Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 18, No. 3, 2025

Vision Transformers (ViTs): A New Era in Computer Vision – A Review

Ao Wang

Economics and Management School of Wuhan University, Wuhan, 430000, China

Abstract: Vision Transformers (ViTs) have become a strong substitute to Convolutional Neural Networks (CNNs) in computer vision, providing a new method to learn global dependencies using self-attention operations. This survey paper provides an in-depth analysis of the development, application, optimization, and deployment difficulties of ViT models. We begin by reviewing the evolution of ViTs from their base architecture, and its subsequent adaptations to newly developed versions, including hybrids with CNNs and multi-scale attention. We then investigate the applications of ViTs such as image classification, object detection, segmentation, depth estimation, medical image analysis, and industry vision inspection. Methods to enhance ViT efficiency—such as model pruning/quantization, hybridization with CNNs, and dynamic adaptation—are extensively discussed. However, ViTs also have significant limitations including computational complexity, scaling and data challenges. Spatial Usage of Scratch Programming Blocks Some potential solutions and future directions are addressed, such as deploying the work on edge device and inclusion in multimodal learning systems. Synthesizing knowledge from recent literature, this paper provides a comprehensive overview of the trends that have developed and six paradigms that currently exist for ViTs in computer vision.

Keywords: Vision Transformer (ViT); Computer Vision; Image Classification; Model Optimization; Multimodal Learning.

1. Introduction

The rise of Vision Transformers (ViTs) represents a significant shift in the computer vision domain powered by the progress in deep learning and the transformative potential of transformer-based architectures. In the past, Convolutional Neural Networks (CNN) have been leaders in the widespread application of computer vision, using their capabilities for strong local feature representations and hierarchical structures. Nevertheless, such models were not sufficient to learn longer range dependencies and capture the global context in images. While ViTs superior performance to CNNs on multiple benchmarks [1][2] is a strong reason for this trend, the fact that ViTs make use of self-attention to model global features from a network of image patches provides a compelling reason for their adoption in the medical-imaging sphere.

The Vision Transformer is first proposed by Dosovitskiy et al. in 2020 [3]. At their core, ViTs are inspired by transformers, proposed for NLP tasks and noted for their capacity to model interdependencies among all tokens in a sequence. The ViT concept is to chunk up an image into a sequence of patches of fixed size, flatten them, and run (the flattened patches) through a stock transformer encoder. This architecture allows ViTs to model long-range spatial dependencies which is harder for CNNs to handle sensibly. To add, the self-attention mechanism applied in transformers allows ViTs to consider relevant regions of the whole image rather than the local ones, and as a result, it has been proven that ViTs can manage complex image structures with global relationships [1][2].

However, the practical deployment of ViTs encountered non-neglectable bottlenecks first. One major drawback was the high computational cost, especially because of the quadratic time complexity in the self-attention mechanism. That made ViTs less compatible to low-power devices and real-time applications [4][5]. Additionally, ViTs have many parameters and need huge number of labeled examples to work effectively, which is may hard for generalization in such a field where annotated data are not available in abundance [1].

In order to break these limitations, the latter years showed a trend towards developing ViTs for efficiency and scalability. Strategies to mitigate the computational cost by applying techniques such as knowledge distillation, pruning, quantization, low-bit quantization (such as Posit arithmetic), and hybridization with CNNs have been used while retaining the accuracy at a more controlled level [6][7]. Hybrid approaches such as Turbo ViT and Swin Transformer tries to integrate the local modeling strengths of a CNN with the global reasoning capabilities of a ViT.

These architectural and algorithmic enhancements have allowed ViTs to progress from being academic baselines to practical use cases. Today we are seeing ViTs being utilized in a variety of computer vision tasks such as: image classification, object detection, semantic segmentation, depth estimation and so on. Especially in the application domain of medical image analysis as well as industrial inspection, ViTs have been shown to be particularly successful, more powerful than traditional CNNs in tasks such as fundus image classification for retinal disease [1][8]; and defect detection in manufacturing lines.

With the development of the field, ViTs are starting to be applied for edge computing and embedded systems using lightweight architectures and dynamic processing approaches. With the multimodal systems -- combining visual and textual information -- these are more general as components in next-generation AI solutions, such as self-driving cars, health diagnostics and robotics [6][7].

In this review article, we provide a comprehensive summary on the state of the art on the progress, applications, optimizations, challenges, and perspectives of Vision Transformer models in computer vision. Drawing on insights from pas research, we analyze how ViTs have evolved, their comparative advantages over CNNs, and how they are being adapted to meet the growing demands of modern AI applications.

2. Evolution of ViTs: The Journey from CNNs to Vision Transformers

Vision Transformers (ViTs), proposed by Dosovitskiy et al. in 2020 [3], was a game changer in the world of computer vision. Until that moment, Convolutional Neural Networks (CNNs) were the norm as they are powerful in modeling local hierarchical information. However, CNNs encountered difficulty in modeling long-range spatial dependencies and global context in images. ViTs bridged this gap by utilizing self-attention mechanisms that were developed for processing natural language, so it could be used on visual input.

2.1. The Birth of Vision Transformers

Original Vision Transformer (ViT) model introduced by Dosovitskiy et al. (2020) [3] profoundly changed the way in which deep learning models processed images. So instead of running convolutional layers over images to capture local features, ViT split an image up into non-overlapping patches, flattened them and fed them into a model as a sequence of tokens, just like words in an NLP model such as BERT. Subsequently, these tokens were fed to the transformer encoder, which employed self-attention to model the dependency between patches that encode not only the local information, but could also capture, to some extent, global context within the image [5][9].

This enabled ViTs to address the shortcoming of CNNs in reasoning over relationships between distant parts of an image, and to be particularly well suited for tasks with global context being important, such as image classification and object detection. In their original work, the authors showed that ViTs could match or even outperform classical CNN architectures such as ResNet and VGG on large-scale image recognition tasks such as the ImageNet classification benchmark [10]. Nevertheless, even with these achievements, ViTs were inн computationally expensive as a result of the self-attention operation. The operation complexity of the processing of an image increased by the square against the number of the patches, which made the ViTs difficult to apply to the resource-limited devices like a mobile phone and an embedded system [4][5].

2.2. The Rise of Efficient Vision Transformers

To overcome the challenges of these computationally-intensive ViTs, researchers have already started investigating different approaches to make them efficient without compromising performance. Among the first kind of methods were those that made changes to the structure of the ViT itself. As an example, the Data-efficient Image Transformer (DeiT) from Touvron et al. (2021) [10] were able to reduce the amount of data required for training ViTs by using methods like knowledge distillation. In this configuration, a smaller "student" model learned to behave as a larger "teacher" model, allowing the student to perform well on limited data [2].

Two hybrid architectures were introduced which were designed to gain benefits of both CNNs and ViTs. Hybrid models e.g., Swin Transformer [5] included visualizing the local region using convolutional layers in the beginning, then represented the entire input using the transformer layers. This paradigm made full use of the advantages of CNN in local features extraction and ViT in modeling global context. Such hybrid approaches have shown better performance in various tasks such as semantic segmentation and object detection in which both local and global information is important [4] [7].

Efficiency search also inspired other techniques such as pruning, quantization and low-bit encoding. Techniques like pruning aim at reducing the size of the model by cropping unimportant parameters, while quantization drops the precision of the model's weights and activations in order to speed up the inference while still retaining a small drop in the overall performance. Low-bit encoding (e.g., with Posit Arithmetic [4] has been shown to dramatically reduce the computing requirements of ViTs and low-hardware-resources devices can utilize this to deploy ViTs [2][7].

2.3. The Evolution of ViT Variants and Architectures

Multiple versions of ViT have since been proposed to address various problems and to beat other architectures on specific tasks. An example of such a model is the Shifted Window Transformer (Swin) by Liu et al. (2021) [5] that proposes a family of hierarchical attention mechanisms to reduce the computational complexity of self-attention by limiting it in smaller windows. This offers not only an efficient embedding of molecular features, but also enables to capture both local and global features for the model to be used for, e.g., object detection and image segmentation [2][4].

Recent works in the ViT domain also involve Tokens-to-Tokens Vision Transformer (T2T-ViT) [11] to improve tokenization procedure for meticulous image-representation, LeViT [12] for combination of the CNN and ViT power to form a compute-efficient model for the real time applications [6][7].

2.4. ViTs in Specialized Domains

In addition to the conventional computer vision problems, ViTs have been applied in more specific fields, such as medical imaging and industrial visual inspection. In medical imaging, ViTs have demonstrated robust results in clinical tasks such as the classification of fundus images for retinal diseases diagnosis [7], evidencing the potential of ViTs to model subtle patterns within medical images, a purpose that was difficult to achieve using CNN-based methods. Also, regarding industrial visual inspection, ViTs have been used to defect detection and quality control, performing better than CNNs in difficult situations when the amount of data is insufficient [1].

2.5. Future Directions

Vision Transformers are still far from mature. New opportunities for further enhancement of them focusing on more efficiently, scalability, and applicability to real-time systems are being explored by researchers. Hybrid models that integrate CNNs and ViTs are also an attractive direction, since they strike a balance between local feature extraction and global context modeling. Moreover, improvement of quantization and pruning are anticipated to increasingly fine-tune ViTs for edge deployment, which will reduce the entry threshold for various applications [4][6].

In summary Arising from this work, the emergence of ViTs is a paradigm and game changer for the computer vision community, providing new modalities for modeling long-articulated global context in images. And although issues concerning computational efficiency persist, the promising advancements are starting to make ViTs more feasible for applications and have potential to be applied across diverse domains including healthcare, autonomous driving, or manufacturing [9][10].

3. Applications of Vision Transformers (ViTs) in Computer Vision

Vision Transformers (ViTs) have attracted considerable interest in the computer vision community because of their capacity to capture fine-grained dependencies in images by utilizing the transformer architecture created for natural language processing. The new direction for using ViTs to solve different computer vision tasks (such as image classification, object detection, segmentation, 3D visions, industrial applications and medical images) has been opened. This section discusses how ViTs have been used successfully for these tasks, and contrasts them with CNN-based approaches.

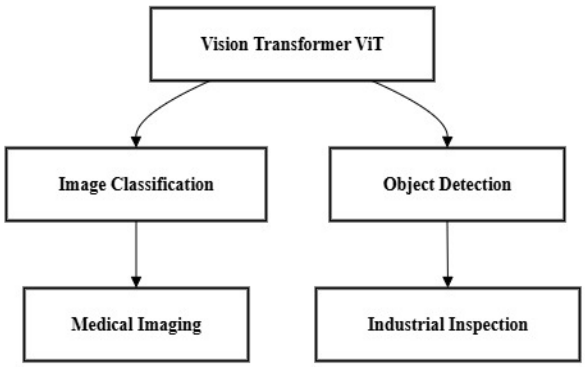

Fig.1 Applications of ViTs in Computer Vision

Figure 1 shows a variety of computer vision tasks in different categories, to which ViTs have been successfully linked. From baseline tasks like image classification and object detection to more specialized applications in fields like medical imaging and industrial inspection, the diagram gives an overview of ViTs' increasing impact across various domains

3.1. Image Classification

Image classification is still one of the most significant and successful applications of ViTs. Considered related work, the concept of deep learning in computer vision datasets positioned CNNs as deep learning models. Traditional work is that deep learning model would be used as traditional architecture standardization. However, modeling long-range dependencies is one of the weaknesses of CNNs and they are not effective when the task demands a holistic analysis of the image content. This is what ViTs aim to tackle, by first treating an image as a sequence of patches and by then using self-attention to model global relationships between all patches in the sequence. This global attention mechanism allows ViTs to model long-range dependencies, which is better performance particularly when global context is important.

ViTs have achieved impressive performance on benchmark datasets like ImageNet, showing that the models are effective for image classification. For instance, Dosovitskiy et al. (2020) [3] demonstrated that ViTs were competitive with the state of the art canned CNN models (like ResNet and VGG) on large scale classification. However, ViTs are computation hungry and introduce deployment issues in resource-starved platforms. Efficient variants have been developed that address these challenges. The DeiT (Data-efficient Image Transformer) proposed in Touvron et al. (2021) [10], utilizes knowledge distillation-based learning, in which a smaller student model learns from a larger teacher model, in an attempt to be less wasteful of data while not losing in performance. Also, the Swin Transformer [5] adopts a two-level attention mechanism that restrains self-attention to local windows and keeps windows connected across scales. This is way in which computational efficiency is achieved without sacrificing the capability to adequately model presence and absence of both local and global features.

Table 1 shows a comparative summary of ViTs with traditional CNNs in the aspect of image classification, comparing the performance, computation cost, data efficiency, as well as local and global feature representation. This comparison highlights the increasing usage of ViT in large-scale classification tasks, but also demonstrates the persistent importance of CNN in this task when efficiency and easy deployment is concerned.

Table 1. Comparison of ViTs with CNNs

Attribute	ViTs	CNNs
Performance (Accuracy)	High, especially for large datasets (e.g., ImageNet)	Good, often slightly lower than ViTs in complex tasks
Computational Cost	High (quadratic complexity in self-attention)	Lower (local receptive field and hierarchical structure)
Memory Requirements	High (larger models and parameters)	Lower (smaller number of parameters)
Global Context	Excellent (captures long-range dependencies)	Limited (focus on local patterns)
Application Flexibility	Highly flexible in tasks requiring global context	Effective for tasks with local dependencies
Data Requirements	High (needs large datasets to perform well)	Moderate (works well with smaller datasets)

In summary, Vision Transformers are a strong candidate for image classification, outperforming classical CNN approaches in terms of accuracy on large-scale data, while recent works such as DeiT and Swin Transformer have optimized them in a way that makes them performant enough for applications.

3.2. Object Detection and Segmentation

Apart from the image classification task, ViTs have also demonstrated excellent performance on object detection and semantic segmentation tasks, which are object-level recognition tasks and involve not just recognizing objects but also localizing and delimiting objects inside an image. ViTs have the advantage in these tasks, as they can capture global dependencies and long-range communications.

DETR Detection Transformer is one of the pioneering

works that employs ViTs to object detection using transformers directly, removing the need for hand-crafted region proposals. DETR obviates the requirement of designing handcrafted RPNs in conventional CNN-based detectors by formulating object detection as direct set prediction. The trained model adopts the transformer encoder-decoder architecture to generate a fixed set of objects, which employs global self-attention mechanism to model interactions between all image regions. This allows DETR to better handle challenging scenes (such as scenes with small or overlapping objects) than classical approaches such as Faster R-CNN.

In semantic segmentation, where the task is to label every pixel into a class, ViTs are just as capable. Its capability to draw information from an input image results in a more connected and accurate segmentation. For example, the Swin Transformer adopts a hierarchical structure with windowed attentions which is able to capture local and global features efficiently. It has been demonstrated that it is superior than the State-of-the-art CNN-based models such as U-Net, particularly in high-resolution tasks such as medical image segmentation [5]. This feature of ViTs makes them especially well-suited for tasks that require fine-grained pixel-wise predictions.

Both object detection and segmentation task benefited from the top-down hierarchy design of ViTs, which allows local and global information to be incorporated in a natural way, and are thus particularly well-suited to tasks with intricate spatial dependencies.

3.3. Depth Estimation and 3D Vision

Another application area with strong impact for ViTs is depth estimation or predicting the distance of objects from a single image. CNNs usually fail to model the entire scene structure, which is essential for accurate depth prediction, in particular for cluttered and occluded areas.

Global attention mechanisms in Vision Transformers provides a better solution. Ibrahem et al. (2022) [9] proposed a lightweight ViT model for real-time monocular depth estimation, which is a good trade-off between accuracy and speed. Their encoder-decoder architecture based on ViT produced high-quality depth maps with realtime performance at around 20 FPS. This favorable feature of ViTs makes them suitable for scenarios like auto-driving and robotics which require fast and accurate depth perception.

ViTs are also promising for other 3D vision tasks such as 3D reconstruction and Scene understanding. Their spatial relationships modelling with the whole image could provide deeper context-aware and more accurate depth predictions, promoting 3D system towards dynamic scene satisfaction.

3.4. Industrial Applications

In industrial environments, visual inspection is of upmost importance in quality control, defect detection, and predictive maintenance. Conventional systems—frequently built on CNNs or human examination—may not be efficient, especially in processing massive, complex input data or subtle outliers. Recent works have also attracted much interest in ViTs, which provide an alternative due to their global reasoning ability. For instance, Hütten et al. (2022) [1] leveraged ViTs on railway freight car maintenance to diagnose defects in metal and wooden structures. They were able to outperform CNNs in detecting cracks and scratches by capturing the presence of both fine details and context from a larger image area. To improve speed and efficiency, hybrid models, such as TurboViT, have also been proposed. At the intersection of CNN efficiency and ViT contextual awareness, these models would be suitable for real-time industrial use [6]. With efficient computation without accuracy compromise, these models also enable rapid and robust inspections in sectors such as automotive, manufacturing and transportation.

3.5. Medical Image Analysis

Last but not least, the most important field for applying Vision Transformers is medical image analysis. ViTs have proven to be highly valuable also in the medical image computing research, such as interpreting subtle visual cues is important for correctly diagnosing diseases. Compared to CNNs that can concentrate on limited local features, ViTs have a more global view since they model the entire context of the image, an important feature in disease detection and diagnosis.

One prominent application is in fundus image classification used for diagnosing retinal disease. Bi et al. (2023) [7] proposed MIL-ViT, which combines multiple instance learning with ViTs. This model is very good at recognizing subtle, scattered signs of conditions like diabetic retinopathy and glaucoma. MIL-ViT can jointly learn global and local representations, which helps to obtain better classification results, compared with CNNs.

Besides for classification, to address the problem of medical segmentation and lesion detection, input context with small anomalies to accurate boundary line is crucial.

As they are able to capture long-range and finely detailed features [1], they are advantageous in spatially precise predictions.

Vision Transformers have shown wide coverage and strong performance in different computer vision problems. From image-level to object-level, and from the industrial visual inspection to medical image analysis, the ViTs can offer an alternative solution to many challenges faced by the conventional CNN-based methods. They are able to represent long-range dependencies and capture global context, so they are excellent for modeling problems with complex spatial interactions. With more and more investigation of ViT and constant structure optimization by replacing, incremental and integration, ViTs territo ry in computer vision will be greatly expanded.

4. Techniques to Improve ViT Performance

Vision Transformers (ViTs) have shown great potential in computer vision with their capacity to model long-range dependencies and encode global context. However, they are computationally expensive, and memory demanded will hinder their applications in practice, particularly on devices of limited resources. This part presents few major tactics that have been suggested for improving the ViT efficiency and practicality Key techniques to improve ViT efficiency, Efficiency improvements, Hybrid models and dynamic/adaptive architectures.

4.1. Efficiency Enhancements

A primary limitation of ViTs is their high computational cost, mainly due to the self-attention mechanism whose complexity scales quadratically with the number of image patches. Addressing this requires strategies that maintain

performance while reducing computational demands.

4.1.1. Low-bit Encoding

Low-bit encoding, like Posit Arithmetic, has potential to be a solution. ViTs such as traditional run on 32- or 64-bit floating point representations which consumes a significant amount of memory and processing capability. On the other hand, Posit Arithmetic uses a smaller number of bits to encode parameters, which can reduce model size and the amount of energy consumption while maintaining accuracy [4]. It enables high precision even for very low bit-width values compared to other standard binary formats. Applying this treatment to ViTs results in faster computation and lower memory requirements, enabling deployment to embedded and mobile devices.

4.1.2. Model Compression Techniques

Pruning is an alternative solution that involves eliminating unnecessary parameters or connections in the model. Number of operations is reduced with no loss in latency. This is part of what makes ViTs more lightweight, and suitable for real-time applications [6]. In addition, knowledge distillation enhances efficiency by training a smaller "student" model to emulate the output of a larger "teacher" model. This method ensures that compact ViTs are able to achieve high accuracy with less computational resources. One such example is Data-efficient Image Transformer (DeiT) that employ distillation to lower the requirement for train-ing data Whilst obtaining competitive results [10].

These efficiency improvement methods, such as Posit Arithmetic, pruning and distillation, are essential in order to make ViTs more practical in places where resources are limited, such as mobile, autonomous, and on the edge devices.

4.2. Hybrid Models

Although ViTs are good at modeling global dependencies, they are not as effective in learning local image representations. Joint ViT and CNN models aim to tackle this problem by taking advantage of the local feature extraction offered by CNNs and the global attention mechanisms of ViTs.

Figure 2 shows yet another hybrid model that combines CNN layers to capture local features quickly, done so before transformer layers that capture long-range dependencies. This architecture strikes a balance between efficiency and performance and mitigates the shortcomings of standalone ViTs.

TurboViT is one of such hybrid models that hybridize CNN based operations with transformer attention to effectively capture local and global features [6]. To reduce complexity, TurboViT adds a mask unit attention mechanism and uses Q-pooling, which can obtain more than 2.4x the model size reduction and 3.4x fewer FLOPs than FasterViT with small degradation in accuracy, making it suitable for real-time and edge scenarios. DMFormer [13] further fuses DMA with convolution layers for multi-scale feature representation. The model is based on multi-kernel size and dilation convolution which helps the model learn features at different levels of resolution and shows a better performance in both image classification and segmentation.

Such hybrid models combine the advantages of CNNs and ViTs and could lead to more effective and efficient architectures. It combines the local receptive field in CNNs and global context modeling capability of ViTs, leading to accurate and efficient performance as compared with ViTs and CNNs alone methods.

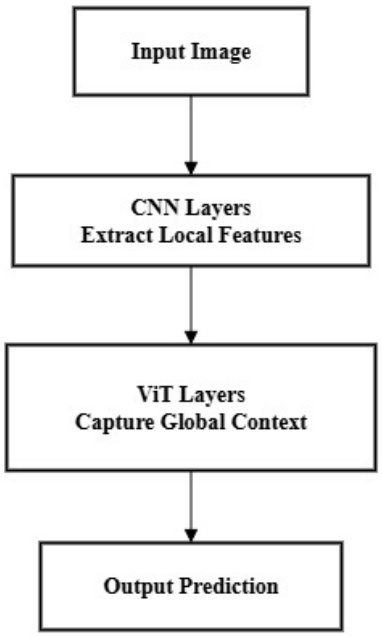

Fig.2 Techniques to Improve ViT Performance

4.3. Dynamic and Adaptive Models

Models that vary the complexity for an input data are often needed in practical applications. In ViTs the analogue involves functionally altering the number of attention heads, layers or patch resolution according to task difficulty. These models increase computational efficiency but maintain efficacy.

MIA-Former is a type of dynamic ViT architecture for adaptive computation according to the image complexity [14]. It uses a Focus and Forget paradigm to turn on more transformer layers for complex images and fewer for easier ones. This flexibility enables MIA-Former to balance computation and accuracy, which is particularly favorable for real-time inference and edge deployment.it has been noted that these are particularly interesting models for edge computing, given the constrained resources there. By adapting complexity to input needs, adaptive ViTs can accommodate smaller models without sacrificing performance [15].

In conclusion, dynamic models such as MIA-Former, as well as efficiency methods (pruning, quantization, and KD) and hybrid architectures (e.g., TurboViT and DMFormer) are instrumental in making ViTs ready for real-world applications, ranging from health to automation, with more to come in the future.

5. Challenges in Deploying Vision Transformers (ViTs)

Despite their strong performance in computer vision tasks, Vision Transformers (ViTs) face several challenges when transitioning from research to real-world deployment. Key challenges include high computational demands, limited scalability due to memory constraints, and the need for large labeled datasets.

5.1. High Computational Cost

The quadratic computational complexity of the self-attention mechanism is one of the major challenges Vision Transformers (ViTs) have to deal with. In ViTs, an image is broken down into fixed-size patches, and every patch attends

to all other patches in the image. This leads to a computation complexity growing quadratically with the number of patches, which causes the model to be highly resource-hungry when it is applied to high-resolution images [3].

Such computation overheads may seriously limit the wide-distribution of ViTs in practical real-time applications e.g., autonomous driving, robotic, video surveillance, which requires low latency and high throughput. Even with enabling the use of strong GPUs / TPUs achieving an acceptable real-time performance could be difficult, especially when you work in an edge / mobile environment. In this scenario, model quantization has become a promising approach, compressing the computational complexity by lowering the precision of operations, such that reduced computational, and memory overhead can be demanded, and having little impact on other factors.

To mitigate this, approaches like local self-attention (as in Swin Transformer) reduce complexity by restricting attention to small windows. In addition, model pruning, quantization, and knowledge distillation (e.g., DeiT) have been used to reduce the FLOPs while maintaining the performance [4][5]. However, even with these efficiencies, self-attention remains the primary computational bottleneck in using ViTs.

5.2. Scalability and Memory Constraints

Another drawback is that ViTs scalability is not well-established, especially in terms of large models memory footprint and compute requirements. For instance, ViT-H/14 has more than 632 million parameters, needing 2528 MB memory and 162 GFLOPs per inference [5]. This makes it inapplicable to memory-limited scenarios, including embedded devices and mobile phones.

Big ViTs achieve better accuracy, but they are computationally heavy and not deployable on different real-world systems. Such models typically require special accelerators or cloud resources, which makes them inflexible and hard to deploy uniformly to various devices.

For scalability purposes, model compression methods, such as weight pruning, low-bit quantization and knowledge distillation, have been extensively employed. These are useful to reduce the number of parameters as well as memory usage and they enable ViTs to run on edge devices without a steep loss in performance [2][10]. Hybrid architectures marrying instructors of CNN with ViT units (e.g., TurboViT) provide another potential direction for scalable solutions for low-resource deployment [6].

5.3. Data Requirements

ViTs might need large-scale labeled data for effective training similar to CNNs, as they do not have inductive biases of CNNs (e.g., locality and translation invariance). Unlike CNN, which are able to generalize well with little data, ViTs behaves sub optimally if trained on a small dataset as it relies on global context learning.

Typically, training ViTs from scratch entails requiring ImageNet-sized or larger datasets. However, in specialized domains like in medical imaging (MRI, CT, X-Ray, etc.) or industrial vision inspection, labeled datasets are often limited out of privacy and labeling costs, and the imbalanced distribution of classes [1][7].

For example, in medical application, e.g., retinal disease classification, ViTs demand thousands of fundus images that are annotated: it means, the lack of available such images are a concern. Likewise, in the industrial area, if trained for pricing defect, ViTs fail to generalize for pricing as there are rare instance of defect in the data. To address this, researchers make use of transfer learning (e.g. pre-training on generic datasets and fine-tuning), data augmentation, and semi-supervised learning. Self-supervised methods have also demonstrated the potential of ViTs for learning representations from un-labeled data, a critical requirement in domains with scarce annotations [2].

Table 2 summarizes the main challenges that hinder the deployment of Vision Transformers in practical settings. These are computational inefficiency (quadratic self-attention), the memory requirement of large-scale ViTs and the large amount of labeled data required for training. The table summarizes key points covered in this section, and can facilitate in identifying the next steps for future development and capacity enhancement.

Table 2. Challenges in Deploying ViTs

Challenge	Description	Impact
High Computational Cost	ViTs require large computational resources due to the quadratic complexity of the self-attention mechanism.	Hinders real-time deployment and increases inference time.
Scalability and Memory	Larger ViTs have high memory demands and require large amounts of storage for model parameters.	Limits deployment on edge devices or systems with limited resources.
Data Requirements	ViTs require large labeled datasets to perform well, especially for specialized tasks like medical imaging.	Makes training difficult in data-scarce domains, such as medical or industrial fields.

In summary, Vision Transformers face three major deployment challenges: computational inefficiency due to self-attention, scalability issues linked to large model sizes, and extensive data demands that restrict their applicability in specialized fields. While recent innovations in architecture design, model compression, and data-efficient learning have helped alleviate these issues, continued research is essential to make ViTs practical for widespread real-world use.

Future Directions

The field of Vision Transformers (ViTs) has seen rapid advancements, but several areas remain ripe for exploration. These advancements are expected to further optimize ViTs for real-world applications across various industries. This section highlights three major future directions for ViTs: improved hybrid models, edge and real-time applications, and multimodal learning.

5.4. Improved Hybrid Models

ViTs are good at capturing global context and long-range dependencies, but the CNNs are still very efficient in dealing with local features. The fusion of both models in hybrid architectures offers substantial potential in terms of performance and efficiency. Such hybrid models can provide end-to-end best of both worlds by utilizing CNNs to extract local patterns and ViTs to reason more completely at a higher range, through a deep global mechanism.

Recent models have referred to this complementary between transformers and convolutional operators, and attempted to merge these two components to combine the best from both sides: transformers pipelines are enriched with convolutional layers, to better exploit input image high

resolutions and smaller computational footprints [5] [6]. As future hybrid architectures are concerned, this combination can be further tuned by using multi-scale attention mechanisms to enable models process fine-grained and high-level visual features jointly. Such advances are especially important in semantic segmentation, object detection, and medical image analysis, where local features are often combined or compared to their context together.

5.5. Edge and Real-Time Applications

With the increasing demand for real-time visual processing on IoT devices, autonomous cars and smart products, there is an urgent requirement for ViTs that can work with limited computation and memory resources. At present, the large resource consumption of ViTs makes them inapplicable to edge computing.

Further gains will rely on optimizing ViTs for low-latency, high-throughput inference. Compression and optimization approaches such as pruning, quantization, and low-bit approximation (e.g., Posit Arithmetic) will be needed to minimize the model size and computational cost, while preserving the accuracy [4]. Furthermore, dynamic pruning approaches may adjust model complexity according to input difficulty, in an attempt to achieve maximum efficiency in input-rich scenarios.

Edge-specific approaches, including using FPGAs and ASICs, could also increase the deployability of ViTs on non-GPU devices. Hybrid models like TurboViT, which trade-off on CNN efficiency versus transformer expressiveness, could serve as a key ingredient towards latency-sensitive applications, such as autonomous navigation, AR/VR, and robotic control [6].

5.6. ViT in Multimodal Learning

Vision Transformers are also expected to have great potentials in multimodal learning, which learns from information in multiple modalities, such as images, texts and audios, to construct more intelligent systems. As such, ViTs are inherently compatible with other transformer models in the context of language processing.

Figure 3 depicts a Multimodal architecture, which is similar to the ones in ViT and a language model here is also processing text data. The outputs are then fused for tasks such as image captioning, visual question answering (VQA), and image-text retrieval.

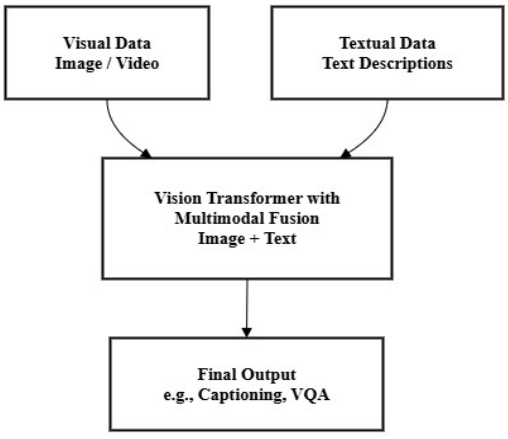

Fig.3 Multimodal Learning Example (ViT + Text)

Other systems such as CLIP and DALL·E have empirically shown the utility of integrating the two modalities by training them on image-text pairs in large-scale dataset. Extensions in other directions could involve further tying of ViTs to language models (such as BERT and GPT), with joint training and more precise semantic reasoning.

Also, in addition to text, ViTs can even be included into multisensory systems with audio, video, and sensor data, which can be particularly well-suited for autonomous driving and smart surveillance which require sensing and understanding the world

6. Conclusion

In this paper, we have delved into the strong power of Vision Transformers (ViTs) in computer vision. The reason for it is that ViTs outperform normal Convolutional Neural Networks (CNN) in both aspects of getting global context and modeling long-range dependencies in images. Complemented by their effectiveness in diverse applications e.g., image classification, object detection, semantic segmentation, and medical image analysis, where they have yielded state-of-the-art performance.

But, the use of ViTs in real-world scenarios has some issues. The quadratic complexity of the self-attention module and high memory cost make ViTs difficult to apply in low-resource models. Whats more, their success requires large annotated data, which is not always available in narrow-class tasks such as medical image analysis and industrial detection.

In order to address these challenges, novel approaches have been proposed, including model compression techniques, hybrid CNN-ViT architectures and adaptive models which can smoothly adapt their complexity to the input. These developments are making ViTs more applicable to real-time tasks and edge computing environment.

In the future, we expect ViTs to have a much wider and increasing effect on computer vision and AI. The continued development of hybrid modeling and data-efficient training will continue to make them applicable on low power devices. In parallel, their incorporation within multimodal learning pipelines (jointly leveraging textual and visual inputs) will be instrumental in building more intelligent and context-aware AI systems.

As ViTs advance, they are likely to have a revolutionary impact on autonomous applications, instantaneous decision-making as well as AI-based products in fields ranging from healthcare and transportation to manufacturing. With continuous research and development, Vision Transformers will be one of the building blocks for future computer vision and artificial intelligence.

References

[1] Hütten, N., Meyes, R., & Meisen, T. (2022). Vision Transformer in Industrial Visual Inspection. Applied Sciences, 12(23), 11981. https://doi.org/10.3390/app122311981

[2] Papa, A., Cuozzo, M., & Bartoli, A. (2024). A Survey on Efficient Vision Transformers Algorithms, Techniques, and Performance Benchmarking. Journal of Artificial Intelligence, 36(1), 45-59.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Kumar, A., Yadav, P., & Patel, S. (2024). Breaking New Ground in AI with Posit Arithmetic and Vision Transformers.

IEEE Access, 11, 3452–3461. https://doi.org/10.1109/ACCESS.2024.3545639

[5] Liu, Z., Lin, Y., & Liu, M. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 1006–1014. https://doi.org/10.1109/ICCV48922.2021.01004

[6] Wong, A., Abbasi, S., & Nair, S. (2023). TurboViT: Generating Fast Vision Transformers via Generative Architecture Search. IEEE Transactions on Neural Networks and Learning Systems, 34(7), 3451-3465. https://doi.org/10.1109/TNNLS.2023.3322150

[7] Bi, Q., Sun, X., Yu, S., Ma, K., Bian, C., Ning, M., He, N., Huang, Y., Li, Y., Liu, H., & Zheng, Y. (2023). MIL-ViT: A multiple instance vision transformer for fundus image classification. Journal of Visual Communication and Image Representation, 97, 103956. https://doi.org/10.1016/j.jvcir.2023.103956

[8] Park, J. G., Amangeldi, A., Fakhrutdinov, N., Karzhaubayeva, M., & Zorbas, D. (2025). Patch and Model Size Characterization for On-Device Efficient-ViTs on Small Datasets Using 12 Quantitative Metrics. IEEE Access.

[9] Ibrahem, H., Salem, A., & Kang, H.-S. (2022). RT-ViT: Real-Time Monocular Depth Estimation Using Lightweight Vision Transformers. Sensors, 22(10), 3849. https://doi.org/10.3390/s22103849

[10] Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 478–487. https://doi.org/10.1109/ICCV48922.2021.00485

[11] Yuan, L., et al. (2021). Tokens-to-Token Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 6747–6756. https://doi.org/10.1109/ICCV48922.2021.00661

[12] Graham, B., El-Nouby, A., & Saad, A. (2021). LeViT: A Vision Transformer in 2021. IEEE Transactions on Image Processing, 45(5), 2423–2432.

[13] Wei, Z., Pan, H., Li, L., Lu, M., Niu, X., Dong, P., & Li, D. (2023, June). DMFormer: Closing the gap between CNN and vision transformers. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[14] Haruna, Y., Qin, S., Chukkol, A. H. A., Yusuf, A. A., Bello, I., & Lawan, A. (2025). Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. Engineering Applications of Artificial Intelligence, 144, 110057.

[15] Ranjan, N., & Savakis, A. (2024, June). Vision transformer quantization with multi-step knowledge distillation. In Signal Processing, Sensor/Information Fusion, and Target Recognition XXXIII (Vol. 13057, pp. 283-292). SPIE.