange between the coarse and fine branches. By integrating multi-scale feature representations, CrossViT achieves better generalization and robustness compared to standard ViTs, especially for classification, object detection, and segmentation tasks. It has been shown to outperform traditional ViTs and CNNs on image classification benchmarks while maintaining comparable efficiency. CrossViT is particularly useful for tasks where multi-scale information is crucial, such as scene understanding, medical imaging, and fine-grained object recognition.

C. EFFICIENCY AND LIGHTWEIGHT MODELS

Efficiency and Lightweight Models in Vision Transformers (ViTs) focus on reducing computational complexity and memory usage while maintaining competitive performance. These models are designed for resource-constrained environments, such as mobile devices, edge computing, and real-time applications. Below is a detailed explanation of the key architectures in this category:

1) MobileViT

The MobileViT is a lightweight Vision Transformer [35] designed for mobile and edge devices, combining the strengths of convolutions and self-attention to achieve high performance while maintaining efficiency. Figure 9 shows the architecture of MobileViT depicting various building blocks of the architecture. Traditional ViTs require extensive computations, making them unsuitable for real-time applications on resource-constrained devices. To address this, MobileViT introduces a hybrid CNN-Transformer architecture, where convolutional layers are used for early feature extraction and MobileViT blocks replace standard convolutional layers in deeper network stages, enabling global self-attention with minimal computational overhead.

A key innovation of MobileViT is its ability to learn both local and global representations efficiently. It first applies convolutional layers to extract local features, then unfolds feature maps into sequences and processes them using transformer-based self-attention before refolding them back into a spatial representation. This design allows MobileViT to retain the inductive biases of CNNs while leveraging the global context-awareness of transformers. MobileViT achieves state-of-the-art performance on image classification (ImageNet), object detection, and segmentation, all while being lightweight, fast, and memory-efficient making it ideal for applications such as real-time AI, autonomous systems, and mobile vision tasks.

2) LeViT (LEAKED VISION TRANSFORMER)

The LeViT (Leaked Vision Transformer) [39] is a hybrid Vision Transformer (ViT) designed to be fast, efficient, and scalable for mobile and edge computing applications. Unlike standard ViTs, which suffer from high computational costs and memory usage, LeViT introduces a convolution-Transformer hybrid approach, optimizing both speed and accuracy. It combines convolutional layers for early feature extraction with lightweight transformer blocks for efficient global self-attention, resulting in a model that is significantly faster than traditional ViTs while maintaining competitive accuracy.

A key innovation in LeViT is the use of hierarchical patch embeddings and downsampling to progressively reduce the number of tokens, thereby reducing computational complexity. Additionally, it replaces standard Multi-Head Self-Attention (MHSA) with an optimized version, making it more efficient for high-resolution images. LeViT achieves better latency and throughput compared to traditional ViTs, making it suitable for real-time applications such as object detection, facial recognition, and edge AI tasks. By striking a balance between performance, efficiency, and computational cost, LeViT serves as a practical alternative to CNNs and standard transformers in low-power environments.

3) CvT (CONVOLUTIONAL VISION TRANSFORMER)

The Convolutional Vision Transformer (CvT) [15] is a hybrid Vision Transformer that integrates convolutional layers into the standard ViT framework to improve efficiency, feature extraction, and inductive biases. Introduced by researchers from Microsoft, CvT enhances tokenization, local feature extraction, and spatial hierarchies by leveraging convolutions in both the patch embedding stage and self-attention layers. This modification helps address the shortcomings of traditional ViTs, which lack local spatial priors and require large datasets for effective training.

A key innovation in CvT is the Convolutional Token Embedding, which applies overlapping convolutions instead of simple linear projections, improving local feature representation and reducing computational overhead. Additionally, CvT modifies the self-attention mechanism by introducing depthwise convolutions within the projection layers, making the transformer more efficient and structurally similar to CNNs. These improvements enable CvT to outperform standard ViTs and CNNs in image classification, object detection, and segmentation tasks, while maintaining a lower computational footprint. By combining the strengths of CNNs and transformers, CvT offers an effective balance between accuracy, efficiency, and scalability for real-world vision applications.

D. ATTENTION MECHANISM INNOVATIONS

Attention Mechanism Innovations in Vision Transformers (ViTs) focus on improving the efficiency, scalability, and effectiveness of self-attention, which is the core component of transformer architectures. These innovations address challenges such as computational complexity, long-range dependency modeling, and task-specific feature extraction. Below is a detailed explanation of the key architectures in this category:

1) CaiT (CLASS-ATTENTION IN IMAGE TRANSFORMERS)

The Class-Attention in Image Transformers (CaiT) is an improved variant of the Vision Transformer (ViT) introduced to enhance training stability and performance in deep

TABLE 2. Comparison of vision transformer (ViT) architectures based on parameters, throughput, accuracy, and computational complexity.

Model	# Parameters (M)	Throughput (img/s)	Top-1 Accuracy (%)	FLOPs (G)
Data-Efficient Image Transformer (DeiT)				
DeiT-Tiny	5.7	3,796	72.7	1.3
DeiT-Small	22.1	1,827	79.7	4.6
DeiT-Base	86.6	799	82.2	17.6
Swin Transformer				
Swin-T	29	755	81.3	4.5
Swin-S	50	437	83.0	8.7
Swin-B	88	278	83.5	15.4
Pyramid Vision Transformer (PVT)				
PVT-Tiny	13.2	820	75.1	1.9
PVT-Small	24.5	510	79.8	3.8
PVT-Medium	44.2	367	81.2	6.7
PVT-Large	61.4	241	81.7	9.8
Twins Transformer				
Twins-SVT-S	24	1,050	81.7	2.9
Twins-SVT-B	56	608	83.2	8.6
Cross-Scale Vision Transformer (CrossViT)				
CrossViT-S	26.7	1,200	81.0	5.6
CrossViT-B	104.7	320	83.5	21.2
CaiT (Class-Attention in Image Transformers)				
CaiT-XXS24	12	1,400	78.4	2.5
CaiT-XS24	26	760	81.8	5.4
CaiT-S36	68	320	83.3	13.9
MaxViT				
MaxViT-T	31	512	83.6	5.6
MaxViT-S	69	256	85.1	11.7
MaxViT-B	120	128	85.7	23.4
MobileViT				
MobileViT-XXS	1.3	3,500	69.0	0.3
MobileViT-XS	2.3	2,500	74.7	0.6
MobileViT-S	5.6	1,200	78.4	1.3
LeViT (Leaked Vision Transformer)				
LeViT-128	7.8	2,500	76.6	0.4
LeViT-192	11	1,800	78.6	0.6
LeViT-256	18.9	1,200	81.6	1.1
Convolutional Vision Transformer (CvT)				
CvT-13	20	1,200	81.6	4.5
CvT-21	32	800	82.5	7.1
BEiT (BERT Pretraining for Vision Transformers)				
BEiT-Base	86	800	85.2	17.6
BEiT-Large	304	275	87.4	61.6
ViTDet				
ViTDet-B	86	800	82.7	17.6
ViTDet-L	304	275	84.0	61.6

transformer models [40]. Unlike standard ViTs, where the classification token (CLS token) interacts with all layers, CaiT introduces class-attention layers that are placed at the end of the transformer stack. These layers exclusively process the CLS token while keeping the patch tokens frozen, reducing gradient propagation issues and allowing deeper transformer architectures to be trained effectively. This design helps mitigate the training instability of deep transformers and improves representation learning for classification tasks.

CaiT also incorporates LayerScale, a technique that introduces trainable scaling parameters in residual connections, further stabilizing deep model training. Compared to conventional ViTs, CaiT achieves higher accuracy without requiring additional pretraining data and scales efficiently to deeper architectures. By refining token interactions and stabilizing gradient flow, CaiT enables Vision Transformers to achieve competitive results in image classification while maintaining computational efficiency.

2) MaxViT (MAXIMALLY PARALLEL ViT)

The MaxViT (Maximally Parallel Vision Transformer) is a highly efficient Vision Transformer that introduces a grid-based attention mechanism to enhance scalability and computational efficiency for vision tasks [41]. Unlike standard ViTs, which rely on global self-attention and suffer from quadratic complexity, MaxViT incorporates a hybrid local-global attention approach that enables efficient processing of high-resolution images. This design allows MaxViT to balance local feature extraction and long-range dependencies while maintaining a high degree of parallelism for fast and scalable computations.

A key innovation in MaxViT is its Grid Attention mechanism, which combines block attention (local processing

TABLE 3. **Description of advanced ViT models. This table provides a description of some of the advanced ViT models, which are categorized based on the key innovation along with its advantages and challenges.**

Category	Key Models	Description	Advantages	Challenges
Patch-Based Approach	• T2T-ViT • TNT-ViT • DPT • CrowdFormer	Uses different methods for patch generation and processing	Enhances feature extraction, retains spatial relationships	May increase computational complexity
Knowledge Transfer-Based	• TaT • TinyViT	Utilizes teacher-student learning for model training	Reduces data dependency, improves efficiency	Performance depends on quality of teacher network
Shifted Window-Based	• CSWinTT	Uses localized and shifted attention windows	Improves efficiency, reduces computational overhead	May lose some global context
Attention-Based	• CaiT • DAT • SeT	Modifies the self-attention mechanism to enhance feature learning	Increases adaptability and feature representation	Higher memory and processing cost
Multi-Transformer-Based	• Dual-ViT • MMViT • MPViT	Employs multiple transformer branches for different scales	Improves multi-scale feature extraction	Complexity in designing and training models

within small regions) and grid attention (global interaction across spatially distant regions) in a structured manner. This hierarchical structure allows MaxViT to efficiently capture both fine-grained details and global semantic information, making it particularly effective for dense vision tasks such as image classification, object detection, and segmentation. Additionally, MaxViT retains CNN-like efficiency while benefiting from transformer-based self-attention, enabling strong performance across various benchmarks. By integrating parallelized attention operations, MaxViT achieves superior efficiency, making it well-suited for high-resolution vision tasks and real-time AI applications.

3) ViTDet

The ViTDet (Vision Transformer for Detection) [42] is a detection-optimized Vision Transformer designed specifically for object detection and instance segmentation tasks. Unlike traditional ViTs, which are primarily used for image classification, ViTDet is fine-tuned for dense prediction tasks using a feature pyramid structure that allows multi-scale feature extraction. Developed for frameworks like Mask R-CNN and Cascade R-CNN, ViTDet enables high-resolution vision processing while maintaining the benefits of self-attention mechanisms.

A key innovation in ViTDet is its use of non-overlapping window attention, similar to Swin Transformer, to improve computational efficiency while retaining global receptive fields. This allows ViTDet to handle large-scale objects and fine-grained details effectively. It is designed to work seamlessly with modern object detection pipelines, demonstrating state-of-the-art performance on benchmarks such as COCO. ViTDet outperforms CNN-based models like ResNet-FPN [43], making it a powerful choice for high-precision detection, segmentation, and dense vision applications by leveraging hierarchical feature extraction and efficient transformer-based attention.

Table 2 provides the comparison of different ViT architectures discussed above based on parameters throughput, accuracy, and computational complexity. The advanced ViT models which are not explained in the above is summarized in Table 3.

VI. APPLICATIONS OF VISION TRANSFORMERS

With the introduction of a novel framework known as Vision Transformers (ViT), a new approach to tackling computer vision tasks has emerged. Unlike traditional convolutional neural networks (CNNs), which have long dominated the field, ViTs reformulate these tasks as an optimization problem involving equality operations, positioning themselves as a formidable alternative. Table 4 summarizes various ViT Applications along with its architectural features, and some real-World use cases,

A. IMAGE CLASSIFICATION

For classification, trainable attention is divided into two primary streams. Paranoid intricate attention is a necessary precondition for large-scale digital interactions, focusing on specific image regions. Soft attention is used to create deformable feature maps. The term "visual attention" was first introduced to select essential regions and locations in an image [44]. Additionally, resizing the input image can help reduce the computational load of the model. The attention heat map can be used to crop a sub-region from a global image.

The first application of AG-CNN was in medical image classification [45]. SENet [46] introduced soft self-attention to reweight convolutional feature channels, while complex attention requires recalibrating selected feature maps. Saumya et al. [47] leveraged attention maps to reweight DNNs using intermediate feature estimations. Han et al. [48] further enhanced CNN representations by employing attribute-aware attention.

TABLE 4. Vision transformer applications: Task descriptions, architectural features, and real-world use cases.

Application	Description	Features	Use Cases
Image Classification	Vision Transformers (ViT) represent images as patches and apply self-attention to transform the image classification process.	• Self-Attention Mechanisms • Hierarchical Representation Learning • Global Context Modeling	• Disease Diagnosis • Fine-Grained Classification • Product Recognition
Object Detection	Detects objects by finding and identifying them using bounding boxes in images/videos.	• Fine-grained Localization • Multimodal Fusion • Scale and Aspect Ratio Invariance	• Autonomous Driving and Traffic Safety • Medical Imaging and Disease Detection
Semantic Segmentation	Recognizes and labels individual pixels in an image using common semantic properties.	• Pixel-level Classification • Encoder-decoder Architecture • Improved Edge Detection	• Land Coverage Analysis • Scene Understanding
Video Processing	Analyzes and understands video content using transformer-based architectures with self-attention mechanisms.	• Temporal Modeling • Hierarchical Representation • Scalability and Parallelism	• Video Surveillance • Video Editing and Production
Action Recognition	Analyzes video footage to recognize and classify human actiViTies using transformer-based models.	• Temporal Attention Mechanisms • Spatial-Temporal Representation Learning • Long-Term Dependency Modeling	• Surveillance and Security • ActiViTy Monitoring

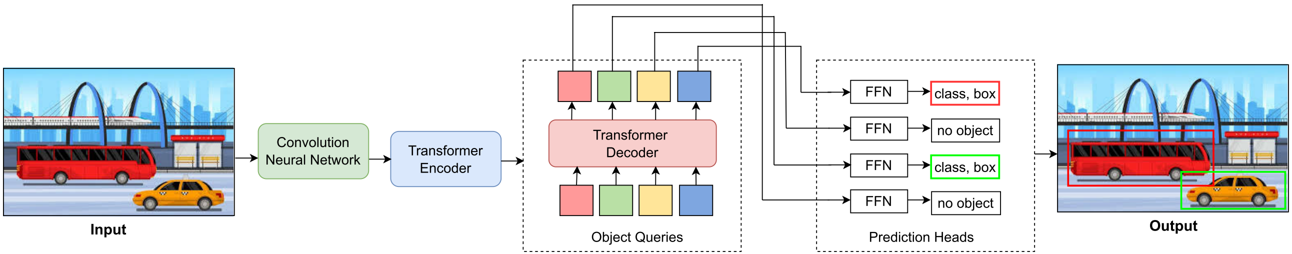

FIGURE 10. The architecture of DETR (based on [49]).

iGPT [50] is a powerful yet non-specialized language model for image classification. Traditional image classification involves training on vast datasets of images and tags. While iGPT learns from image descriptions or captions, it cannot process or analyze image content directly.

Computer vision has long utilized pre-training approaches, as revisited and integrated into self-supervised methods by Chen et al. [50]. The pre-training phase is followed by fine-tuning. The auto-regressive and BERT-based methods are revealed during the pre-training stage. Unlike NLP, where recursion is applied to language tokens, pixel prediction is implemented using a sequence transformer architecture. Pre-training, similar to early stopping methods, benefits from favorable initializations like incremental PCA. A small classification head is added in the final tuning stage to adjust all weights and optimize classification objectives. The following are the key observations and achievements:

- *Understanding 2D Image Characteristics:* Despite being trained on 1D pixel sequences, iGPT understands 2D image features such as object appearance and category.
- *Generative Capabilities:* It generates coherent image completions and samples without human-provided labels.
- *Competitive Features:* State-of-the-art performance is achieved by features extracted from iGPT across various image classification datasets, including CIFAR-10 [51], CIFAR-100 [52], STL-10, and ImageNet.

B. OBJECT DETECTION

Object detection is a fundamental task in computer vision that requires simultaneous localization and classification of potential objects within a single image. It is essential for various applications like facial recognition, autonomous driving, pedestrian detection, and medical detection. It significantly influences scene understanding, environment perception, and object tracking. Deep learning-based object detection methods are becoming more popular due to their rapid development. However, several challenges remain in handling varying scales, creating lightweight models, and balancing efficiency with precision.

Like Faster R-CNN [53], convolutional neural networks (CNNs) have been the foundation of most traditional mainstream object detection techniques. YOLO [54], and SSD [55] are notable examples. The success of transformers in natural language processing has led researchers to try and modify Transformer topologies for computer vision problems.

Transformers have gained significant interest in object detection recently, leading to high-performance models like Deformable DETR, DETR, Swin Transformer, and

DINO. In-depth analysis and evaluation will be necessary for subsequent research as these models represent a new paradigm in object detection.

The Detection Transformer (DETR), introduced by Carion et al. [56], is a transformer-based object detection framework that eliminates the need for hand-crafted components like region proposals. It extracts image features using a CNN backbone, adds fixed positional encodings, and processes them through an encoder-decoder transformer. The decoder produces N output embeddings using learned positional encodings (object queries), where N is predefined rather than dependent on the number of objects. Final predictions bounding boxes and class labels are computed using simple feed-forward networks (FFNs). Unlike traditional detectors, DETR employs bipartite matching to assign predictions to ground truth objects. The architecture of DETER is shown in Figure 10.

Although DETR enables end-to-end object detection, it has drawbacks like long training times and poor small-object performance. Zhu et al. [57] introduced Deformable DETR, which improves efficiency by using a deformable attention module that focuses on key locations instead of all spatial features. This accelerates convergence, reduces computational costs, and integrates multi-scale features, achieving 1.6× faster inference and 10× lower training cost than DETR.

To further reduce computational complexity, Zheng et al. [58] proposed the Adaptive Clustering Transformer (ACT), which replaces DETR's self-attention with locality sensitivity hashing (LSH) for efficient query clustering, minimizing accuracy loss. Multi-Task Knowledge Distillation (MTKD) [59] mitigates performance drops with fine-tuning.

Sun et al. [60] identified cross-attention as the main cause of DETR's slow convergence. They proposed an encoder-only variant with improved training stability, introducing TSP-FCOS and TSP-RCNN models, which enhance performance with feature pyramids.

Dai et al. [61] developed UP-DETR, an unsupervised pretraining approach inspired by NLP, using random query patch detection. This improves DETR's accuracy, especially on small datasets like PASCAL VOC.

MaX-DeepLab is designed explicitly for a sub-category of semantic segmentation known as panoptic segmentation. The object detection using MaX-DeepLab is shown in Figure 11, which uses a dual-path transformer architecture for object detection. The regular semantic segmentation method provides each pixel in an image with a unique class label (e.g., car, person, background), but it doesn't differentiate between individual instances of objects. While panoptic segmentation combines both semantic segmentation and instance segmentation [62]. MaX-DeepLab directly predicts class-labeled masks, eliminating the need for separate steps like object detection and merging. It achieves advanced performance using a "dual-path transformer" that combines CNNs with transformers [62].

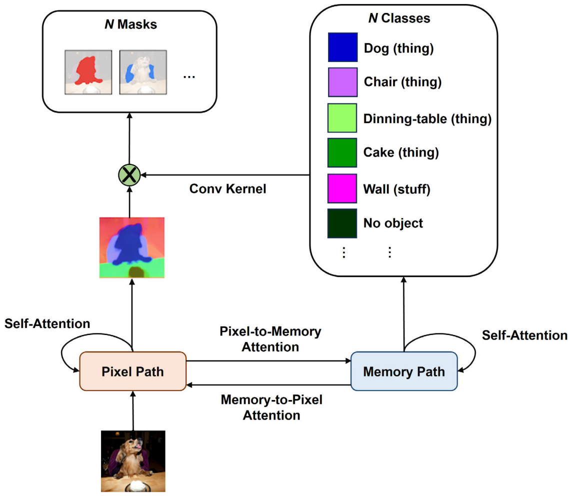

FIGURE 11. Object detection using MaX-DeepLab using dual-path transformer architecture(based on [63]).

C. SEMANTIC SEGMENTATION

Semantic segmentation requires rich spatial information and a large receptive field. Unfortunately, newer techniques typically sacrifice spatial resolution to reach real-time inference speed, resulting in subpar performance. A brand-new network for bilateral segmentation called as BiSeNet was proposed by C.Yu et al. [64]. A Small Stride Spatial Path was built first to maintain the spatial information and produce high-resolution features. A quick downsampling Context Path is used in the interim to obtain an adequate receptive field. A new Feature Fusion Module was introduced to efficiently combine the characteristics on top of the two paths. The suggested architecture strikes the ideal compromise between speed and segmentation performance on the Cityscapes, CamVid, and COCO-Stuff datasets. 68.4% of the mean IOU for the 2048 × 1024 input.

Deformable VisTR is an extension of the VisTR framework, an end-to-end transformer-based approach for VIS. One of the major challenges with VisTR is its training efficiency. It is computationally expensive, requiring around 1000 GPU hours during training. Another important challenge is its slow convergence. To resolve these potential challenges, Deformable VisTR uses the Spatio-Temporal Deformable Attention module [65]. The key idea is that instead of attending to all points, it focuses on a small, fixed number of crucial spatiotemporal sampling locations. The key benefit is that it achieves linear computation in the size of spatiotemporal feature maps while maintaining performance comparable to the original VisTR but with significantly fewer GPU training hours (10 × less).

D. ACTION RECOGNITION

Action recognition involves analyzing human movements from video sequences by capturing both spatial (object presence) and temporal (motion dynamics) information. Vision Transformers (ViT) have shown remarkable performance in this domain by leveraging self-attention to model long-range

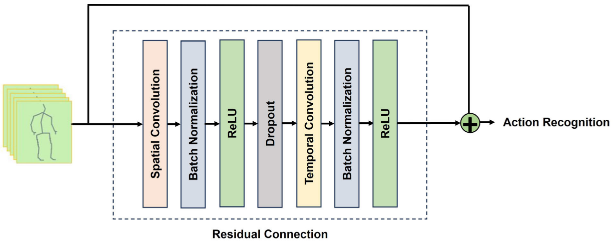

FIGURE 12. ST-GCN block structure (based on [69]).

dependencies, leading to greater robustness and reduced data requirements compared to traditional methods [66].

Despite their success, ViT faces computational challenges when applied to videos. Researchers have explored various strategies to improve efficiency including processing shorter video segments instead of full sequences, using skeleton-based representations (body joint positions) to reduce complexity, and designing specialized ViT architectures tailored for video tasks [67]. Effective action recognition requires capturing both spatial structures (object or joint positions) and temporal patterns (movement dynamics). One approach is skeleton-based modeling, where human movements are represented as graphs, providing a compact and viewpoint-invariant representation [68]. Graph-based methods offer advantages over raw video-based techniques by focusing on motion patterns rather than pixel-based details.

ST-GCN [69] is a skeleton-based action recognition that relies on modeling human movement as a graph, where joints act as nodes and their connections as edges. This approach is more robust to variations in viewpoint, occlusions, and appearance changes compared to traditional RGB-based methods. Earlier techniques used handcrafted features such as SURF [70] and SIFT [71], but these struggled to capture long-range dependencies and required extensive tuning. Recent advances have shifted towards Graph Convolutional Networks (GCNs), which effectively model spatial and temporal relationships in skeletal data. The block structure of ST-GCN is shown in Figure 12.

Sijie Yan et al. [72] pioneered the use of GCNs for action recognition by constructing graphs from human skeletons and applying graph convolutions to extract movement features. Li et al. [73] extended this concept by introducing the Action Structure Graph Convolutional Network (AS-GCN) [74], which incorporated additional structural and actional connections to improve recognition accuracy. However, many GCN-based methods focus only on natural joint connections, overlooking important relationships between non-adjacent joints. ST-GCN enhances traditional GCN models by incorporating extended skeleton graphs with functional connections and a refined partitioning strategy, significantly improving action recognition performance on large-scale datasets [69], [75]. These advancements open new possibilities for more efficient and accurate skeleton-based action recognition, paving the way for further research in graph-based approaches.

While these methods are widely used in video processing, ViT offers a more powerful alternative by directly learning spatial and temporal representations through self-attention, reducing reliance on handcrafted features. Recent research focuses on optimizing ViT for efficient action recognition, making them increasingly viable for large-scale applications.

E. VIDEO PROCESSING

Video processing is another important computer vision task that involves analyzing and manipulating video streams to extract meaningful information. It plays a crucial role in applications that require understanding dynamic scenes, such as action recognition, autonomous driving, and surveillance. ViTs have significantly advanced video processing by modeling complex temporal dynamics and long-range dependencies. Compared to traditional methods, they achieve superior accuracy and quality, leading to remarkable progress in tasks like video completion and translation. ViT facilitates tasks that demand a deep understanding of video content and temporal coherence by effectively capturing spatial and temporal information.

1) VIDEO COMPLETION

ViT excels in video completion by leveraging self-attention to capture long-range dependencies and contextual information, ensuring temporal coherence and realistic results. Their ability to analyze spatial and temporal features makes them well-suited for video inpainting and extrapolation, significantly enhancing video continuity, and visual quality [101].

Existing approaches to video inpainting often struggle with maintaining temporal consistency due to their limited temporal receptive fields. Many methods rely on adjacent frames, leading to artifacts and inconsistencies when handling complex motion. These approaches typically assume global affine transformations, which can result in inaccurate reconstructions [102]. Additionally, without dedicated temporal coherence optimization, frame-by-frame processing requires extensive post-processing, which may fail under severe artifacts. The Spatial-Temporal Transformer Network (STTN) was proposed to overcome these challenges, which formulates video inpainting as a "multi-to-multi" problem. STTN fills missing regions across multiple frames by leveraging both nearby and distant frames. A multi-scale patch-based attention module identifies coherent content across spatial and temporal dimensions, with transformer heads computing similarity across spatial patches at different scales. Stacking multiple layers, STTN dynamically enhances attention mechanisms to refine missing regions. Additionally, a spatial-temporal adversarial loss guides the model in learning to generate cohesive and visually realistic content [103].

2) VIDEO CAPTIONING

ViTs are essential for video translation due to their self-attention mechanism, which captures complex

TABLE 5. Benchmarking ViT performance across tasks: Objectives, Models, datasets, and evaluation criteria.

Task	Objective	Model	Dataset	Evaluation Metrics
Image Classification	Categorize images into predefined classes. Enhance performance on fine-grained classification tasks, transfer learning, and domain adaptation. Improve robustness against adversarial perturbations. [76]	• ViT-B/16 • Swin-T [31] • DeiT-S [77]	• ImageNet • CIFAR-100	• Top-1 Acc: 84.0% • Top-5 Acc: 96.8%
Segmentation	Partition images into meaningful regions. Improve boundary delineation for complex objects, reduce computational overhead, and achieve higher segmentation accuracy in medical and remote sensing applications. [78]	• DeepLabv3+ • SegFormer • Swin-UNet [79]	• Pascal VOC • ADE20K	• Mean IoU: 82.1% • Pixel [81]: 95.5%
Anomaly Detection	Identify rare and unusual patterns. Ensure high sensitivity in detecting subtle anomalies. Improve defect detection in industrial applications and enhance anomaly localization in complex environments. [81]	• GANomaly [82] • PatchCore • FastFlow [83]	• MVTec AD • VisA	• AUC: 98.2%
Image Enhancement	Improve image quality by reducing artifacts, restoring details in low-resolution images, and enhancing contrast in low-light conditions. Useful in satellite imagery and medical imaging. [84]	• SwinIR [85] • Restormer • HAT [86]	• DIV2K • Real-ESRGAN	• PSNR: 32.4 dB • SSIM: 0.91
Image Synthesis	Generate new images from data distribution. Create high-fidelity images with diverse styles, improve texture generation, and enhance realism in virtual environments. [87]	• StyleGAN2 [88] • VQ-VAE-2 • DiT [89]	• FFHQ • CelebA-HQ	• FID: 2.84
Medical Image Analysis	Analyze medical images for diagnosis. Assist in early detection of diseases through automated analysis, improve segmentation of medical scans, and enhance lesion detection. [90]	• Swin UNETR • TransUNet [91] • MedViT [92]	• BraTS • NIH Chest X-ray	• Dice Score: 85.6% • Sensitivity: 88.2%
Image De-noising	Remove noise from corrupted images. Enhance clarity in images captured under poor conditions, restore images affected by motion blur, and improve detail preservation. [93]	• DnCNN • SwinIR [94] • NAFNet [95]	• BSD500 • SIDD	• PSNR: 29.2 dB • SSIM: 0.87
Depth Estimation	Predict depth information from images. Improve 3D scene understanding in autonomous systems, refine depth perception for augmented reality, and enhance robot navigation accuracy. [96]	• DPT • AdaBins [97] • MonoViT [98]	• NYU Depth V2 • KITTI	• RMSE: 0.32 • Abs Rel: 0.12
Image Captioning	Generate descriptive text for images. Support applications in accessibility, content management, and automated image understanding. Improve coherence and accuracy in generated descriptions.	• ViT-GPT2 • BLIP • OFA [99] [100]	• MS COCO • Flickr30k	• BLEU-4: 36.8 • CIDEr: 117.5

transformations and inter-frame interactions. This enables precise frame interpolation, style transfer, and resolution enhancement. By efficiently gathering spatial and temporal information, transformers produce accurate, and superior video translations [104].

Current video captioning methods for untrimmed videos typically involve two phases: proposal localization and caption generation. The Masked Transformer (MT) [105], inspired by machine translation, integrates and optimizes these steps. The MT model uses a proposal encoder to predict event proposals and a caption decoder to generate captions. It replaces recurrent neural networks (RNNs) with stacked self-attention blocks to better capture long-range dependencies.

The Accelerated Masked Transformer (AMT) [106] improves upon MT with enhanced efficiency. It introduces acceleration strategies for both stages. First, faster localization via local attention and a lightweight, anchor-free proposal predictor, and second, single-shot feature masking and average attention for quicker caption generation. In testing, AMT is nearly twice as fast as a 3-layer Masked Transformer, with slightly better performance. AMT demonstrates the power of self-attention mechanisms in video captioning by effectively handling long-range dependencies and optimizing computational efficiency. The introduction of local attention and single-shot feature masking in AMT significantly enhances processing speed without compromising accuracy, making it a strong contender for real-time applications.

Table 5 provides the summary of benchmarking ViT performance across differnt computer vision tasks. The table also have reference to various benchmarked datasets and the top evaluation metrics.

VII. COMPARISON WITH CNNs AND HYBRID APPROACHES

Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) represent two distinct paradigms for visual understanding. While CNNs excel due to their inductive biases, such as spatial locality and translation invariance, ViTs leverage self-attention to capture long-range dependencies and model global relationships. This section compares their performance, strengths, and trade-offs, while also highlighting the emergence of hybrid models that aim to combine the best aspects of both architectures.

A. PERFORMANCE BENCHMARKS ON STANDARD DATASETS

ViTs have demonstrated competitive or superior performance to CNNs on major computer vision benchmarks, particularly

TABLE 6. Image classification performance on ImageNet-1K.

Model Type	Model	Top-1 Acc.	Params (M)	Pretraining Data
CNN	ResNet-50	76.1%	25.6	ImageNet-1K
CNN	EfficientNet-B7	84.3%	66	ImageNet-1K
ViT	ViT-B/16	77.9%	86	ImageNet-21K
ViT	DeiT-B	83.1%	86	ImageNet-1K
Hybrid	ConvNeXt-L	85.5%	198	ImageNet-1K
ViT (SOTA)	EVA-02-L	**89.6%**	305	LAION-2B

TABLE 7. Performance on COCO (Detection) and ADE20K (Segmentation).

Task	Model Type	Model	Metric	Score
Object Detection	CNN	Mask R-CNN (ResNet-50)	mAP	41.0
	ViT	ViTDet (ViT-L)	mAP	53.0
	Hybrid	Swin-T + Mask R-CNN	mAP	50.5
Segmentation	CNN	DeepLabV3+ (ResNet-101)	mIoU	44.1
	ViT	SETR (ViT-L)	mIoU	50.3
	Hybrid	Swin-B + UPerNet	mIoU	53.5

when large-scale datasets are available. Table 6 shows the comparison of ViT and CNN for image classification tasks. The CNNs, such as EfficientNet, have been found to be more parameter-efficient when working with small datasets [107]. In contrast, ViTs and DeiT typically require large-scale pretraining on extensive datasets, such as ImageNet-21K/JFT, to match the performance of CNNs. However, hybrid models that combine the strengths of both architectures, like ConvNeXt and CoAtNet, have shown great promise in bridging this gap, often outperforming both pure ViTs and CNNs [108].

Table 7 provides the comparison for object detection and segmentation tasks. The ViTs have been shown to excel in high-compute scenarios, such as object detection, where models like ViTDet outperform traditional CNN-based detectors when sufficient data is available. However, hybrid models, including Swin and PVT, have demonstrated exceptional performance by leveraging the strengths of both architectures, particularly through their ability to learn multi-scale features, ultimately dominating the landscape in various computer vision tasks.

B. STRENGTHS AND WEAKNESSES: CNNs VS. ViTs

Both Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) are foundational architectures in computer vision, each with distinct advantages and limitations. ViTs, inspired by the Transformer model in natural language processing (NLP), leverage self-attention to process images, while CNNs utilize convolutional layers to extract hierarchical features. Below is an expanded comparison highlighting their strengths and weaknesses across various dimensions.

1) STRENGTHS AND WEAKNESSES OF CNN

CNNs excel in parameter efficiency and local feature extraction, making them ideal for resource-constrained applications like mobile devices and small datasets. Their convolutional operations, honed by decades of research, efficiently capture hierarchical patterns (e.g., edges, textures) and benefit from robust frameworks like PyTorch [109]. However, their reliance on local receptive fields limits global context understanding, and their sequential layer structure reduces parallelism, hindering scalability. While techniques like dilated convolutions help, CNNs often underperform ViTs in tasks requiring long-range dependencies, such as scene understanding or multimodal learning.

2) STRENGTHS AND WEAKNESSES OF ViT

ViTs surpass CNNs in global reasoning and scalability, leveraging self-attention to model distant relationships and achieve state-of-the-art results on large datasets. Yet, their quadratic computation cost and data hunger make them impractical for edge deployment or small-scale tasks where CNNs remain dominant. For applications demanding fine-grained local analysis (e.g., medical imaging) or efficient inference, CNNs retain an edge, while ViTs thrive in high-complexity domains like video analysis or multimodal systems [110]. The choice hinges on trade-offs between computational resources, data availability, and task requirements.

The key strengths and weaknesses discussed above about CNNs and ViTs are summarized in Table 8. While CNNs remain a robust choice for resource-efficient and small-scale applications, ViTs dominate large-scale and complex tasks. Hybrid approaches like Swin Transformer and BoTNet [111]

TABLE 8. Comparison of vision transformers (ViTs) and convolutional neural networks (CNNs) across various aspects.

Aspect	ViTs	CNNs
Global Context	Captures long-range dependencies	Limited by local receptive fields
Efficiency	Computationally expensive (quadratic)	Parameter-efficient (linear complexity)
Data Requirements	Requires large-scale pretraining	Performs well on small datasets
Scalability	Scales well with increasing model size	Limited scalability for large models
Parallelization	Highly parallelizable on modern hardware	Less parallelizable due to convolutions
Performance	SOTA on large-scale datasets	Competitive on small and medium datasets
Interpretability	Complex attention maps	Easier to interpret feature hierarchies

TABLE 9. Popular hybrid architectures combining CNNs and transformers.

Model	Key Innovation	Backbone	Params (M)	ImageNet Acc.
CoAtNet	Depthwise Conv + Relative Attention	CoAtNet-0	25	81.6%
		CoAtNet-7	2,440	**88.6%**
ConvNeXt	Modernized CNN (ViT-inspired)	ConvNeXt-XL	198	87.8%
BoTNet	ResNet + Self-Attention	BoTNet-S1	33	84.7%
MaxViT	Multi-axis Attention	MaxViT-Tiny	31	83.6%
		MaxViT-Base	120	86.5%
Swin Transformer	Shifted Window Attention	Swin-B	88	85.2%
PVT	Pyramid Vision Transformer	PVTv2-B5	82	84.0%

combine the strengths of both paradigms, offering promising solutions across diverse domains.

C. HYBRID MODELS: COMBINING CNNs AND TRANSFORMERS

Hybrid models aim to leverage the strengths of both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to create a more powerful and efficient architecture. By combining the local feature extraction capabilities of CNNs with the global understanding and parallelization capabilities of ViTs, hybrid models can achieve state-of-the-art performance in various computer vision tasks. Hybrid models can learn features at multiple scales, combining the local features extracted by CNNs with the global features learned by ViTs [112]. Hybrid models can create a hierarchical representation of the input data, with early layers focusing on local features and later layers focusing on global features. Hybrid models can benefit from the parallelization capabilities of ViTs, making them more efficient for large-scale computations.

Hybrid models can achieve state-of-the-art performance in various computer vision tasks, surpassing the performance of pure CNNs and ViTs. Some of the advanced and popular hybrid architectures available in the literature are summarized in Table 9.

D. KEY TRADE-OFFS

The choice of architecture depends on the specific application and requirements. For edge devices and small datasets, Convolutional Neural Networks (CNNs) such as MobileNet and EfficientNet are suitable due to their efficiency and parameter frugality. For example, in medical imaging and robotics, where data is often limited and computational resources are constrained, CNNs have been shown to be effective. In contrast, Vision Transformers (ViTs) are better suited for large-scale pretraining tasks, such as those using the JFT or LAION datasets, and for tasks that require a global understanding of the input data, like panoramic segmentation. Hybrid models, which combine the strengths of both CNNs and ViTs, offer a balanced trade-off between efficiency and accuracy, making them a good choice for applications where both are important, such as in the Swin and ConvNeXt models. Additionally, hybrid models have been shown to be effective for detection and segmentation tasks, such as in the PVT and Mask2Former models, where their ability to capture both local and global features is particularly useful. Table 10 summarizes the key trade-off between CNNs, ViTs and hybrid models on various aspects.

VIII. CHALLENGES AND OPEN ISSUES

Vision Transformers (ViTs) have demonstrated remarkable success across various computer vision tasks, but several critical challenges remain unresolved. These limitations present important opportunities for future research and development. Figure 13 provides the summary of various challenges and limitations of ViT in image processing and computer vision, which is briefly explained in the below subsections.

TABLE 10. Key trade-offs between CNNs, ViTs, and hybrid models.

Aspect	CNNs	ViTs	Hybrid Models
Data Efficiency	High (works with small datasets)	Low (needs large datasets)	Moderate
Training Cost	Low (optimized convolutions)	High (quadratic attention)	Moderate
Interpretability	Clear filter semantics	Noisy attention maps	Mixed
Scalability	Plateaus at scale	Excels with big data	Balanced
Global Context	Limited (local receptive field)	Excellent (self-attention)	Good
Local Feature Extraction	Excellent (convolution)	Weak (patch processing)	Good
Hardware Optimization	Mature (cuDNN)	Emerging	Moderate

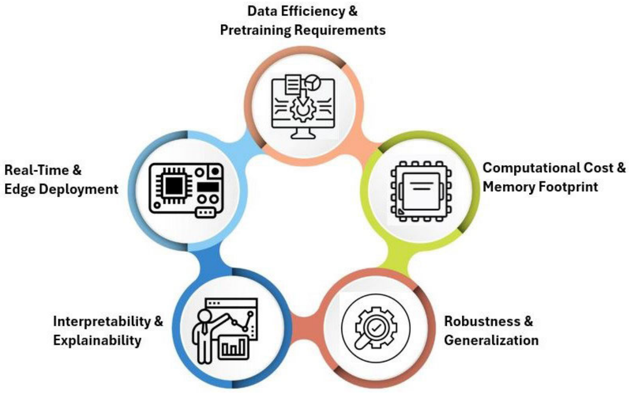

FIGURE 13. ViT challenges and open issues.

A. DATA EFFICIENCY AND PRETRAINING REQUIREMENTS

One of the significant challenges facing Vision Transformers (ViTs) is their requirement for massive datasets to achieve competitive performance. In contrast to Convolutional Neural Networks (CNNs), which can learn effectively from smaller datasets, ViTs typically necessitate large-scale pretraining datasets, such as JFT-300M or LAION [113], to reach state-of-the-art performance. This raises several key issues and open problems that must be addressed to improve the data efficiency of ViTs. The following are the key issues:

- *Weak inductive biases:* Compared to CNNs, ViTs possess weaker inductive biases, including a lack of built-in translation equivariance and locality. This results in reduced performance when trained on smaller datasets, where the model's ability to generalize is crucial.
- *Poor sample efficiency:* ViTs demonstrate poor sample efficiency on medium-sized and small datasets, such as those encountered in medical imaging applications. This limitation hinders the adoption of ViTs in domains where data is scarce or expensive to collect.

The following are some of the open challenges on data efficiency:

- *Self-supervised learning:* Can self-supervised learning (SSL) methods, such as Masked Autoencoders (MAE) or Dense Contrastive Learning (DINO), eliminate the need for supervised pretraining? SSL methods have shown promise in reducing the reliance on large labeled datasets, but their effectiveness in ViTs remains an open question.
- *Designing ViT architectures with stronger priors:* How can ViT architectures be designed to incorporate stronger built-in priors, allowing them to perform well in small data regimes? This may involve introducing additional constraints or biases into the model, such as spatial hierarchies or attention mechanisms, to improve its ability to generalize from limited data.
- *Knowledge distillation:* Are there effective distillation techniques to transfer knowledge from large ViTs to compact versions, enabling the deployment of ViTs in resource-constrained environments? Distillation methods have been successfully applied to CNNs, but their applicability to ViTs remains an open problem.

B. COMPUTATIONAL COST AND MEMORY FOOTPRINT

The Vision Transformer (ViT) architecture has shown remarkable performance in various computer vision tasks, but its computational cost and memory footprint pose significant challenges, particularly when dealing with high-resolution images and video. The quadratic complexity of self-attention, a key component of ViTs, limits their applicability to large-scale inputs. The key issues are the following:

- *Quadratic complexity:* The self-attention mechanism [114] in ViTs has a computational complexity of $O(N^2)$ and memory requirements of $O(N^2)$ for N patches, where N is the number of patches in the input image. This quadratic complexity leads to significant computational costs and memory consumption, making it challenging to apply ViTs to high-resolution images and video.
- *Heavy memory consumption during training:* The large memory requirements of ViTs during training can lead to significant memory consumption, making it difficult to train large models on standard hardware.
- *Inefficient inference:* Compared to optimized Convolutional Neural Networks (CNNs), ViTs can be less efficient during inference, which can limit their deployment in real-time applications.

Some of the open challenges on computational and memory efficiency are listed below:

- *Linear attention variants:* Can linear attention variants, such as linear attention or attention with linear complexity, achieve parity with softmax attention in terms of performance? Linear attention variants can potentially reduce the computational complexity and memory requirements of ViTs, making them more applicable to large-scale inputs.
- *Optimization for hardware accelerators:* How can ViTs be optimized for hardware accelerators, such as Tensor Processing Units (TPUs) or Graphics Processing Units (GPUs), to improve their computational efficiency and reduce memory consumption? Optimizing ViTs for hardware accelerators can enable their deployment in a wide range of applications, from cloud-based services to edge devices.
- *Dynamic token sparsification:* Are there dynamic token sparsification methods that can maintain the accuracy of ViTs while reducing their computational cost and memory requirements? Token sparsification methods, such as pruning or quantization, can potentially reduce the number of tokens processed by the self-attention mechanism, leading to significant computational savings and memory reduction.

C. ROBUSTNESS AND GENERALIZATION

Vision Transformers (ViTs) have demonstrated impressive performance in various computer vision tasks, but their robustness and generalization capabilities are still not well understood. Recent studies have revealed that ViTs can exhibit unexpected failure modes when faced with distribution shifts and adversarial attacks, which raises significant concerns about their reliability and trustworthiness. The following are some of the key challenges:

- *Susceptibility to adversarial patches:* ViTs have been shown to be vulnerable to adversarial patches, which are small, specially crafted patches that can be added to an image to mislead the model [115]. This vulnerability highlights the need for more robust and secure ViT architectures.
- *Poor out-of-distribution generalization:* ViTs often struggle to generalize to new, unseen distributions, which can lead to poor performance in real-world applications. This limitation is particularly concerning in domains where data distribution shifts are common, such as in medical imaging or autonomous driving.
- *Attention collapse in deep architectures:* Deep ViT architectures can suffer from attention collapse, where the attention mechanism becomes less effective or even collapses, leading to poor performance. This phenomenon is not yet fully understood and requires further investigation.

The following are some of the open issues still present in ViT for future exploration and study:

- *Robustness paradox:* Why are ViTs simultaneously robust to some perturbations (e.g., random noise) but vulnerable to others (e.g., adversarial patches)? Understanding this paradox is crucial for developing more robust ViT architectures.
- *ViT-specific regularization techniques:* Can we develop regularization techniques that are specifically designed for ViTs, which can help improve their robustness and generalization capabilities? Regularization techniques, such as dropout or weight decay, are commonly used in CNNs, but their effectiveness in ViTs is still an open question.
- *Stabilizing attention mechanisms:* How can we make attention mechanisms more stable and robust, particularly in deep architectures? Stabilizing attention mechanisms is crucial for improving the overall robustness and generalization capabilities of ViTs.

D. INTERPRETABILITY AND EXPLAINABILITY

While Vision Transformers (ViTs) have achieved remarkable performance across computer vision tasks, their decision-making processes remain significantly less interpretable than conventional CNNs [116]. This "black box" nature poses substantial challenges for deployment in sensitive domains (e.g., medical imaging, autonomous systems) where model transparency is crucial. Some of the key challenges are listed below:

- *Noisy Attention Maps:* Unlike CNN filters that often correspond to semantically meaningful features (e.g., edge detectors), ViT attention maps frequently exhibit diffuse or counterintuitive patterns. Unlike CNN filters that often correspond to semantically meaningful features (e.g., edge detectors), ViT attention maps frequently exhibit diffuse or counterintuitive patterns.
- *Lack of Feature Visualization Analogues:* Patch embeddings are high-dimensional and non-linear. Attention operates on abstract token relationships rather than spatial hierarchies. While CNNs enable visualization through Filter activation patterns and Class activation maps (e.g., Grad-CAM).
- *Cross-Patch Interaction Complexity:* Global self-attention creates intricate dependency graphs where any patch can influence any other, making it difficult to trace decision pathways. The dynamic nature of attention (context-dependent weights) prevents static analysis of feature importance.

The following are the open issues that pave the way for future research directions on ViT interpretability and explainability:

- *Quantifying Interpretability:* Metrics Needed: Formal measures to evaluate whether attention aligns with human-annotated regions of interest and Causal relationships. Standardized datasets with ground-truth explanations for model decisions.

- *ViT-Specific Visualization Tools:* Methods to filter noisy attention (e.g., sparsification, clustering) while preserving salient features. Techniques to map attention patterns to visualize semantic concepts and temporal analysis.
- *Human-Aligned Attention:* Training objectives that encourage attention to focus on semantically meaningful regions (e.g., integrating eye-tracking data). Intermediate layers that enforce alignment with human-defined concepts. Architectures that distinguish correlation from causation in patch relationships.

E. REAL-TIME AND EDGE DEPLOYMENT

While Vision Transformers (ViTs) have demonstrated state-of-the-art performance on many vision tasks [117], their computational demands make deployment in latency-sensitive and resource-constrained environments particularly challenging. Unlike CNNs that benefit from decades of hardware optimization, ViTs face fundamental architectural hurdles for efficient edge deployment [118]. Some of the key issues are as follows:

- *High Inference Latency:* The quadratic complexity of self-attention leads to 2-5× slower inference than optimized CNNs for comparable accuracy. This is particularly problematic for high-resolution inputs (greater than 512px) where the patch count grows rapidly. Hardware-unfriendly operations include Irregular memory access patterns in attention and a lack of optimized kernels for token mixing operations.
- *Memory Constraints:* ViTs require 3-10× more memory than CNNs due to the storage of full attention matrices ($O(N^2)$) and large intermediate activations in feed-forward layers. This adds limitations to developing ViT models in mobile and memory-constrained edge devices.
- *Framework Limitations:* Current deployment stacks (TensorRT, ONNX Runtime) are optimized for CNNs. Missing compiler optimizations for dynamic token sparsification, mixed-precision attention, and hardware-specific acceleration of matrix multiplications.

The following are some of the open challenges that exist in deploying ViT in real-time and edge deployment:

- *Real-Time Video Processing:* Architectural innovations needed for temporal attention compression for video, keyframe-based attention sharing, and recurrent ViT variants with memory buffers. Latency targets should be <30ms/frame for 1080p video (real-time at 30FPS) and <100mW power consumption for embedded vision.
- *Model Compression Strategies:* Attention-specific quantization schemes and 4-bit integer vs floating-point tradeoffs should be considered specifically for ViT deployment. Pruning techniques should be improved for Token pruning (adaptive computation), head pruning for multi-head attention and block-wise sparsity patterns. Need improvement in distillation methods to CNN-to-ViT knowledge transfer.
- *Mobile Deployment Optimization:* Hardware-aligned designs to patch embeddings optimized for mobile NPUs, and Attention approximation for DSP acceleration. The existing or new framework should support TFLite delegates for ViT ops and ONNX extensions for dynamic attention.

IX. RECENT ADVANCEMENTS IN ViT

Recent Vision Transformers (ViTs) advancements have significantly enhanced their performance, efficiency, and applicability across various computer vision tasks. Researchers have developed lightweight and efficient variants like MobileViT and TinyViT, which reduce computational costs while maintaining high accuracy, making them suitable for resource-constrained devices. Techniques such as masked autoencoders (MAE) and self-supervised learning frameworks (e.g., DINO, BEiT) have improved data efficiency, enabling ViTs to generalize well even with limited labeled data. Hierarchical architectures like Swin Transformers and PVT have addressed scalability challenges, allowing ViTs to handle high-resolution images effectively for dense prediction tasks like segmentation and object detection. Additionally, multimodal models such as CLIP and Flamingo have expanded ViTs' capabilities to bridge vision and language understanding, opening new possibilities for cross-modal applications. These advancements, coupled with ongoing research into interpretability, robustness, and hardware optimization, continue to solidify ViTs as a powerful and versatile tool in modern computer vision.

1) **Efficient Vision Transformers:** One of the main challenges with ViTs is their computational cost and memory footprint. Several advancements have focused on reducing computational complexity while maintaining high performance [119]:
   a) *Swin Transformer:* Introduces hierarchical attention with shifted windows, reducing computational cost while preserving spatial locality.
   b) *Pyramid Vision Transformer (PVT):* Uses progressively smaller attention windows, making it efficient for dense prediction tasks.
   c) *Tokens-To-Token (T2T-ViT):* Refines tokenization by aggregating local features before feeding them into the transformer, improving inductive biases.
   d) *LeViT & MobileViT:* Optimized for edge and mobile devices, reducing energy consumption while maintaining competitive accuracy.
   e) *Token Merging (ToMe):* Reduces redundant tokens dynamically to enhance efficiency while preserving accuracy.
2) **Self-Supervised & Contrastive Learning in ViTs:** Self-supervised learning (SSL) has enabled

transformers to learn from unlabeled data, improving robustness and generalization [120]:

   a) *DINO (Self-Distillation with No Labels):* Uses knowledge distillation to train ViTs without labels, producing high-quality, semantic feature representations.
   b) *MAE (Masked Autoencoders):* Adapts masked image modeling (similar to BERT's masked token prediction) for vision tasks, reconstructing missing patches of an image.
   c) *SimMIM & iBOT:* Extend contrastive and masked modeling techniques to enhance self-supervised learning efficiency.

3) **Multimodal Vision Transformers:** ViTs are increasingly used in multimodal tasks [121] by integrating vision with text, speech, or other modalities:
   a) *CLIP (Contrastive Language-Image Pretraining):* Learns vision representations from natural language supervision, enabling zero-shot classification.
   b) *DALL-E and Parti:* Vision-language models that generate images from textual descriptions, leveraging transformer-based architectures.
   c) *BEiT (Bidirectional Encoder Representation from Images):* Inspired by BERT, it learns bidirectional representations using self-supervised objectives.
4) **Vision Transformers in Medical Imaging:** ViTs have shown promise in medical imaging applications, improving diagnosis accuracy in radiology, pathology, and ophthalmology [91]:
   a) *TransUNet:* Combines UNet-like convolutional layers with transformers for medical image segmentation.
   b) *Swin UNETR:* Uses hierarchical attention mechanisms to improve medical image processing.
   c) *ViTs in Pathology:* Applied in histopathology for cancer detection, anomaly localization, and cell segmentation.
5) **Robustness and Generalization of ViTs:** ViTs exhibit improved robustness against adversarial attacks compared to CNNs. However, research focuses on enhancing generalization across different datasets [122]:
   a) *Adversarially Trained ViTs:* Improve robustness by integrating adversarial learning techniques.
   b) *RobustViT & Efficient Fine-Tuning:* Reduce overfitting by integrating domain adaptation strategies.
   c) *Few-Shot Learning with ViTs:* Enhance learning capabilities in data-scarce environments using meta-learning approaches.
6) **Hardware Acceleration for ViTs:** Efforts have been made to optimize ViTs for real-time applications on GPUs, TPUs, and specialized AI accelerators [123]:
   a) *Sparse Transformers:* Reduce computational complexity by processing only essential tokens.
   b) *Quantized ViTs:* Lower bit precision models for energy-efficient inference.
   c) *Neuromorphic ViTs:* Exploring biologically inspired spiking neural networks (SNNs) for ultra-low-power vision tasks.

X. CONCLUSION

Transformers have introduced a paradigm shift in computer vision by replacing traditional, localized feature extractors with globally attentive mechanisms capable of modeling intricate relationships across an image. Through this survey, we have provided a comprehensive exploration of the foundational principles of transformers, their adaptation to vision-specific tasks, and the evolution of architectures such as ViT, DeiT, Swin Transformer, PVT, CrossViT, and others. These models demonstrate that vision transformers can achieve state-of-the-art performance across a wide spectrum of applications including image classification, object detection, segmentation, medical image analysis, video understanding, and cross-modal learning. Moreover, we have compared transformers with CNNs and hybrid models to illustrate their strengths in capturing global context, flexibility in design, and potential for multi-task learning. However, their superior performance often comes at the cost of higher computational demands, memory usage, and reliance on large-scale datasets for effective training factors that limit their widespread adoption in resource-constrained environments.

Despite these limitations, recent advancements have shown promising directions to make Vision Transformers more practical and adaptable. Techniques such as pruning, quantization, knowledge distillation, and the design of efficient transformer variants are helping reduce computational overhead. In parallel, self-supervised and contrastive learning approaches are improving data efficiency and model generalization. Furthermore, the growing interest in multimodal architectures where transformers serve as a unifying backbone across text, vision, and audio indicates their expanding role in broader AI systems. Looking ahead, future research should focus on improving the interpretability of ViTs, enhancing robustness to adversarial inputs, and optimizing their deployment for real-time and edge computing. With these developments, Vision Transformers are well-positioned to become foundational components of intelligent systems across diverse domains, combining accuracy, flexibility, and scalability in ways that go beyond the capabilities of traditional convolutional architectures.

REFERENCES

[1] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[2] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their CNN-transformer based variants," *Artif. Intell. Rev.*, vol. 56, no. S3, pp. 2917–2970, Dec. 2023.

[3] S. Jamil, M. Jalil Piran, and O.-J. Kwon, "A comprehensive survey of transformers for computer vision," *Drones*, vol. 7, no. 5, p. 287, Apr. 2023.

[4] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, ‘‘Vision transformers for image classification: A comparative survey,’’ *Technologies*, vol. 13, no. 1, p. 32, Jan. 2025. [Online]. Available: https://www.mdpi.com/2227-7080/13/1/32

[5] K. Islam, ‘‘Recent advances in vision transformer: A survey and outlook of recent work,’’ 2022, *arXiv:2203.01536*.

[6] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, ‘‘Video Swin transformer,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, ‘‘Masked autoencoders are scalable vision learners,’’ 2021, *arXiv:2111.06377*.

[8] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, ‘‘A survey of visual transformers,’’ *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7478–7498, Jun. 2024.

[9] Y. Li, J. Wang, X. Dai, L. Wang, C.-C. Michael Yeh, Y. Zheng, W. Zhang, and K.-L. Ma, ‘‘How does attention work in vision transformers? A visual analytics attempt,’’ 2023, *arXiv:2303.13731*.

[10] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, ‘‘Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,’’ in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5797–5808. [Online]. Available: https://aclanthology.org/P19-1580/

[11] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, ‘‘A survey on vision transformer,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[12] Y. Gündüç, ‘‘Tensor-to-image: Image-to-image translation with vision transformers,’’ 2021, *arXiv:2110.08037*.

[13] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, ‘‘Conditional positional encodings for vision transformers,’’ 2021, *arXiv:2102.10882*.

[14] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, ‘‘On layer normalization in the transformer architecture,’’ 2020, *arXiv:2002.04745*.

[15] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, ‘‘CvT: Introducing convolutions to vision transformers,’’ 2021, *arXiv:2103.15808*.

[16] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, ‘‘AdaptFormer: Adapting vision transformers for scalable visual recognition,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 16664–16678.

[17] A. Shokouhmand, H. Wen, S. Khan, J. A. Puma, A. Patel, P. Green, F. Ayazi, and N. Ebadi, ‘‘Diagnosis of coexisting valvular heart diseases using image-to-sequence translation of contact microphone recordings,’’ *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2540–2551, Sep. 2023.

[18] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, ‘‘Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[19] F. D. Keles, P. M. Wijewardena, and C. Hegde, ‘‘On the computational complexity of self-attention,’’ in *Proc. 34th Int. Conf. Algorithmic Learn. Theory*, 2023, pp. 597–619.

[20] C. Esteves, M. Suhail, and A. Makadia, ‘‘Spectral image tokenizer,’’ 2024, *arXiv:2412.09607*.

[21] K. Jiang, P. Peng, Y. Lian, and W. Xu, ‘‘The encoding method of position embeddings in vision transformer,’’ *J. Vis. Commun. Image Represent.*, vol. 89, Nov. 2022, Art. no. 103664.

[22] C. J. B. Hernndez, D. A. Sierra, S. Varrette, and D. L. Pacheco, ‘‘Energy efficiency on scalable computing architectures,’’ in *Proc. IEEE 11th Int. Conf. Comput. Inf. Technol.*, Aug. 2011, pp. 635–640.

[23] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, ‘‘Evo-ViT: Slow-fast token evolution for dynamic vision transformer,’’ in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2964–2972.

[24] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, ‘‘Scaling vision transformers,’’ 2021, *arXiv:2106.04560*.

[25] R. Xu, S. Hu, H. Wan, Y. Xie, Y. Cai, and J. Wen, ‘‘A unified deep learning framework for water quality prediction based on time-frequency feature extraction and data feature enhancement,’’ *J. Environ. Manage.*, vol. 351, Feb. 2024, Art. no. 119894.

[26] J. Pang and S. Dong, ‘‘A novel ensemble system for short-term wind speed forecasting based on hybrid decomposition approach and artificial intelligence models optimized by self-attention mechanism,’’ *Energy Convers. Manage.*, vol. 307, May 2024, Art. no. 118343.

[27] Z. Yang, H. Du, D. Niyato, X. Wang, Y. Zhou, L. Feng, F. Zhou, W. Li, and X. Qiu, ‘‘Revolutionizing wireless networks with self-supervised learning: A pathway to intelligent communications,’’ *IEEE Wireless Commun.*, pp. 1–8, 2025, doi: 10.1109/MWC.002.2400197.

[28] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, ‘‘How to train your ViT? Data, augmentation, and regularization in vision transformers,’’ 2021, *arXiv:2106.10270*.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, ‘‘Training data-efficient image transformers & distillation through attention,’’ in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, ‘‘Training data-efficient image transformers & distillation through attention,’’ 2021, *arXiv:2012.12877*.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[32] T. Jumphoo, K. Phapatanaburi, W. Pathonsuwan, P. Anchuen, M. Uthansakul, and P. Uthansakul, ‘‘Exploiting data-efficient image transformer-based transfer learning for valvular heart diseases detection,’’ *IEEE Access*, vol. 12, pp. 15845–15855, 2024.

[33] H. Bao, L. Dong, S. Piao, and F. Wei, ‘‘BEiT: BERT pre-training of image transformers,’’ 2021, *arXiv:2106.08254*.

[34] A. Ashourvan, Q. K. Telesford, T. Verstynen, J. M. Vettel, and D. S. Bassett, ‘‘Multi-scale detection of hierarchical community architecture in structural and functional brain networks,’’ *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0215520.

[35] S. Mehta and M. Rastegari, ‘‘MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer,’’ 2021, *arXiv:2110.02178*.

[36] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, ‘‘Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,’’ 2021, *arXiv:2102.12122*.

[37] J. Li, Y. Bao, W. Liu, P. Ji, L. Wang, and Z. Wang, ‘‘Twins transformer: Cross-attention based two-branch transformer network for rotating bearing fault diagnosis,’’ *Measurement*, vol. 223, Dec. 2023, Art. no. 113687.

[38] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, ‘‘CrossFormer++: A versatile vision transformer hinging on cross-scale attention,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3123–3136, May 2024.

[39] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, ‘‘LeViT: A vision transformer in ConvNet's clothing for faster inference,’’ 2021, *arXiv:2104.01136*.

[40] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, ‘‘Going deeper with image transformers,’’ 2021, *arXiv:2103.17239*.

[41] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, ‘‘MaxViT: Multi-axis vision transformer,’’ 2022, *arXiv:2204.01697*.

[42] L. Wang and A. Tien, ‘‘Aerial image object detection with vision transformer detector (ViTDet),’’ in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 6450–6453.

[43] A. Abdelrahman and S. Viriri, ‘‘FPN-SE-ResNet model for accurate diagnosis of kidney tumors using CT images,’’ *Appl. Sci.*, vol. 13, no. 17, p. 9802, Aug. 2023.

[44] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, ‘‘Stand-alone self-attention in vision models,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.

[45] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, ‘‘Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification,’’ 2018, *arXiv:1801.09927*.

[46] J. Hu, L. Shen, and G. Sun, ‘‘Squeeze-and-excitation networks,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[47] J. Saumya, A. L. Nicholas, N. Lee, and H. T. Philip, ‘‘Learn to pay attention,’’ in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1209–1222.

[48] K. Han, J. Guo, C. Zhang, and M. Zhu, ‘‘Attribute-aware attention model for fine-grained representation learning,’’ in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 2040–2048.

[49] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[50] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 1691–1703.

[51] F. O. Giuste and J. C. Vizcarra, "CIFAR-10 image classification using feature ensembles," 2020, *arXiv:2002.03846*.

[52] Y. Shima, "Image augmentation for object image classification based on combination of pre-trained CNN and SVM," in *Proc. J. Phys., Conf.*, 2018, vol. 1004, no. 1, Art. no. 012001.

[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[54] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[57] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[58] M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," 2020, *arXiv:2011.09315*.

[59] W.-H. Li and H. Bilen, "Knowledge distillation for multi-task learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*. Glasgow, U.K.: Springer, 2020, pp. 163–176.

[60] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3611–3620.

[61] Z. Dai, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised pretraining for object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1601–1610.

[62] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-end panoptic segmentation with mask transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5463–5474.

[63] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9404–9413.

[64] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 325–341.

[65] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3D ConvNets with attention for action recognition," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107037.

[66] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1366–1401, May 2022.

[67] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

[68] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017.

[69] Q. Wang, K. Zhang, and M. A. Asghar, "Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy," *IEEE Access*, vol. 10, pp. 41403–41410, 2022.

[70] E. Oyallon and J. Rabin, "An analysis of the SURF method," *Image Process. Line*, vol. 5, pp. 176–218, Jul. 2015.

[71] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of SIFT and its variants," *Meas. Sci. Rev.*, vol. 13, no. 3, pp. 122–131, Jun. 2013.

[72] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[73] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[74] Z. Yu, D. Jin, Z. Liu, D. He, X. Wang, H. Tong, and J. Han, "AS-GCN: Adaptive semantic architecture of graph convolutional networks for text-rich networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2021, pp. 837–846.

[75] O. Keskes and R. Noumeir, "Vision-based fall detection using ST-GCN," *IEEE Access*, vol. 9, pp. 28224–28236, 2021.

[76] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[77] M. Li, "Transformer-based self-supervised learning and distillation for medical image classification: Improving colorectal cancer detection on NCT-CRC-HE-100K with Swin-T V2," in *Proc. 3rd Int. Conf. Cloud Comput., Big Data Appl. Softw. Eng. (CBASE)*, Oct. 2024, pp. 644–648.

[78] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[79] Q. Tong, Z. Zhu, M. Zhang, K. Cao, and H. Xing, "Cross former embedding DeepLabv3+ for remote sensing images semantic segmentation," *Comput., Mater. Continua*, vol. 79, no. 1, pp. 1353–1375, 2024.

[80] Y. Xu, Y. Xia, Q. Zhao, K. Yang, and Q. Li, "A road crack segmentation method based on transformer and multi-scale feature fusion," *Electronics*, vol. 13, no. 12, p. 2257, Jun. 2024.

[81] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," 2018, *arXiv:1805.06725*.

[82] A. Luiz B. Vieira e Silva, F. Simões, D. Kowerko, T. Schlosser, F. Battisti, and V. Teichrieb, "Attention modules improve modern image-level anomaly detection: A DifferNet case study," 2024, *arXiv:2401.08686*.

[83] F. Wu and S. Xu, "Mask-patchcore: A robust anomaly detection model focusing on interested region," in *Proc. 16th Int. Conf. Graph. Image Process. (ICGIP)*, vol. 13539. Bellingham, WA, USA: SPIE, 2025, pp. 88–99.

[84] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[85] X. Kong, C. Dong, and L. Zhang, "Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy," 2024, *arXiv:2401.03379*.

[86] H. Choi, C. Na, J. Oh, S. Lee, J. Kim, S. Choe, J. Lee, T. Kim, and J. Yang, "Reciprocal attention mixing transformer for lightweight image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 5992–6002.

[87] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[88] A. Bhattad, D. McKee, D. Hoiem, and D. A. Forsyth, "StyleGAN knows normal, depth, albedo, and more," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 73082–73103.

[89] J. Wang, Y. Jiang, Z. Yuan, B. Peng, Z. Wu, and Y.-G. Jiang, "OmniTokenizer: A joint image-video tokenizer for visual generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 28281–28295.

[90] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," 2022, *arXiv:2201.01266*.

[91] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. Lungren, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "3D TransUNet: Advancing medical image segmentation through vision transformers," 2023, *arXiv:2310.07781*.

[92] D. Saadati, O. Nejati Manzari, and S. Mirzakuchaki, "Dilated-UNet: A fast and accurate medical image segmentation approach using a dilated transformer and U-Net architecture," 2023, *arXiv:2304.11450*.

[93] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[94] Z. Wu, W. Shi, L. Xu, Z. Ding, T. Liu, Z. Zhang, and B. Zheng, "DIFNet: Dual-domain information fusion network for image denoising," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Singapore: Springer, 2024, pp. 279–293.

[95] M. Wang, P. Yuan, S. Qiu, W. Jin, L. Li, and X. Wang, "Dual-encoder UNet-based narrowband uncooled infrared imaging denoising network," *Sensors*, vol. 25, no. 5, p. 1476, Feb. 2025.

[96] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021, *arXiv:2103.13413*.

[97] J. Bae, K. Hwang, and S. Im, "A study on the generality of neural network structures for monocular depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2224–2238, Apr. 2024.

[98] Y. Li and X. Wei, "MobileDepth: Monocular depth estimation based on lightweight vision transformer," *Appl. Artif. Intell.*, vol. 38, no. 1, Dec. 2024, Art. no. 2364159.

[99] A. Anagnostopoulou, T. Gouvea, and D. Sonntag, "Enhancing journalism with AI: A study of contextualized image captioning for news articles using LLMs and LMMs," 2024, *arXiv:2408.04331*.

[100] S. Bianco, L. Celona, M. Donzella, and P. Napoletano, "Improving image captioning descriptiveness by ranking and LLM-based fusion," 2023, *arXiv:2306.11593*.

[101] C. Wang, H. Huang, X. Han, and J. Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5232–5239.

[102] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3D gated convolution and temporal PatchGAN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9066–9075.

[103] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial–temporal transformations for video inpainting," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 528–543.

[104] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6489–6498.

[105] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8739–8748.

[106] Z. Yu and N. Han, "Accelerated masked transformer for dense video captioning," *Neurocomputing*, vol. 445, pp. 72–80, Jul. 2021.

[107] A. Sarkar, Y. Yang, and M. Vihinen, "Variation benchmark datasets: Update, criteria, quality and applications," *Database*, vol. 2020, Jan. 2020, Art. no. baz117.

[108] S. Dhar and L. Shamir, "Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks," *Vis. Informat.*, vol. 5, no. 3, pp. 92–101, Sep. 2021.

[109] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.

[110] X. Huang, N. Bi, and J. Tan, "Visual transformer-based models: A survey," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.* Springer, 2022, pp. 295–305.

[111] S. Y. Yerima and M. K. Alzaylaee, "Mobile botnet detection: A deep learning approach using convolutional neural networks," in *Proc. Int. Conf. Cyber Situational Awareness, Data Anal. Assessment (CyberSA)*, Jun. 2020, pp. 1–8.

[112] H. Long, "Hybrid design of CNN and vision transformer: A review," in *Proc. 7th Int. Conf. Comput. Inf. Sci. Artif. Intell.*, Sep. 2024, pp. 121–127.

[113] A. Birhane, V. Prabhu, S. Han, V. Naresh Boddeti, and A. Sasha Luccioni, "Into the LAIONs den: Investigating hate in multimodal datasets," 2023, *arXiv:2311.03449*.

[114] F. Duman Keles, P. Mahesakya Wijewardena, and C. Hegde, "On the computational complexity of self-attention," 2022, *arXiv:2209.04881*.

[115] Y. Fu, S. Zhang, S. Wu, C. Wan, and Y. Celine Lin, "Patch-fool: Are vision transformers always robust against adversarial perturbations?" 2022, *arXiv:2203.08392*.

[116] W. Bousselham, A. Boggust, S. Chaybouti, H. Strobelt, and H. Kuehne, "LeGrad: An explainability method for vision transformers via feature formation sensitivity," 2024, *arXiv:2404.03214*.

[117] X. Liu, Y. Song, X. Li, Y. Sun, H. Lan, Z. Liu, L. Jiang, and J. Li, "Efficient partitioning vision transformer on edge devices for distributed inference," 2024, *arXiv:2410.11650*.

[118] G. Xu, Z. Hao, Y. Luo, H. Hu, J. An, and S. Mao, "DeViT: Decomposing vision transformers for collaborative inference in edge devices," 2023, *arXiv:2309.05015*.

[119] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.

[120] J. Rabarisoa, V. Belissen, F. Chabot, and Q.-C. Pham, "Self-supervised pre-training of vision transformers for dense prediction tasks," 2022, *arXiv:2205.15173*.

[121] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," 2022, *arXiv:2204.08721*.

[122] Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang, "When adversarial training meets vision transformers: Recipes from training to architecture," 2022, *arXiv:2210.07540*.

[123] Z. Li and Q. Gu, "I-ViT: Integer-only quantization for efficient vision transformer inference," 2022, *arXiv:2207.01405*.

BALAMURUGAN PALANISAMY is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India. His research interests include natural language processing, deep learning, and generative models.

VIKAS HASSIJA received the B.Tech. degree from M. D. U. University, Rohtak, India, in 2010, the M.S. degree in telecommunication and software engineering from the Birla Institute of Technology and Science (BITS), Pilani, India, in 2014, and the Ph.D. degree in IoT security and blockchain from the Jaypee Institute of Information Technology (JIIT), Noida. He has done his postdoctoral research with the National University of Singapore, Singapore. He is currently an Associate Professor with KIIT, Bhubaneswar. He has also worked as an Assistant Professor with JIIT for four years. He has eight years of industry experience and has worked with various telecommunication companies, such as Tech Mahindra and Accenture. His research interests include IoT security, network security, blockchain, and distributed computing.

ARPITA CHATTERJEE is currently pursuing the bachelor's degree in computer science and engineering with the Kalinga Institute of Industrial Technology. Also, she is currently pursuing her research internship with the Department of Electricals and Electronics, Birla Institute of Technology and Science (BITS), Pilani, under Dr. GSS Chalapathi. She have a keen interest in algorithms and data structures. In addition to academic pursuits, she actively participates in Hackathons like the Smart India Hackathon (SIH). She is passionate about data science and aims to contribute to the field through innovative research and hands-on experience.

ARPITA MANDAL is currently pursuing the bachelor's degree in computer science and engineering with the Kalinga Institute of Industrial Technology. Also, she is currently pursuing her research internship with the Department of Electricals and Electronics, Birla Institute of Technology and Science (BITS), Pilani, under Dr. GSS Chalapathi. Through coursework and coding practices, she has developed an adequate foundation in programming, algorithms, data structures, etc.. She has contributed to several research and academic initiatives With a focus on Full Stack and Web Development. She is eager to explore the field of CNN and ViT further and apply her skills to real-world challenges in the IT industry.

DEBANSHI CHAKRABORTY is currently pursuing the bachelor's degree in computer science and engineering with the Kalinga Institute of Industrial Technology. Also, she is currently pursuing her research internship with the Department of Electricals and Electronics, Birla Institute of Technology and Science (BITS), Pilani, under Dr. GSS Chalapathi. She is deeply interested in machine learning and is actively researching generative AI using ML tools. She has developed expertise in Full Stack Web Development. Through academic projects and research initiatives, she aims to apply her skills to real-world challenges and further explore advancements in AI and web technologies.

AMIT PANDEY received the M.Tech. degree in computer science from the Deenbandhu Chhotu Ram University of Science and Technology, Sonipat, Haryana, India. He is currently a Research Scholar with the School of Computer Science, Engineering and Technology, Bennett University, Greater Noida, India. His research interests include semantic segmentation for biomedical image segmentation and explainable artificial intelligence.

G. S. S. CHALAPATHI (Senior Member, IEEE) received the B.E. degree (Hons.) in electrical and electronics engineering from the Birla Institute of Technology and Science (BITS), Pilani, in 2009, and the M.E. degree in embedded systems and the Ph.D. degree from BITS Pilani, in 2011 and 2019, respectively. He carried out his postdoctoral research with The University of Melbourne, Australia, under the supervision of Prof. Rajkumar Buyya, and a Distinguished Professor with The University of Melbourne. During his doctoral studies, he was a Visiting Researcher with the National University of Singapore and Johannes Kepler University, Austria. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, BITS Pilani. He has published in reputed journals, such as IEEE Wireless Communication Letters, IEEE Sensors Journal, and Future Generation Computing Systems. His research interests include UAVs, precision agriculture, and embedded systems. He is a member of ACM. He is a Reviewer of IEEE Internet of Things Journal and IEEE Access.

DHRUV KUMAR received the Ph.D. degree in the computer science, focusing on large scale data analytics from the University of Minnesota (UMN), USA. He is currently an Assistant Professor with the Department of Computer Science and Information Systems, BITS Pilani, Rajasthan, India. He has published over 20 papers in internationally recognized journals and conferences. His current research interests include generative AI and AI agents.