A Comprehensive Literature Review with Self-Reflection

Literature Review

October 7, 2025

Abstract

This literature review provides a comprehensive analysis of recent research in the field. The review synthesizes findings from 367 research papers, identifying key themes, methodological approaches, and future research directions.

Contents

1 Introduction to Visual Transformers

1.1 The Rise of Transformers in Deep Learning

The landscape of deep learning for sequence modeling underwent a profound transformation with the advent of the Transformer architecture, which effectively addressed critical limitations inherent in prior recurrent and convolutional neural networks. Before this paradigm shift, models like Recurrent Neural Networks (RNNs) and their more sophisticated variants, Long Short-Term Memory (LSTM) networks **?**, faced significant challenges in capturing long-range dependencies. These challenges stemmed from issues such as vanishing or exploding gradients and their inherently sequential processing nature, which fundamentally hindered parallelization and scalability for longer sequences. Convolutional Neural Networks (CNNs), while powerful for local feature extraction and exhibiting strong inductive biases like locality and translation equivariance, were not primarily designed for global context modeling across extended sequences. They often required complex architectural designs, such as deeper stacks or dilated convolutions, to achieve broader receptive fields, which could still be limited compared to the entire sequence.

The seminal paper "Attention Is All You Need" by Vaswani et al. (2017) introduced the Transformer, a novel architecture that entirely eschewed recurrence and convolutions, relying instead on a multi-head self-attention mechanism to draw global dependencies between input and output elements **?**. This groundbreaking work fundamentally altered the approach to sequence-to-sequence tasks, particularly in Natural Language Processing (NLP). The core innovation of the Transformer lies in its ability to dynamically weigh the importance of different parts of the input sequence when processing each element, irrespective of their positional distance **??**. Unlike RNNs, which process tokens one by one, the self-attention mechanism enables parallel processing of all tokens in a sequence, significantly accelerating training times and improving scalability for longer sequences **?**. Beyond self-attention, the Transformer architecture also incorporates crucial components such as positional encodings to inject information about the relative or absolute position of tokens in the sequence, and feed-forward networks for non-linear transformations, all contributing to its robust representational capacity.

The self-attention mechanism computes a weighted sum of all input elements, where the weights are dynamically calculated based on the similarity between a query (representing the current element being processed) and all keys (representing other elements in the sequence). More precisely, the attention weights are derived by computing the scaled dot-product between the query vector (Q) and key vectors (K) for all elements, followed by a softmax function to normalize these scores, and then multiplying by value vectors (V) to obtain the output: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d\_k}})V$, where d_k is the dimension of the keys **??**. This direct, unconstrained interaction between any two positions in the input sequence ensures that long-range dependencies are effectively captured, a significant improvement over the limited receptive fields of CNNs or the vanishing gradient issues prevalent in RNNs. By allowing each element to attend to all other elements, the Transformer can build a rich contextual representation for every token, reflecting its global relationships within the sequence. Furthermore, the multi-head aspect of self-attention allows the model to jointly attend to information from different representation subspaces at different positions, enriching the model's ability to focus on various aspects of the input simultaneously and enhancing its overall expressiveness **?**.

This breakthrough capability of the Transformer to efficiently model global dependencies and process sequences in parallel profoundly impacted deep learning, establishing a new state-of-the-art across numerous NLP tasks, including machine translation, text summarization, and question answering **?**. The subsequent development of large-scale pre-trained models like BERT **?** further demonstrated the Transformer's power, showcasing its ability to learn highly generalizable language representations. The conceptual elegance and computational efficiency of self-attention laid the critical groundwork for its eventual adaptation and widespread adoption beyond NLP. Critically, the success in NLP prompted researchers to challenge the long-held axiom that CNNs' inherent inductive biases, such as locality and translation equivariance, were indispensable for computer vision tasks **?**. The Transformer, by design, lacks these built-in biases, and overcoming this perceived limitation became a key conceptual hurdle. This motivated the direct application of a sequence-processing architecture to image data, where images could be re-conceptualized as sequences of patches, allowing the powerful self-attention mechanism to capture global visual relationships ?. This paradigm shift directly challenged the long-standing dominance of recurrent and convolutional networks across various deep learning applications, particularly paving the way for Vision Transformers (ViTs).

The Transformer architecture, with its innovative self-attention mechanism and parallel processing capabilities, thus marked a pivotal moment in deep learning. Its capacity for robust capture of long-range dependencies not only revolutionized NLP but also established a foundational conceptual framework that would eventually inspire a new generation of models in computer vision and other domains, fundamentally altering how deep learning models perceive and process complex data. This conceptual leap set the stage for the subsequent exploration of Vision Transformers, which sought to leverage these advantages for visual understanding, despite the architectural differences from traditional CNNs.

1.2 Motivation for Vision Transformers: Beyond CNNs

Convolutional Neural Networks (CNNs) have historically dominated computer vision, largely owing to their strong inductive biases of locality and translation equivariance. These biases, inherent in convolutional operations, enable efficient extraction of local features and parameter sharing, leading to remarkable success in tasks like image classification and object recognition over the past decade ?. However, the very nature of these local operations, while beneficial for capturing fine-grained patterns, inherently restricts CNNs' ability to effectively model global relationships and long-range dependencies across an entire image. This limitation is particularly pronounced when understanding complex scenes or objects where interactions between spatially distant parts are crucial ??.

To compensate for this restricted receptive field, CNN architectures often resort to complex designs such as stacking many convolutional layers, employing large kernel sizes, using dilated convolutions, or incorporating pyramid structures to aggregate broader context and simulate a global view ?. While these strategies have pushed the boundaries of CNN performance, they often introduce significant computational overhead, increase model depth, or still rely on an indirect, hierarchical aggregation of information rather than direct global interaction. For instance, in medical imaging, the "inadequacy of long-range relation modeling" in CNNs poses challenges for comprehensive understanding of complex anatomical structures **?**, and their "restricted local receptive field limits their ability to learn global context information" in breast ultrasound image classification **?**. The need for a more direct and flexible mechanism to capture these global interactions served as a primary motivation for exploring alternative architectures.

The advent of Vision Transformers (ViTs) offered a fundamentally different paradigm, directly addressing the CNNs' limitations in global context modeling. Inspired by the success of Transformers in Natural Language Processing (NLP) for handling long-range dependencies in sequential data, ViTs re-conceptualized image processing. By treating images as sequences of flattened patches, which are then linearly embedded and processed by a standard Transformer encoder, ViTs leverage the multi-head self-attention mechanism to establish direct relationships between any two patches, irrespective of their spatial distance **?**. This design inherently provides a global receptive field from the very first layer, allowing the model to weigh the importance of all input patches when processing each individual patch, thereby enabling a more holistic understanding of the visual scene. This radical departure from convolutional locality represented a significant architectural shift, promising to overcome the architectural constraints that plagued CNNs in capturing broad context.

However, this paradigm shift was not without its own challenges, which in turn motivated subsequent waves of research within the ViT domain. By largely abandoning the strong inductive biases of locality and translation equivariance inherent in CNNs, the original ViTs exhibited a significant dependence on vast amounts of pre-training data to achieve competitive performance **???**. This "data hunger" stemmed from their more general-purpose architecture, which required extensive exposure to diverse visual patterns to learn robust representations without the built-in priors of convolutions. Furthermore, the initial ViT models, with their uniform patch processing, lacked the inherent hierarchical feature extraction capabilities that CNNs naturally provided, which are crucial for dense prediction tasks like object detection and semantic segmentation **?**. The quadratic computational complexity of global self-attention with respect to image resolution also presented a practical hurdle for high-resolution inputs. These new challenges, arising directly from ViTs' novel approach to global context modeling, became the driving force for subsequent architectural innovations aimed at improving data efficiency, computational scalability, and the ability to generate multi-scale features, thereby making ViTs more robust and versatile for a broader range of computer vision applications.

1.3 Scope and Structure of the Review

This literature review is meticulously structured to provide a comprehensive and navigable roadmap through the rapidly evolving landscape of Visual Transformer (ViT) research. Its organizational framework is designed to facilitate a pedagogical progression, tracing the intellectual trajectory of ViTs from their foundational concepts to their most advanced applications and future challenges. This approach ensures a coherent narrative that highlights the chronological development, interconnectedness of ideas, and the continuous evolution of this transformative field.

The review commences with **Section 1: Introduction to Visual Transformers**, which establishes the historical context of deep learning for vision, critically examines the limitations of traditional Convolutional Neural Networks (CNNs), and introduces the paradigm shift instigated by Transformers. This foundational section sets the stage by delineating the overall scope of the review, emphasizing the journey from initial concepts to sophisticated architectures and diverse applications.

Building upon this introduction, **Section 2: Foundational Vision Transformer Architectures and Early Optimizations** delves into the seminal works that introduced ViTs to computer vision. This section covers the core methodology of processing images with standard Transformer encoders and critically analyzes their initial potential and inherent limitations, such as significant data requirements. It then explores immediate efforts to address these practical constraints through early innovations in training efficiency, stability, and tokenization, including initial investigations into ViT robustness and transferability.

The narrative then transitions to **Section 3: Hierarchical and Efficient Vision Transformer Architectures**. This pivotal section focuses on architectural innovations that transformed ViTs into versatile backbones capable of handling a broader range of computer vision tasks. It details the development of hierarchical structures, multi-scale feature representations, and efficient attention mechanisms designed to overcome the original ViT's quadratic computational complexity and its limitations in dense prediction tasks. The discussion here encompasses various design techniques aimed at enhancing efficiency and scalability of attention mechanisms.

Following the architectural advancements, **Section 4: Self-Supervised Learning Paradigms for Vision Transformers** investigates the transformative role of self-supervised learning (SSL) in unlocking ViTs' full potential. This section details various SSL methodologies, including masked image modeling and self-distillation, which enable ViTs to learn powerful visual representations from vast amounts of unlabeled data. It highlights how these sophisticated pre-training strategies mitigate ViT's initial reliance on massive labeled datasets, paving the way for scalable and robust models, and discusses the scaling of these self-supervised ViTs towards foundation models.

Section 5: Hybrid Architectures and Beyond Self-Attention explores the fascinating convergence of ViTs with CNNs and examines radical alternatives to the self-attention mechanism. This section details hybrid architectures that strategically combine the inductive biases of convolutions with the global context modeling of Transformers, aiming for improved efficiency and robustness. Furthermore, it investigates research that questions the absolute necessity of complex self-attention, proposing simpler yet effective token mixing mechanisms, and demonstrates how modern CNNs, inspired by ViT design principles, can achieve competitive performance.

The review then broadens its scope in **Section 6: Applications and Domain-Specific Adaptations of Visual Transformers**. This section showcases the broad applicability and versatility of ViTs across various computer vision tasks and specialized domains, including object detection, semantic segmentation, medical image analysis, and remote sensing. It illustrates how ViTs have been adapted to excel in these diverse real-world scenarios, often through customized integration strategies, and touches upon their utility in general image classification and emerging 3D vision tasks.

Finally, **Section 7: Future Directions and Open Challenges** provides a forward-looking perspective on the field. It identifies key future research directions, unresolved theoretical questions, and practical challenges, such as the continuous quest for more efficient and scalable architectures, the potential of multimodal and new foundational models, and the exploration of novel architectures beyond traditional attention mechanisms. This section also critically addresses ethical considerations and the broader societal impact of increasingly powerful Vision Transformers, including concerns related to bias, robustness, and responsible AI development. The review concludes with **Section 8: Conclusion**, synthesizing the key findings and offering a final outlook on the trajectory of visual intelligence.

Through this structured progression, the review aims to provide readers with a deep understanding of the intellectual journey, current state, and future potential of Visual Transformer research, emphasizing the dynamic interplay between theoretical advancements and practical applications.

2 Foundational Vision Transformer Architectures and Early Optimizations

2.1 The Original Vision Transformer (ViT) Paradigm

The Vision Transformer (ViT) architecture operationalized a profound shift in computer vision, directly applying the Transformer architecture, originally developed for natural language processing, to image recognition tasks ?. This groundbreaking work fundamentally re-conceptualized image processing, moving away from local convolutional operations to a sequence-to-sequence approach, treating images as sequences of flattened patches rather than relying on inherent spatial hierarchies.

At its core, the ViT architecture processes an image by first dividing it into non-overlapping, fixed-size square patches (e.g., 16x16 pixels). Each patch is then flattened into a 1D sequence of pixel values and linearly projected into a higher-dimensional embedding space. To re-introduce crucial spatial information lost during the flattening, learnable positional encodings are added to these patch embeddings, allowing the model to understand the relative positions of different patches within the original image **?**. These embedded patches, along with an additional learnable "class token" (a direct borrowing from BERT's `[CLS]` token in NLP, intended to aggregate global information for classification), form the input sequence to a standard Transformer encoder. This encoder, composed of multiple layers of multi-head self-attention (MSA) and feed-forward networks (FFN), enables the model to capture long-range dependencies and global contextual relationships across the entire image by allowing each patch to attend to all other patches. The final classification is typically performed by a multi-layer perceptron (MLP) head attached to the output of the class token.

The original ViT demonstrated impressive performance, achieving state-of-the-art results on large-scale image classification benchmarks, notably surpassing CNN-based models when pre-trained on massive datasets such as JFT-300M **?**. This success underscored the power of the self-attention mechanism to learn rich, global representations without relying on strong, hard-coded convolutional inductive biases like locality and translation equivariance.

However, despite its groundbreaking performance, the original ViT paradigm critically highlighted significant limitations. Unlike CNNs, which possess strong, hard-coded inductive biases such as local receptive fields and translation equivariance, the original ViT architecture largely foregoes these explicit priors. While the initial patchification step introduces a rudimentary form of local processing, the subsequent global self-attention mechanism is designed to learn relationships across arbitrary distances without inherent spatial constraints **??**. This design choice, while enabling unprecedented flexibility and global context modeling, critically meant that ViTs had to learn these fundamental visual priors from data itself. Consequently, ViTs exhibited substantial data hunger, requiring extensive pre-training on colossal datasets to learn robust visual representations from scratch, making them less competitive than CNNs when trained on smaller, more common datasets like ImageNet-1K without such pre-training ?.

Furthermore, the computational cost associated with the original ViT posed practical challenges. The global self-attention mechanism scales quadratically with respect to the number of input tokens (patches) ?. For instance, doubling the linear resolution of an image (e.g., from 224x224 to 448x448) quadruples the number of patches, leading to a sixteen-fold increase in the computational cost of the attention layers ??. This prohibitive computational and memory overhead, especially for high-resolution inputs, limited its practical deployment and accessibility ??. The reliance on a single `[CLS]` token for global representation, while effective, also represented a direct porting of an NLP mechanism, which would later be critically re-evaluated and often replaced by more vision-centric aggregation strategies in subsequent ViT variants.

In conclusion, the original ViT paradigm successfully demonstrated the viability of Transformers for computer vision, showcasing their unparalleled ability to model global dependencies and learn powerful representations from data. Yet, it simultaneously exposed a fundamental tension: the trade-off between the expressive power of pure self-attention and the practical demands of data efficiency, computational cost, and the need for implicit inductive biases. This initial work laid the foundation for subsequent research to address these limitations, either by developing more data-efficient training strategies or by integrating architectural priors to make Vision Transformers more robust and versatile across various scales and data regimes.

2.2 Data-Efficient Training and Distillation

The initial introduction of Vision Transformers (ViTs) marked a significant paradigm shift in computer vision, demonstrating that pure Transformer architectures could achieve state-of-the-art performance on image recognition tasks ?. However, a critical limitation of the original ViT was its substantial data hunger, necessitating pre-training on colossal datasets like JFT-300M to outperform traditional Convolutional Neural Networks

(CNNs). This reliance on proprietary, extremely large datasets posed a major barrier to broader adoption and practical application. Consequently, early research efforts focused intensely on developing strategies to mitigate this data dependency, making ViTs more accessible and efficient for researchers and practitioners without access to such extensive resources.

A pivotal advancement in addressing ViT's data hunger was the introduction of data-efficient training techniques, most notably knowledge distillation, exemplified by the Data-efficient Image Transformer (DeiT) **?**. DeiT demonstrated that a ViT student model could achieve competitive performance, even with significantly smaller training datasets (e.g., ImageNet-1K), by learning from a pre-trained CNN teacher. This teacher-student paradigm effectively transfers the rich representations and strong inductive biases (such as locality and translation equivariance) inherent in CNNs to the ViT, thereby bootstrapping its performance without requiring massive amounts of labeled data. The distillation process in DeiT involved a specialized distillation token that interacts with the class token and patch tokens, learning to reproduce the teacher's output. Crucially, DeiT's success was also attributed to a sophisticated training recipe that included aggressive data augmentation techniques like RandAugment, Mixup, and CutMix. These augmentations effectively expanded the diversity of the training data, preventing overfitting and enabling the ViT to learn robust features from a comparatively smaller dataset, thus making ViTs practical for a wider range of applications.

Complementing distillation, other approaches sought to imbue ViTs with CNN-like inductive biases directly into their architecture to improve data efficiency. For instance, **?** proposed ConViT, which introduced soft convolutional inductive biases into the self-attention mechanism. By incorporating a gated positional self-attention that leverages local information, ConViT improved the performance of ViTs, particularly when trained on smaller datasets. This method aimed to combine the local feature extraction strengths of CNNs with the global reasoning capabilities of Transformers, making the models more robust to data scarcity by providing better architectural priors. Similarly, **?** introduced Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) as generic add-on modules to enhance the locality inductive bias of ViTs. SPT reorganizes patches to capture local information more effectively, while LSA restricts attention to local neighborhoods, enabling ViTs to learn effectively from scratch even on small-size datasets like Tiny-ImageNet, demonstrating significant performance improvements. These architectural modifications offer an alternative to distillation by intrinsically improving the ViT's ability to learn from limited data.

The concept of distillation itself also evolved, moving beyond CNN teachers. **?** explored "Attention Distillation" for self-supervised Vision Transformers (ViT-SSKD), demonstrating that distilling information directly from the attention mechanism of a teacher ViT to a student ViT could significantly narrow the performance gap between them. This approach proved effective for smaller ViT models (e.g., ViT-Tiny and ViT-Small) and was independent of the specific self-supervised learning algorithm, highlighting the versatility and continued relevance of distillation for improving the data efficiency and performance of ViT students, even when the teacher is another ViT. Furthermore, the challenge of training ViTs from scratch extended beyond classification. **?** investigated training ViT-based object detectors from scratch without large-scale ImageNet pre-training. Their findings revealed that specific architectural changes and extended training epochs played critical roles in achieving competitive performance, underscoring that data efficiency for ViTs is not solely about distillation but also about optimizing the architectural design and training regimen for specific tasks when large pre-training datasets are unavailable.

In summary, the early strategies for data-efficient training and distillation were crucial in overcoming the initial barrier of ViT's immense data requirements. Techniques like knowledge distillation **??** successfully transferred valuable inductive biases and rich representations, enabling ViTs to achieve strong performance with standard datasets. Concurrently, architectural innovations that integrated convolutional priors **?** and enhanced locality through refined tokenization and attention mechanisms **?** further improved ViT's ability to learn from less data and even train from scratch **?**. These advancements collectively transformed ViTs from a theoretical curiosity requiring massive resources into a practical and widely applicable architecture, accelerating their integration into the broader computer vision community. While these methods successfully addressed the data-hunger problem, making ViTs viable on ImageNet, achieving stable training for deeper and more powerful variants presented a new set of challenges related to optimization and token representation, which are explored in the subsequent section.

2.3 Enhancing Stability and Tokenization for Deeper Models

The introduction of Vision Transformers (ViTs) marked a significant paradigm shift in computer vision, demonstrating the potential of self-attention mechanisms for image recognition **?**. However, early ViT models faced considerable challenges, notably their heavy reliance on vast pre-training datasets and inherent instability when scaled to greater depths, which hindered their practical application and ability to train effectively from scratch on standard datasets like ImageNet. Addressing these limitations became crucial for the broader adoption of ViTs.

Initial efforts focused on mitigating the data dependency. For instance, DeiT **?** introduced data-efficient training strategies, including knowledge distillation from a convolutional neural network (CNN) teacher, enabling ViTs to achieve competitive performance on ImageNet with significantly less training data. While this improved data efficiency, the fundamental architectural challenges related to training very deep Transformer models, such as vanishing or exploding gradients and general instability, persisted.

A pivotal advancement in enabling deeper and more stable ViT architectures was presented by CaiT (Class-attention in Image Transformers) **?**. This work directly addressed the instability issues encountered when scaling ViTs to hundreds of layers, a common problem in deep neural networks. CaiT introduced **LayerScale**, a simple yet highly effective mechanism that rescales the residual connections within each Transformer block. By adaptively learning per-channel scaling factors for the output of self-attention and feed-forward layers, LayerScale stabilized the training dynamics, allowing the successful development and training of much deeper ViTs without performance degradation. Furthermore, CaiT proposed **class-attention layers**, where a dedicated class token attends to all image tokens to aggregate global information, but image tokens do not attend to the class token. This asymmetric attention mechanism further enhanced stability and improved the representation learning capabilities of deeper models, demonstrating that ViTs could indeed "go deeper" with appropriate architectural modifications.

Complementing these architectural stability improvements, refinements in the initial tokenization process were explored to better capture fine-grained local image structures and optimize sequence length, thereby facilitating more effective training from scratch. The Tokens-to-Token ViT (T2T-ViT) **?** introduced a novel **Tokens-to-Token (T2T) module** designed to progressively structure local tokens. Unlike the original ViT's approach of simply flattening non-overlapping image patches, the T2T module iteratively aggregates neighboring tokens, effectively modeling local connectivity and reducing the sequence length. This hierarchical tokenization strategy allowed T2T-ViT to capture fine-grained details and local contextual information more efficiently, leading to improved performance. Crucially, this enhanced tokenization, combined with the general advancements in ViT training stability, enabled T2T-ViT to achieve state-of-the-art results when trained from scratch on ImageNet, without requiring extensive pre-training on larger external datasets.

The innovations introduced by CaiT **?** and T2T-ViT **?** collectively marked a significant step forward in the development of Vision Transformers. CaiT's LayerScale and class-attention layers provided the necessary architectural stability for scaling ViTs to unprecedented depths, while T2T-ViT's progressive tokenization improved the capture of local features and reduced computational overhead, making ViTs more amenable to training from scratch on standard datasets. These advancements significantly enhanced the practicality, robustness, and scalability of ViTs, moving them beyond their initial limitations and paving the way for their broader application in diverse computer vision tasks.

Despite these critical advancements, the inherent quadratic complexity of global self-attention, even with reduced sequence lengths from improved tokenization, remained a computational bottleneck for very high-resolution inputs or real-time applications. This limitation spurred further research into more efficient attention mechanisms and hierarchical designs, which would become a subsequent major focus in the evolution of Vision Transformers.

3 Hierarchical and Efficient Vision Transformer Architectures

3.1 Shifted Window-Based Attention for Hierarchical Processing

The initial Vision Transformer (ViT) architecture, while demonstrating the power of self-attention for image classification, suffered from two significant limitations: a quadratic computational complexity with respect to image size, making it impractical for high-resolution inputs, and an inability to generate multi-scale feature maps, which are crucial for dense prediction tasks like object detection and semantic segmentation. Addressing these challenges, the concept of shifted window-based attention emerged as a pivotal innovation, most notably introduced by the Swin Transformer **?**. This mechanism fundamentally transformed ViTs into efficient, general-purpose backbones capable of hierarchical processing.

The Swin Transformer **?** proposed a hierarchical Vision Transformer that computes representations with shifted windows. Its core innovation lies in partitioning images into non-overlapping local windows and applying self-attention only within these windows, thereby reducing computational complexity from quadratic to linear with respect to image size. Crucially, the 'shifted window' approach introduces cross-window connections by cyclically shifting the window partitions between successive self-attention layers. This ingenious mechanism allows for information flow across different windows, effectively expanding the receptive field and enabling the construction of a hierarchical feature representation akin to those in convolutional neural networks (CNNs). This design made Swin Transformer highly efficient and effective, achieving state-of-the-art results across image classification, object detection, and semantic segmentation, establishing it as a versatile backbone for various vision tasks.

The effectiveness of the Swin Transformer's shifted window design quickly led to its adoption and adaptation across various computer vision domains. For instance, SwinIR **?** leveraged the Swin Transformer as a strong baseline for image restoration tasks, demonstrating its capability in low-level vision by incorporating Residual Swin Transformer Blocks. Similarly, for semantic segmentation of remote sensing images, the Class-Guided Swin Transformer (CG-Swin) **?** utilized Swin as an encoder, further validating its suitability for dense prediction by designing a class-guided Transformer block in the decoder. The Lightweight Dual-Branch Swin Transformer (LDBST) **?** for remote sensing scene classification refined the local connection within the Swin framework by integrating a depthwise convolutional layer into the MLP, boosting connections with neighboring windows and demonstrating efforts to make Swin-based architectures more efficient.

While shifted window attention proved highly effective, alternative strategies for achieving multi-scale features and efficient global context were also explored. CrossViT **?**, for example, proposed a dual-branch transformer that processes image patches of different sizes and fuses them using a linear-complexity cross-attention mechanism, offering an alternative to Swin's window-shifting for multi-scale interaction. MobileViT **?** aimed for light-weight, mobile-friendly ViTs by presenting "transformers as convolutions," combining CNN strengths with ViTs for efficient global processing. Interestingly, ViTDet **?** explored plain, non-hierarchical ViT backbones for object detection and found that simple feature pyramids from single-scale feature maps, combined with *non-shifted* window attention aided by very few cross-window propagation blocks, could achieve competitive results. This finding directly challenges the absolute necessity of the *shifted* component for all cross-window interactions, suggesting that sparse global connections can sometimes suffice. Further, the Deformable Attention Transformer (DAT) **?** introduced deformable self-attention, where key and value positions are selected in a data-dependent manner, offering a more flexible and focused alternative to fixed windowing for capturing informative features. Hiera **?** argued that with strong pretraining like Masked Autoencoders (MAE), a simpler hierarchical vision transformer, stripped of some "bells-and-whistles," could be equally effective and faster, implying that the core hierarchical structure might be more critical than the specific shifted window mechanism for certain applications. Other works, such as GSC-ViT **?**, achieved a balance between local and global feature extraction through groupwise separable multihead self-attention, echoing the principles of local processing with global interaction.

The success of the Swin Transformer cemented its role as a foundational backbone, influencing many subsequent works. Numerous studies in remote sensing image analysis, including QAGA-Net **?**, ODDL-Net **?**, CTNet **?**, P2FEViT **?**, and those exploring quantitative regularization **?**, have adopted Swin Transformer (or similar hierarchical ViTs like Next-ViT) as their backbone, often integrating it with CNNs or specialized training strategies to address the unique challenges of remote sensing data. This highlights Swin's versatility and its ability to serve as a robust feature extractor. Beyond direct adoption, the principles of hierarchical processing and efficient attention inspired new hybrid architectures. Next-ViT **?** and TRT-ViT **?** are examples of "next-generation" vision transformers that combine convolutional and transformer blocks in hierarchical designs, specifically optimized for efficient deployment in industrial scenarios. MambaVision **?** represents a recent hybrid Mamba-Transformer backbone that also emphasizes hierarchical architecture for efficient modeling of visual features and capturing long-range spatial dependencies. Even in video action recognition, the need for efficient spatio-temporal processing led to models like TP-ViT **?** and ViT-Shift **?**, which adapt ViT principles to handle temporal dynamics, often benefiting from hierarchical feature extraction.

In conclusion, the introduction of shifted window-based attention, pioneered by the Swin Transformer, marked a paradigm shift in Vision Transformer research. It effectively mitigated the original ViT's limitations of quadratic complexity and fixed-scale processing, making Transformers practical and highly effective for a broad spectrum of computer vision tasks, especially dense prediction. This innovation established a new standard for hierarchical feature representation in ViTs, enabling linear scaling with image size and facilitating cross-window information exchange. While the core shifted window mechanism remains influential, subsequent research has explored various refinements, alternatives, and hybrid architectures, continuously seeking to optimize the balance between computational efficiency, inductive biases, and the expressive power of global attention. The ongoing tension lies in determining the optimal integration of localized processing (like windows or convolutions) with global reasoning, and whether the specific "shifted" mechanism is always necessary or if simpler cross-window connections suffice, particularly when coupled with powerful pre-training strategies.

3.2 Pyramid Structures for Multi-Scale Feature Representation

The initial success of Vision Transformers (ViTs) **?** primarily in image classification highlighted their powerful global reasoning capabilities. However, their inherent design, which typically produces fixed-resolution feature maps and suffers from quadratic computational complexity due to global self-attention, presented significant challenges for dense prediction tasks like object detection and segmentation. These tasks critically demand fine-grained localization and multi-scale contextual understanding, capabilities traditionally dominated by Convolutional Neural Networks (CNNs) through their hierarchical feature extraction. To bridge this gap, a pivotal direction in ViT research has focused on developing pyramid-like architectures that can generate multi-scale feature representations efficiently and without explicit convolutional layers, thereby expanding ViT's applicability beyond simple image classification.

A seminal contribution in this domain is the Pyramid Vision Transformer (PVT) **?**. PVT introduced a novel pyramid structure to progressively reduce the resolution of feature maps, moving from high-resolution, fine-grained features at early stages to low-resolution, semantically rich features at deeper layers. This hierarchical design is achieved through a spatial reduction attention mechanism, which efficiently downsamples the key and value matrices in the self-attention computation. This approach allows PVT to capture both low-level details and high-level semantic information, making it a versatile backbone for dense prediction tasks and a direct competitor to CNN-based feature pyramid networks (FPNs). Crucially, PVT maintains the global context modeling strengths of Transformers while significantly reducing computational costs compared to the original ViT's global attention.

Concurrently, the Swin Transformer **?** emerged as another highly influential hierarchical Vision Transformer, also designed to overcome the computational and multi-scale limitations of earlier ViTs. While its core mechanism of shifted window-based attention is detailed in Subsection 3.1, it is important to note its parallel contribution to generating multi-scale feature representations. By restricting self-attention to non-overlapping local windows and introducing a shifted window mechanism for cross-window connections, Swin Transformer achieves linear computational complexity and a hierarchical feature pyramid. This design has proven exceptionally effective across a wide array of dense prediction tasks, including object detection **??** and semantic segmentation, and even low-level vision tasks like image restoration **?** and monocular depth estimation **?**, underscoring the versatility of such hierarchical designs.

Beyond PVT and Swin, other architectures have explored diverse strategies for generating multi-scale features. The Hierarchical Vision Transformer (HiViT) **??** further refines hierarchical ViT designs, demonstrating advantageous performance, particularly in self-supervised pre-training methods like masked image modeling. HiViT emphasizes architectural simplicity, arguing that many "bells-and-whistles" added to hierarchical ViTs are unnecessary when combined with strong pre-training, leading to faster and more accurate models. In contrast to fixed pyramid structures, the Deformable Attention Transformer (DAT++) **?** introduces a novel deformable multi-head attention module. This mechanism adaptively allocates key and value positions in a data-dependent way, allowing the model to dynamically focus on relevant regions and overcome the data-agnostic nature of handcrafted attention patterns in some pyramid designs. This offers a more flexible approach to capturing multi-scale and long-range relationships.

Interestingly, some research has challenged the necessity of *inherently hierarchical* ViT backbones for dense prediction. ViTDet **?** demonstrates that a simple feature pyramid can be effectively built from a *plain*, single-scale Vision Transformer feature map, especially when the ViT is pre-trained with Masked Autoencoders (MAE). This suggests that the rich representations learned by plain ViTs can be adapted for multi-scale tasks with minimal architectural modifications, such as using window attention with few cross-window propagation blocks. This approach simplifies the backbone design while achieving competitive results in object detection.

Furthermore, hybrid approaches have emerged, combining the strengths of ViTs with traditional CNN-based FPNs to enhance multi-scale feature learning. For instance, the Feature Pyramid Vision Transformer (FPViT) **?** integrates Transformers with a ResNet backbone and a feature pyramid, allowing the Transformers to capture global contexts from CNN-extracted features while leveraging multi-scale maps for better adaptability in medical image classification. Similarly, Transformer-Based YOLOX **?** employs a pre-trained ViT as a backbone and integrates an FPN decoder to effectively aggregate multi-level features for object detection, demonstrating superior performance on challenging datasets. Other specialized models like the Visual Saliency Transformer (VST) **?** leverage multi-level token fusion and upsampling within a pure transformer framework to generate high-resolution saliency maps, while SENet **?** incorporates a local information capture module to compensate for the patch-level attention mechanisms in pixel-level tasks like camouflaged object detection and salient object detection, further reinforcing the critical need for fine-grained multi-scale features.

In conclusion, the development of pyramid Vision Transformer architectures, spearheaded by models like PVT and Swin Transformer, represents a significant evolution in computer vision. By ingeniously designing hierarchical structures and employing efficient attention mechanisms, these models have successfully addressed the critical need for multi-scale feature representation and improved computational efficiency. This has expanded ViT's applicability far beyond simple image classification, enabling them to serve as versatile and powerful backbones for a broad spectrum of dense prediction and other complex vision tasks, directly challenging the long-standing dominance of CNNs. The ongoing research continues to explore novel attention mechanisms, architectural simplifications, and hybrid integrations to further enhance their efficiency and adaptability across diverse applications.

3.3 Multi-Scale and Efficient Attention Mechanisms

The initial promise of Vision Transformers (ViTs) was significantly constrained by their quadratic computational complexity with respect to image resolution and their inherent limitations in capturing multi-scale features, which are indispensable for dense prediction tasks. While hierarchical architectures like the Swin Transformer **?** and Pyramid Vision Transformer (PVT) **?** (as discussed in Sections 3.1 and 3.2, respectively) laid foundational groundwork by introducing window-based and spatial reduction attention for linear complexity and multi-scale feature generation, subsequent research has delved deeper into refining the core attention mechanism itself to further enhance efficiency and multi-scale processing capabilities. This continuous effort is crucial for enabling ViTs to serve as versatile and deployable backbones across diverse vision applications.

A primary avenue of innovation has focused on optimizing spatial attention to balance global context and local detail more effectively, often by intelligently structuring local interactions. *Multiscale Vision Transformers* (MViT) **?** addressed the challenge of scaling to higher resolutions by progressively expanding channel dimensions while reducing spatial resolution within a hierarchical structure, enabling robust performance on high-resolution inputs, including video. This approach efficiently aggregates information across scales. Complementing this, the *Twins* architecture **?** refined spatial attention by explicitly combining both global and local attention mechanisms. This hybrid strategy aims to leverage the benefits of fine-grained local processing and broad contextual understanding, thereby creating a robust general-purpose backbone. Further expanding on localized attention, the *CSWin Transformer* **?** introduced cross-shaped window attention, which captures richer contextual information more efficiently than square windows by attending to features along horizontal and vertical strips. Similarly, *RegionViT* **?** proposed a regional-to-local attention strategy, allowing efficient processing of large images by first attending to broader regional features and then refining with local attention. This hierarchical attention within a single block effectively balances computational cost with comprehensive feature extraction. *Focal Attention* **?** also contributes to this theme by efficiently capturing both fine-grained local details and broader contextual information, proving particularly beneficial for tasks like document understanding where varied scales of information are critical. These methods collectively demonstrate a trend towards more sophisticated, yet computationally constrained, local-global attention interactions, moving beyond simple windowing to more adaptive spatial sampling.

Beyond static windowing, a significant advancement in efficient attention involves making the attention mechanism data-dependent and dynamic. *Deformable Attention Transformers* (DAT) **?** and its enhanced version DAT++ **?** introduced a novel deformable multi-head attention module. Unlike fixed-grid or window-based attention, deformable attention adaptively allocates the positions of key and value pairs in a data-dependent manner. This flexible scheme allows the model to dynamically focus on relevant regions, overcoming the limitations of handcrafted attention patterns (like those in Swin or PVT) and maintaining the representational power of global attention while significantly reducing computational cost and memory usage. This approach represents a critical step towards more intelligent and adaptive attention mechanisms that can dynamically adjust their receptive fields based on image content, thereby enhancing both efficiency and feature discriminability.

Another crucial direction for efficiency has been the development of linear attention mechanisms and strategies for reducing spatial redundancy. Traditional self-attention's quadratic complexity stems from the softmax operation and the dense interaction matrix. *UFO-ViT* **?** proposed a novel self-attention mechanism with linear complexity by eliminating non-linearity and factorizing the matrix multiplication without complex linear approximations. This direct approach to linear scaling offers substantial computational benefits, especially for high-resolution inputs. Concurrently, methods focusing on reducing spatial redundancy by pruning or merging tokens have gained traction. The *Localization and Focus Vision Transformer* (LF-ViT) **?** strategically curtails computational demands by processing a reduced-resolution image in a "Localization" phase. If a definitive prediction is elusive, it triggers a Neighborhood Global Class Attention (NGCA) mechanism to identify and spotlight class-discriminative regions from the original image in a "Focus" phase. This selective processing significantly reduces FLOPs. Similarly, *CP-ViT* **?** introduced a cascade pruning framework that dynamically predicts sparsity in ViT models, progressively pruning uninformative patches and heads. By defining a cumulative score and adjusting pruning ratios based on layer-aware attention range, CP-ViT achieves substantial FLOPs reduction with minimal accuracy loss. Furthermore, *LTM-Transformer* **?** proposes a novel block with Learnable Token Merging (LTM), which reduces FLOPs and inference time by merging tokens in a learnable scheme, compatible with various existing Transformer networks. These token-level optimization strategies demonstrate that efficiency can be gained not just by altering the attention computation itself, but also by intelligently reducing the amount of information that needs to be processed.

The broader landscape of efficient hierarchical designs also continues to evolve. *Hiera* **?** exemplifies how architectural simplicity, when combined with strong self-supervised pretraining (like MAE), can yield highly efficient and accurate hierarchical Vision Transformers. By stripping away many "bells-and-whistles" commonly added for supervised performance, Hiera achieves faster inference and training while maintaining competitive accuracy, suggesting that computational overhead can be reduced through a holistic approach encompassing both architectural design and pre-training strategy.

In conclusion, the continuous pursuit of multi-scale and efficient attention mechanisms has been paramount for the practical deployment of Vision Transformers **?**. From sophisticated spatial attention refinements that balance local and global context **?????**, to dynamic and data-dependent mechanisms like deformable attention **??**, and radical approaches that achieve linear complexity or reduce spatial redundancy through token pruning/merging **????**, the field strives to balance the expressive power of global attention with computational feasibility. Unresolved issues include finding the optimal trade-off between incorporating inductive biases (like locality) and maintaining the flexibility of pure attention, as well as exploring novel attention-free or highly simplified architectural components that can further reduce computational overhead without sacrificing performance.

4 Self-Supervised Learning Paradigms for Vision Transformers

4.1 Masked Image Modeling (MIM) for Representation Learning

Masked Image Modeling (MIM) has emerged as a highly effective self-supervised learning paradigm for Vision Transformers (ViTs), drawing direct inspiration from the success of BERT in Natural Language Processing (NLP). This approach trains a ViT to learn rich visual representations by reconstructing masked portions of an image, effectively predicting missing pixels or discrete visual tokens. By forcing the model to infer global context from partial observations, MIM significantly reduces the reliance on extensive human annotations, enabling ViTs to leverage vast amounts of unlabeled data for pre-training. A key distinction often drawn is that MIM's generative reconstruction task encourages learning rich, localized features, whereas contrastive methods (discussed in Section 4.2) tend to yield more globally semantic, linearly separable features.

One of the pioneering works in this domain is **?**, which introduced BEiT, a BERT-like pre-training framework for image Transformers. BEiT tokenizes images into discrete visual tokens using a discrete Variational AutoEncoder (dVAE) and then trains a standard ViT encoder to predict these tokens for randomly masked image patches. This method demonstrated that a masked language modeling objective could be effectively transferred to the visual domain, yielding strong representations for various downstream tasks. The discrete token prediction objective forces the model to operate at a higher level of semantic abstraction, potentially leading to more semantically meaningful feature learning by abstracting away low-level pixel noise.

Building upon this foundation, **?** presented Masked Autoencoders (MAE), a highly scalable and efficient MIM approach. MAE trains a Vision Transformer encoder to reconstruct the original raw pixel values of masked image patches. A key innovation of MAE is its asymmetric encoder-decoder architecture: only the visible patches are fed into the encoder, while a lightweight decoder reconstructs the missing pixels from the encoder's output and the masked token embeddings. This design, coupled with a remarkably high masking ratio (e.g., 75

Beyond these foundational works, the MIM landscape has diversified. SimMIM ? demonstrated that a simple linear head for pixel reconstruction could be highly effective, simplifying the decoder design even further. MaskFeat ? explored predicting hand-crafted features like Histograms of Oriented Gradients (HOG) instead of raw pixels, suggesting that the reconstruction target itself can be varied to influence learned representations.

A significant challenge for MIM, particularly MAE, has been its application to hierarchical Vision Transformers (e.g., Swin Transformer, PVT) which incorporate local window-based attention. The original MAE design, relying on the global attention of plain ViTs to handle randomly masked sequences, struggled with the local inductive biases of hierarchical architectures. To address this, HiViT ? proposed a new design for hierarchical ViTs that maintains efficiency and performance in MIM by modifying Swin Transformer to make mask-units serializable. Similarly, Uniform Masking (UM-MAE) ? enabled MAE pre-training for pyramid-based ViTs with locality. UM-MAE introduces a Uniform Sampling strategy that selects one random patch from each 2x2 grid, combined with Secondary Masking, to preserve equivalent elements across local windows, significantly improving pre-training efficiency and transferability for these architectures.

The effectiveness of MIM, particularly MAE, lies in its ability to learn robust and transferable representations. For instance, pre-trained MAE backbones have been shown to be highly effective for downstream tasks like object detection. ? demonstrated that plain, non-hierarchical ViT backbones pre-trained with MAE could achieve competitive results in object detection (ViTDet) with minimal architectural adaptations, even outperforming methods based on hierarchical backbones. Similarly, ? leveraged MAE pre-training to simplify hierarchical ViT architectures, showing that a strong visual pretext task can allow for the removal of many "bells-and-whistles" without sacrificing accuracy, resulting in faster and more efficient models like Hiera. This highlights MIM's capacity to imbue even simpler ViT designs with powerful representational capabilities. Furthermore, MIM can serve as a powerful auxiliary task; MAT-VIT ? uses an MAE-based self-supervised auxiliary task to improve medical image classification, effectively leveraging both unlabeled and labeled data. In the context of deep reinforcement learning from pixels, reconstruction-based ViT methods, including MIM, have been found to significantly outperform ViT contrastive-learning approaches, suggesting their utility in learning rich visual states for control tasks **?**.

While MIM has proven highly effective in self-supervised pre-training, the computational cost associated with training these large models, even with efficient designs like MAE's asymmetric architecture, remains a significant consideration. The optimal masking strategy, including masking ratio and pattern, continues to be an active area of research to maximize learning efficiency and representation quality. For example, BEiT v2 explores more sophisticated block-wise masking schemes to encourage learning richer semantics. Future work is investigating more adaptive masking techniques or hybrid approaches that combine MIM's generative loss with a discriminative loss on the visible tokens to capture both local detail and high-level semantics. The trade-off between pixel-level and discrete token-level reconstruction also remains a topic of investigation, as each approach offers distinct advantages in terms of simplicity, efficiency, and the semantic richness of the learned features.

4.2 Self-Distillation and Contrastive Learning without Labels

The inherent data hunger of Vision Transformers (ViTs) for pre-training with vast, labeled datasets has spurred significant research into self-supervised learning (SSL) paradigms. These approaches enable ViTs to acquire robust and semantically meaningful representations from unlabeled data, thereby mitigating the bottleneck of costly human annotation. Within SSL, two prominent and often intertwined methodologies—contrastive learning and self-distillation—have demonstrated exceptional efficacy, revealing profound emergent properties in ViT features that underscore their intrinsic capacity for visual understanding.

Contrastive learning, a foundational SSL paradigm, operates by maximizing the agreement between different augmented views of the same image (positive pairs) while simultaneously pushing apart representations of different images (negative pairs) in the embedding space. Early adaptations for ViTs, such as MoCo-v3 **?**, successfully integrated the momentum encoder concept with contrastive objectives, demonstrating that ViTs could learn powerful representations competitive with CNNs on ImageNet. Similarly, frameworks inspired by SimCLR **?** were adapted, emphasizing the importance of large batch sizes or memory banks for effective negative sampling. While effective, the reliance on explicit negative pairs and the associated computational overhead or architectural complexities (e.g., large memory banks or distributed training for large batches) presented practical challenges for scaling and maintaining training stability, motivating the search for alternative non-contrastive approaches.

In response to these complexities, non-contrastive self-supervised methods emerged, circumventing the explicit need for negative pairs altogether. A pioneering work in this category is BYOL (Bootstrap Your Own Latent) **?**, which demonstrated that high-quality representations could be learned by simply predicting the representation of one augmented view from another. BYOL employs a momentum encoder for the teacher network and a crucial predictor head on the student branch. The combination of the predictor network and a stop-gradient operation applied to the teacher's output was instrumental in preventing representational collapse, ensuring the student learned meaningful, non-trivial features. Following BYOL, SimSiam **?** further simplified non-contrastive learning by showing that even without a momentum encoder, a stop-gradient operation alone, when applied to one branch of a Siamese network, could effectively prevent collapse, making the training process more straightforward and efficient.

Building upon these non-contrastive principles, DINO (Self-Distillation with No Labels) **?** emerged as a seminal work specifically tailored for ViTs. DINO employs a student-teacher architecture where a student network is trained to match the output distribution of a teacher network for different augmented views of the same image. Crucially, the teacher network's weights are an exponential moving average of the student's weights (a momentum encoder), providing a stable yet evolving target. To prevent representational collapse, a common challenge in non-contrastive methods, DINO incorporates techniques like centering and sharpening the output distributions, which effectively regularize the learning process. This self-distillation encourages the student to learn features invariant to various augmentations and to capture deep semantic information. A remarkable emergent property observed in ViTs trained with DINO is their ability to perform object segmentation without any explicit supervision; the attention maps of the self-attention layers spontaneously highlight object boundaries and coherent semantic regions. This capability illustrates that ViTs can intrinsically learn to parse visual scenes into meaningful entities through unsupervised means, significantly reducing the need for human annotation for tasks like segmentation.

Beyond foundational self-distillation, further refinements and applications have broadened its impact. EsViT **?** extended self-distillation by exploring different view generation strategies and leveraging attention-based distillation to enhance the learning process for ViTs. Furthermore, the concept of distilling knowledge from a self-supervised teacher to a smaller student has gained traction for efficiency and deployment on resource-constrained devices. AttnDistill **?** specifically addresses self-supervised knowledge distillation for ViTs by directly distilling information from the crucial attention mechanism of a teacher to a student. This method demonstrates that by guiding the student with the teacher's attention, the performance gap between models can be significantly narrowed, enabling the deployment of high-performing ViTs on memory and compute-constrained devices, even down to tiny ViT models. Similarly, distillation principles are leveraged in contexts like ViT quantization, where a teacher-student framework can rectify issues such as attention distortion in binarized ViTs, improving their performance on resource-limited devices **?**. While this application of distillation focuses on model compression and enhancing the utility of already learned representations rather than initial self-supervised representation learning, it underscores the versatility of the teacher-student paradigm in improving ViT practicality.

A significant evolution in this space involves the convergence of self-distillation with masked image modeling (MIM), bridging concepts from this subsection with those discussed in Section 4.1. iBOT (Image BERT Pre-training with Online Tokenizer) **?** exemplifies this hybrid approach. iBOT combines the self-distillation framework of DINO with masked image modeling, where a student ViT is trained to predict the output of a teacher ViT for masked image patches. This dual objective allows iBOT to leverage the strengths of both paradigms: the global context learning from MIM and the robust feature learning from self-distillation, leading to highly effective visual representations that exhibit both fine-grained detail and strong semantic understanding. Such self-supervised techniques are also being adapted for domain-specific applications, such as medical image analysis, where models like MAT-VIT explore MAE-based self-supervised auxiliary tasks within a Vision Transformer framework to leverage abundant unlabeled medical images, improving classification performance in data-scarce scenarios **?**.

In summary, self-distillation and contrastive learning paradigms have been instrumental in unlocking the full potential of Vision Transformers by enabling them to learn from vast amounts of unlabeled data. While contrastive methods like MoCo-v3 provided early successes, the subsequent development of non-contrastive self-distillation methods such as BYOL, SimSiam, and DINO simplified the training process by avoiding explicit negative pairs, often leading to more stable learning and revealing remarkable emergent properties like unsupervised object segmentation. The continuous refinement of distillation techniques for efficiency and the development of hybrid approaches like iBOT, which integrate self-distillation with masked image modeling, further enhance the representational capacity and practical applicability of ViTs. These advancements not only drastically reduce the reliance on extensive human annotation but also highlight the profound inherent representational capacity of Transformer architectures for visual data, paving the way for more data-efficient, generalizable, and universally applicable vision models.

4.3 Scaling Self-Supervised ViTs to Foundation Models

The landscape of computer vision is undergoing a profound transformation with the emergence of "Vision Foundation Models," a paradigm driven by the aggressive scaling of self-supervised Vision Transformers (ViTs) to unprecedented sizes. Drawing inspiration from large language models, a foundation model is broadly defined as a large model pre-trained on a vast quantity of diverse data, designed to be adaptable to a wide range of downstream tasks ?. In vision, this translates to developing highly robust, general-purpose visual backbones capable of transferring effectively across an extremely broad spectrum of tasks with minimal fine-tuning, thereby significantly reducing the need for task-specific model development and accelerating progress towards universal visual intelligence.

The initial success of ViTs ? in image recognition, despite their significant data hunger, underscored the potential of Transformer architectures for visual data. This limitation spurred extensive research into self-supervised learning (SSL) techniques, which became indispensable for enabling ViTs to scale by leveraging vast quantities of unlabeled data. As discussed in previous subsections, Masked Image Modeling (MIM) ?? and self-distillation methods like DINO ? proved instrumental in allowing ViTs to learn rich, generalizable representations without explicit human annotations. Building on these foundational SSL advancements, the field has now entered an era of aggressive scaling, pushing model capacities and data volumes to new extremes.

A prime illustration of this scaling trend is the work by ?, which systematically explored scaling Vision Transformers to up to a billion parameters. This research demonstrated that with sufficient model capacity and effective self-supervised pre-training, particularly MIM, ViTs can learn exceptionally powerful and transferable representations. Crucially, this scaling revealed emergent properties and improved scaling laws, where performance gains continue with increasing model size and data, setting new benchmarks for general-purpose visual backbones. These models are typically pre-trained on massive, diverse, and often curated web-scale datasets, far surpassing the scale of traditional benchmarks like ImageNet. Complementing this, ? pushed the boundaries of self-supervised learning further, focusing on learning highly robust and generalizable visual features without any supervision. DINOv2's success lies in its ability to produce features that are readily usable for a wide array of downstream tasks, significantly reducing the need for task-specific labeled data and extensive fine-tuning. This approach exemplifies how refined SSL strategies, combined with large-scale pre-training, enable models to learn intrinsic visual understanding, often exhibiting remarkable emergent properties such as object segmentation without explicit supervision. Furthermore, research like Hiera ? demonstrates that strong visual pretext tasks, such as MAE, can simplify hierarchical ViT designs, allowing for the removal of architectural "bells-and-whistles" while maintaining or improving accuracy and efficiency post-pre-training, suggesting the power often lies more in the robust SSL pre-training strategy than in architectural complexity alone.

However, this aggressive scaling is not without significant challenges and costs. The engineering hurdles associated with training such colossal models, including distributed computing, memory optimization, and stable optimization techniques, are substantial. Moreover, the exorbitant computational costs and environmental impact of training billion-parameter models raise concerns about accessibility and sustainability, creating a potential barrier to entry for academic research. To mitigate these issues, research into more efficient architectures and compression techniques is vital. For instance, approaches like UFO-ViT ? propose linear complexity self-attention mechanisms to alleviate the quadratic computational burden, while DeepViT ? addresses attention collapse in deeper models through "Re-attention" to enable more effective scaling. Furthermore, structured pruning methods like GOHSP ? and multi-dimensional compression paradigms ? aim to reduce the model size and computational cost of ViTs for practical deployment without significant accuracy loss, making the benefits of foundation models more accessible.

The hallmark of these Vision Foundation Models is their exceptional transferability and generalization capabilities. A systematic investigation by ? revealed consistent advantages of Transformer-based backbones over ConvNets in transfer learning across a majority of downstream tasks, including fine-grained classification, scene recognition, and open-domain classification. This inherent transferability is amplified at the foundation model scale. For instance, Prompt Vision Transformers ? leverage prompt learning to embed domain-specific knowledge, enabling ViTs to generalize effectively to unseen domains. Similarly, Transferable Vision Transformers (TVT) ? demonstrate superior generalization ability and can be further optimized for unsupervised domain adaptation by focusing on transferable and discriminative features through specialized modules. The concept of attention distillation ? further extends the utility of these large foundation models by enabling the transfer of learned knowledge, particularly from attention mechanisms, to smaller student ViTs, making the benefits of scale accessible to more resource-constrained deployments.

The impact of these Vision Foundation Models is transformative across various applications, primarily by providing highly robust, general-purpose visual backbones that significantly reduce the need for developing task-specific models from scratch. Their ability to learn rich, semantically meaningful features from vast, diverse data makes them uniquely suited for domains with distinct data characteristics or limited labeled data. For example, in medical imaging, models leveraging foundation backbones can achieve high accuracy in tasks like white blood cell classification **?** or even medical image classification with MAE-based auxiliary tasks **?**, often with minimal task-specific data and fine-tuning, demonstrating their strong transferability to data-scarce domains. In autonomous systems, these powerful backbones can be adapted for critical tasks such as traversable area detection **?**, leveraging their robust feature learning for complex environmental understanding. Their capacity to discern subtle visual artifacts and latent data distributions also makes them invaluable for challenging adversarial applications like deepfake detection **?**. Furthermore, their general-purpose nature extends to tasks like camouflaged and salient object detection, where a simple ViT-based encoder-decoder can yield competitive results across both distinct tasks **?**, and to remote sensing image classification, where quantitative regularization can enhance ViT performance even with limited training samples **?**.

In conclusion, the scaling of self-supervised ViTs to foundation models represents a monumental achievement in computer vision. This trend, driven by advanced SSL techniques like MIM and self-distillation, combined with massive model capacities and diverse unlabeled datasets, is creating highly robust and general-purpose visual backbones. These models offer unprecedented generalization and transferability, significantly streamlining the development of high-performance vision systems across a multitude of tasks. While offering immense potential, future research will continue to focus on further enhancing their efficiency, exploring novel architectural refinements, and critically addressing the ethical considerations inherent in such broadly applicable and powerful AI systems, particularly concerning bias amplification from vast, uncurated web-scale data.

5 Hybrid Architectures and Beyond Self-Attention

5.1 Integrating Convolutional Inductive Biases into Transformers

While Vision Transformers (ViTs) have demonstrated impressive capabilities in capturing global dependencies, their initial lack of inherent inductive biases, such as translation equivariance and locality, often necessitates vast training data and can lead to sub-optimal performance on smaller datasets or for fine-grained local tasks ??. To address these limitations, a significant research direction has focused on hybrid architectures that explicitly embed convolutional layers or integrate convolutional inductive biases within the Transformer framework, aiming to synergistically combine the local feature extraction strengths of Convolutional Neural Networks (CNNs) with the global reasoning capabilities of Transformers. This synergistic integration often leads to improved efficiency, better performance, and enhanced robustness, particularly on smaller datasets where pure ViTs might struggle due to their lack of inherent inductive priors. These hybrid designs represent a pragmatic effort to leverage the best attributes of both paradigms, moving beyond a 'CNN vs. Transformer' dichotomy towards more powerful combined architectures.

Early efforts to imbue Transformers with locality often explored architectural modifications without direct convolutions. For instance, the Swin Transformer ? introduced a hierarchical architecture that limits self-attention computation to non-overlapping local windows, with shifted windowing enabling cross-window connections. While this design provides a form of local inductive bias and improves efficiency for dense prediction tasks, it achieves locality through windowing rather than explicit convolutional operations. In contrast, other foundational hybrid models directly integrated convolutions. CoaT (Co-scale Conv-Attentional Image Transformers) ? integrates convolution and attention at co-scales, allowing for a dynamic interplay between local and global features. CvT (Introducing Convolutions to Vision Transformers) ? explicitly embeds convolutions into the Transformer architecture by replacing the linear patch embedding with a strided convolution and using depthwise-separable convolutions for the key, query, and value projection layers within the self-attention mechanism. This approach leverages convolutions for efficient tokenization and local feature aggregation directly within the attention process. ConViT (Improving Vision Transformers with Soft Convolutional Inductive Biases) **?** introduces "soft convolutional biases" by initializing the attention mechanism to prioritize local neighborhoods, gradually expanding its receptive field in deeper layers, thereby mimicking the inductive bias of convolutions without hard-coding them. LeViT (a Vision Transformer in ConvNet's Clothing for Faster Inference) **?** further blurs the lines by optimizing ViTs for speed through the incorporation of attention bias and convolution-like structures, demonstrating that careful architectural design can yield benefits traditionally associated with CNNs.

The integration strategies for convolutional inductive biases can be broadly categorized into several architectural patterns. One common approach involves embedding convolutional layers directly within Transformer blocks or using them for initial feature extraction. For example, some models utilize convolutions for the initial patch embedding, similar to CvT, or integrate them into feed-forward networks (FFNs) to enhance local feature mixing. Another prominent pattern involves parallel CNN and ViT streams, where each branch specializes in different aspects of feature extraction before their outputs are fused. This allows the CNN branch to capture fine-grained local details and translation equivariance, while the Transformer branch focuses on global contextual relationships. For instance, CTNet **?** proposes a joint framework with separate CNN and ViT streams to extract local structural and global semantic features, respectively, for remote sensing scene classification, fusing them for comprehensive understanding. Similarly, GLNS **?** for high-resolution SAR image classification utilizes a lightweight CNN and a compact ViT in parallel, fusing their outputs to leverage complementary local and global features.

Beyond full architectural integration, convolutional inductive biases can also be introduced through lightweight adaptation modules or specialized components. Convpass **?** proposes "Convolutional Bypasses" as plug-and-play adaptation modules for pre-trained

ViTs. These bypasses inject convolutional layers, benefiting from their hard-coded inductive bias, particularly in low-data regimes, without altering the original ViT parameters. This method offers a parameter-efficient way to enhance ViTs with local priors. For pixel-level tasks like camouflaged and salient object detection, SENet **?** incorporates a "local information capture module" within its ViT-based encoder-decoder structure. This module is specifically designed to compensate for the limitations of patch-level attention in capturing fine-grained local details, which are crucial for precise pixel-level predictions.

These hybrid strategies have proven particularly effective in specialized domains where data scarcity or the need for robust local features is paramount. In remote sensing, for example, the fusion of local and global information is critical for tasks like land use and land cover classification or hyperspectral image analysis. Several works adopt parallel or integrated convolutional modules to enhance local feature extraction and reduce reliance on massive pre-training. P2FEViT **?** introduces a plug-and-play CNN feature embedding into ViT, allowing synchronous capture and fusion of global context with local multimodal information. ExViT **?** extends conventional ViTs for multimodal land use and land cover classification by processing image patches with parallel branches of position-shared ViTs augmented with separable convolution modules, fusing their embeddings via cross-modality attention. GSC-ViT **?** for hyperspectral image classification employs a groupwise separable convolution (GSC) module to efficiently capture local spectral-spatial information and a groupwise separable multihead self-attention (GSSA) module for both local and global spatial feature extraction, significantly reducing parameters. Furthermore, the lightweight dual-branch Swin Transformer (LDBST) **?** combines a ViT branch with a CNN branch, integrating a Conv-MLP structure into the ViT branch to enhance connections with neighboring windows, showcasing the versatility of these hybrid designs.

In conclusion, the integration of convolutional inductive biases into Transformers represents a pragmatic and effective strategy to overcome the limitations of pure ViTs, particularly concerning data efficiency, local feature extraction, and robustness on diverse datasets. These hybrid designs, ranging from embedding convolutional layers directly within Transformer blocks and attention mechanisms to employing parallel CNN-ViT streams and lightweight convolutional adaptation modules, successfully combine the local feature extraction strengths of CNNs with the global reasoning capabilities of Transformers. This synergistic approach has consistently led to improved performance and enhanced robustness across various vision tasks, especially in scenarios with limited data or requiring fine-grained local understanding. Future research will likely continue to explore more sophisticated and dynamic integration strategies, a deeper theoretical understanding of their combined inductive biases, and the optimal balancing of computational cost with performance gains across diverse applications, moving beyond a simplistic 'CNN vs. Transformer' dichotomy towards more powerful and versatile combined architectures.

5.2 Rethinking Token Mixing: Alternatives to Self-Attention

While self-attention mechanisms have been instrumental in the success of Vision Transformers (ViTs) by providing global receptive fields, their quadratic computational complexity with respect to sequence length and their initial lack of inherent inductive biases have stimulated extensive research into simpler, more computationally efficient alternatives for global token mixing. This line of inquiry, often unified under the "MetaFormer" architectural paradigm **?**, provocatively suggests that the overall design—comprising a token mixer followed by a feed-forward network (FFN)—can be highly effective even with substantially simpler mixing operations, challenging the notion that complex self-attention is indispensable for achieving state-of-the-art performance in vision.

Among the earliest and most influential works to question the necessity of self-attention were those proposing MLP-based mixers. The *MLP-Mixer* **?** demonstrated that a pure multi-layer perceptron (MLP) architecture, without any explicit attention mechanism, could achieve competitive performance on image classification tasks. This model segments an image into patches, then applies channel-mixing MLPs and token-mixing MLPs alternately. The token-mixing MLP operates across all patches, effectively providing a global receptive field through fully connected layers, albeit with a high computational cost if not carefully implemented. Following this, *gMLP* **?** further refined the MLP-only approach by introducing a Spatial Gating Unit (SGU) that adaptively controls information flow across spatial locations. Both MLP-Mixer and gMLP were foundational in establishing that the general Transformer-like block structure, rather than the specific self-attention operation, was a key driver of performance, paving the way for further exploration of non-attention mixers.

Building on the surprising efficacy of simpler operations, subsequent research explored even more basic alternatives. The *PoolFormer* architecture **?** famously demonstrated that simple pooling operations, such as average or max pooling, could effectively replace self-attention layers within a MetaFormer block while achieving competitive performance. This finding underscored that a parameter-free, local aggregation operation, when stacked within the MetaFormer framework, could surprisingly capture sufficient global context for robust image representation. In a different vein, Global Filter Networks (GFNet) **?** proposed replacing self-attention with global filters applied in the frequency domain, specifically using 2D Fourier transforms. This approach offers a computationally efficient way to achieve global receptive fields by treating tokens as a signal and performing global mixing via frequency-domain multiplication, thereby avoiding the quadratic complexity of self-attention while providing a distinct mechanism for global interaction.

Beyond parameter-free pooling or frequency-domain filters, other works have explored structured or localized alternatives that retain some global interaction without full self-attention. The Vision Permutator (ViP) **?** introduced a permutable self-attention mechanism, which simplifies the attention process by performing attention along different axes (height and width) sequentially, rather than across all tokens simultaneously. While still using an attention-like mechanism, its structured permutation significantly reduces computational cost and provides a form of global information exchange with linear complexity. Focal Modulation Networks (FocalNet) **?** proposed focal modulation, a novel spatial modulation mechanism that captures both fine-grained local context and long-range dependencies in a hierarchical manner. Unlike self-attention, focal modulation explicitly models interactions at different scales through a series of modulators, providing an efficient way to aggregate information across varying receptive fields without the explicit pairwise token comparison of attention.

The exploration of alternatives continues with emerging paradigms, such as state-space models (SSMs). Recent work like *MambaVision* **?** proposes a hybrid Mamba-Transformer backbone, where the Mamba architecture, based on structured state-space sequences, is adapted for visual feature modeling. MambaVision demonstrates the feasibility of integrating SSMs as a distinct, non-attention-based token mixer, particularly in its initial layers, to efficiently capture long-range spatial dependencies. This represents a significant new direction, suggesting that sequence modeling mechanisms beyond attention can be effectively adapted for vision, potentially offering superior efficiency and inductive biases for certain tasks.

In conclusion, the research on rethinking token mixing fundamentally challenges the premise that self-attention is an indispensable component of high-performing vision models. Works like **??????** have collectively demonstrated that simpler, more efficient operations—ranging from basic MLPs and pooling to frequency-domain filters, structured permutations, and focal modulation—can effectively replace self-attention within the robust MetaFormer paradigm. These models achieve competitive performance while significantly reducing computational overhead, pushing the boundaries of efficiency and architectural simplicity. The emergence of state-space models further diversifies this landscape, indicating a continuous quest for novel token mixing strategies. This shift suggests that the general block-based, token-processing architecture of Transformers is a robust design, capable of leveraging diverse mixing strategies. The ongoing challenge lies in balancing the theoretical expressiveness and adaptive capacity of full self-attention with the practical demands for computational efficiency, hardware friendliness, and inherent inductive biases offered by these increasingly varied alternative token mixing approaches.

5.3 Modernizing CNNs with Vision Transformer Design Principles

The emergence of Vision Transformers (ViTs) initially presented a significant challenge to the long-standing dominance of Convolutional Neural Networks (CNNs) in computer vision, primarily due to their superior global context modeling capabilities. However, this perceived architectural dichotomy has evolved into a profound convergence, where insights and design principles from ViTs are now being strategically applied to modernize and revitalize traditional CNNs. This trend effectively blurs the lines, fostering a new generation of powerful convolutional networks that benefit from lessons learned in the Transformer era, thereby challenging the absolute necessity of self-attention for state-of-the-art performance.

A crucial precursor to this modernization was the *Swin Transformer* **?**. While fundamentally a Transformer, Swin's hierarchical architecture, employing shifted window-based attention, introduced a more CNN-like inductive bias by processing local regions and progressively building global interactions. This design demonstrated that Transformers could achieve efficiency and scalability comparable to CNNs, inspiring researchers to investigate whether the performance gains of ViTs stemmed primarily from their attention mechanism or from their broader architectural structure and training methodologies. The Swin Transformer thus served as a blueprint, prompting a re-evaluation of CNN design principles.

The most seminal work exemplifying this modernization is the *ConvNeXt* architecture **?**. The authors embarked on a systematic 'metamorphosis' of a standard ResNet, incrementally incorporating design choices prevalent in state-of-the-art Vision Transformers, particularly those observed in Swin Transformer. This meticulous process involved several key modifications:

- **Macro Design:** The overall stage-wise downsampling and channel ratios were adjusted to align with those of Swin Transformer, moving from a large stem to a patchify stem and adopting similar block repetition patterns.

- **ResNeXt-ification:** Efficient depthwise convolutions and inverted bottleneck designs, characteristic of efficient Transformer blocks, were integrated. This allowed for increased channel capacity within blocks while maintaining computational efficiency.

- **Larger Kernel Sizes:** The kernel size of depthwise convolutions was significantly increased (e.g., from 3x3 to 7x7). This allowed convolutional layers to capture a larger receptive field, mimicking the broader context captured by attention mechanisms without incurring their quadratic computational cost.

- **Layer Normalization:** Batch Normalization layers were replaced with Layer Normalization, a standard practice in Transformers that stabilizes training, especially with smaller batch sizes, and improves generalization.

- **Activation Functions:** The ReLU activation function was substituted with GELU, another activation function widely used in Transformers, contributing to improved performance.

- **Downsampling and Pooling:** Adjustments were made to downsampling layers, and the final classification head adopted a single global average pooling layer, streamlining the network's output stage.

By systematically applying these ViT-inspired design principles, ConvNeXt demonstrated that a pure convolutional network could achieve competitive or even superior performance to leading Transformers across various benchmarks, including ImageNet classification, COCO object detection, and ADE20K semantic segmentation. This work critically challenged the prevailing notion that self-attention was indispensable for achieving state-of-the-art results, instead highlighting the profound importance of macro-architectural design and training strategies.

Beyond architectural modifications, the influence of Vision Transformers extends to training methodologies. Transformer-inspired self-supervised learning strategies, particularly Masked Image Modeling (MIM), have proven highly effective for pre-training convolutional networks, further blurring the methodological distinctions. *ConvMAE* **?** directly applied the Masked Autoencoder (MAE) paradigm, originally developed for ViTs, to pure CNN architectures. It demonstrated that CNNs, like Transformers, could learn powerful visual representations by reconstructing masked image patches, achieving significant performance gains on downstream tasks. This showed that the effectiveness of MIM was not exclusive to attention-based models but could also benefit architectures with strong inductive biases like convolutions. Building on this, *ConvNeXt V2* **?** further showcased this synergy by pre-training the modernized ConvNeXt architecture using a fully convolutional masked autoencoder (FCMAE). This approach not only boosted ConvNeXt's performance but also solidified the argument that the benefits of Transformer-era training techniques are transferable to well-designed CNNs, leading to more robust and data-efficient convolutional backbones.

The revitalization of ConvNets through ViT design principles has found application in various domains. For instance, *GenConViT* **?** for deepfake detection leverages both ConvNeXt and Swin Transformer models for robust feature extraction, illustrating how modernized CNNs can synergistically contribute to complex vision tasks alongside hierarchical ViTs. This ongoing convergence suggests that the future of powerful vision backbones will likely be a synthesis of the best elements from both worlds, leading to more versatile and robust models that effectively balance global context modeling with local inductive biases. The optimal balance and specific architectural configurations remain an active area of research, continually pushing the boundaries of what pure convolutional networks can achieve when inspired by Transformer innovations.

5.4 Efficient and Lightweight Hybrid Designs for Deployment

6 Applications and Domain-Specific Adaptations of Visual Transformers

6.1 Object Detection and Semantic Segmentation

Object detection and semantic segmentation are fundamental dense prediction tasks in computer vision, demanding not only accurate classification but also precise localization and pixel-level understanding of objects within an image. While Convolutional Neural Networks (CNNs) historically dominated these areas, the advent of Vision Transformers (ViTs) has introduced powerful new paradigms, successfully adapting the global context modeling capabilities of Transformers to these intricate visual challenges.

A significant breakthrough in object detection was the introduction of the Detection Transformer (DETR) ?. This pioneering work revolutionized the object detection pipeline by proposing an end-to-end approach that directly predicts a set of objects without the need for traditional components like Non-Maximum Suppression (NMS) or anchor boxes. DETR frames object detection as a set prediction problem, leveraging a Transformer encoder-decoder architecture to learn direct set predictions of bounding boxes and class labels, thereby simplifying the overall process. However, early pure ViT architectures, such as the original Vision Transformer ?, faced limitations when directly applied to dense prediction tasks due to their fixed-resolution inputs and lack of inherent hierarchical feature representation, which are crucial for capturing objects at various scales and for fine-grained pixel-level analysis.

To address these challenges, subsequent research focused on developing hierarchical Vision Transformers capable of generating multi-scale feature maps, essential for dense prediction. The Swin Transformer ? emerged as a powerful backbone, introducing a hierarchical architecture built upon shifted windows. This design allows for local self-attention within non-overlapping windows while enabling cross-window connections through shifted window partitioning, thereby efficiently capturing both local and global context and achieving state-of-the-art performance across various dense prediction benchmarks. Similarly, the Pyramid Vision Transformer (PVT) ? offered another effective solution by constructing a pure Transformer-based pyramid structure. PVT progressively reduces the resolution of feature maps and aggregates contextual information across different scales, making it a versatile backbone for tasks requiring multi-scale feature extraction, such as object detection and semantic segmentation, often surpassing traditional CNN-based methods in accuracy for complex scene analysis.

Further explorations into the architectural design for dense prediction have challenged some established conventions. For instance, ViTDet ? investigated the efficacy of plain, non-hierarchical Vision Transformer backbones for object detection. This work demonstrated that with minimal adaptations for fine-tuning, a simple feature pyramid constructed from a single-scale feature map, combined with window attention and limited cross-window propagation, could achieve competitive results on the COCO dataset. This suggests that the complex hierarchical designs, while effective, might not always be strictly indispensable. Pushing this simplification even further, ShiftViT **?** proposed an extremely simple alternative to the attention mechanism itself. By replacing attention layers with parameter-free shift operations that exchange channels between neighboring features, ShiftViT achieved performance on par with or even superior to the Swin Transformer in classification, detection, and segmentation, questioning the fundamental role of attention as the sole key to ViT's success in dense prediction.

Beyond pure Transformer designs, hybrid architectures have also proven effective in dense prediction tasks, combining the strengths of ViTs with those of CNNs or other mechanisms. For semantic segmentation, particularly in challenging environments, models integrating Vision Transformers with Multilayer Perceptrons (MLPs) and CNNs have shown promise. For example, **?** developed bilateral models for traversable area detection in autonomous vehicles, leveraging a hybrid approach to enhance prediction accuracy by capturing distant details more effectively while maintaining real-time operational capabilities. These models demonstrate how combining the global context understanding of ViTs with the local feature extraction and inductive biases of CNNs can lead to robust solutions for pixel-level tasks.

In conclusion, Vision Transformers have profoundly impacted object detection and semantic segmentation, moving from pioneering end-to-end models like DETR to sophisticated hierarchical backbones such as Swin Transformer and PVT that generate multiscale features crucial for dense prediction. While these advancements have significantly improved accuracy and simplified pipelines, ongoing research continues to explore architectural efficiencies, question the necessity of complex attention mechanisms, and investigate optimal hybrid designs. The field is still actively seeking to balance the computational cost and data efficiency of global attention with the need for precise local detail and the inductive biases traditionally offered by CNNs, paving the way for even more versatile and robust dense prediction models.

6.2 Medical Image Analysis and 3D Segmentation

Accurate and robust segmentation of medical images, particularly in three dimensions, is paramount for diagnosis, treatment planning, and surgical guidance. This task presents unique challenges due to the volumetric nature of the data, often limited availability of annotated datasets, and the critical need for fine-grained detail alongside broad anatomical context. While Convolutional Neural Networks (CNNs) have traditionally excelled in medical image analysis, their inherent limitations in capturing long-range dependencies across large fields of view have motivated the exploration of Vision Transformers (ViTs).

The initial success of Vision Transformers in general image recognition, as demonstrated by models like **?**, paved the way for their adaptation to medical imaging. However, pure Transformer architectures often require extensive datasets and can struggle with capturing fine-grained local details crucial for precise segmentation. This led to the development of hybrid CNN-ViT architectures, which strategically combine the strengths of both paradigms. A foundational step in this direction for medical imaging was *TransUNet* **?**, which integrated a Transformer encoder to capture global contextual information with a CNN decoder for precise localization, showcasing the benefits of this hybrid approach for 2D medical image segmentation.

Extending these concepts to the more complex domain of 3D medical imaging, where volumetric data demands efficient processing of spatial and contextual information, became a critical area of research. Early hybrid models for 3D segmentation, such as *SwinBTS* **?**, leveraged the hierarchical feature extraction capabilities of the 3D Swin Transformer as the encoder and decoder backbone within a U-Net-like structure. This approach aimed to address the CNN's weakness in modeling global and remote semantic information by employing the self-attention mechanism, while still utilizing convolutional operations for efficient upsampling and downsampling. *SwinBTS* demonstrated improved performance in 3D multimodal brain tumor segmentation by focusing on extracting contextual data through the Transformer while enhancing detail feature extraction.

Building upon these advancements, more sophisticated hybrid architectures emerged to further refine the integration of CNNs and ViTs. *Swin Unet3D* **?** represents a significant step forward by proposing a novel parallel feature extraction mechanism within each stage of its 3D U-Net-like encoder-decoder. Unlike prior models that might use Transformers as a bottleneck or sequentially, *Swin Unet3D* employs parallel 3D Swin Transformer Blocks and 3D Convolutional Blocks. This allows the network to simultaneously learn both long-range global dependencies (via the Transformer) and short-range local details (via the CNN) throughout the entire network, addressing the limitations of both pure CNNs (limited receptive fields) and pure ViTs (high parameters, poor local detail learning with limited data). The outputs from these parallel branches are then fused, notably through multiplication, to combine their distinct representations effectively. This parallel processing strategy for complementary features echoes architectural innovations seen in other domains, such as the two-pathway approach in *TP-VIT* **?** for video action recognition, which processes different types of information (spatial and temporal) in parallel.

The integration of 3D Swin Transformer Blocks, which utilize a 3D windowed and shifted-window multi-head self-attention mechanism, is crucial for efficiently processing volumetric data in *Swin Unet3D*. This design, coupled with depth-wise separable convolutions for local feature learning, provides a better balance between segmentation accuracy and model parameters, which is vital for clinical deployment. While these hybrid models significantly advance segmentation accuracy, the broader field of Vision Transformers continues to explore efficiency improvements. For instance, *ViT-Shift* **?**, though focused on video action recognition, demonstrates how modules like the Temporal Shift Module can be integrated into ViTs to reduce computational costs while preserving performance, a consideration that remains highly relevant for resource-intensive 3D medical imaging tasks.

In conclusion, the evolution of medical image analysis for 3D segmentation has seen a clear intellectual trajectory from pure CNNs to sophisticated hybrid CNN-ViT architectures. These models, exemplified by *TransUNet* **?**, *SwinBTS* **?**, and particularly *Swin Unet3D* **?**, effectively combine the local inductive biases of CNNs with the global context modeling capabilities of Transformers. Despite these advancements, challenges persist in optimizing the fusion mechanisms, reducing computational overhead for real-time applications, and developing robust self-supervised learning strategies to mitigate the impact of limited annotated medical datasets. Future research will likely focus on further refining these hybrid designs and exploring more data-efficient training paradigms to achieve even more accurate and clinically viable 3D medical image segmentation solutions.

6.3 Remote Sensing and Environmental Monitoring

Remote sensing is indispensable for environmental monitoring, providing critical data for observing Earth's surface across vast geographical scales. Vision Transformers (ViTs) have emerged as a transformative technology in this domain, offering robust solutions for analyzing complex satellite and radar imagery, particularly for tasks like land use and land cover (LULC) classification and specialized object detection under challenging conditions. Their inherent ability to capture global contextual information across large geographical areas, coupled with their capacity to model long-range dependencies, makes them particularly well-suited for these tasks, often surpassing traditional convolutional neural network (CNN) methods in accuracy and robustness.

One significant application area is **object detection in Synthetic Aperture Radar (SAR) imagery**, which presents unique challenges due to speckle noise, strong scattering, and multi-scale objects against complex backgrounds. Traditional methods often struggle to maintain performance in such cluttered environments. To address this, **?** introduced ST-YOLOA, a hybrid model that integrates the global context modeling capabilities of the Swin Transformer with the efficient YOLOX framework for real-time SAR ship detection. This approach leverages a novel STCNet backbone, an enhanced PANet with SE and CBAM attention for multi-scale feature fusion, and a decoupled detection head, demonstrating a notable accuracy improvement (e.g., 4.83

In the realm of **Land Use and Land Cover (LULC) classification and semantic segmentation**, ViTs are particularly adept at processing multimodal remote sensing data and capturing intricate spatial patterns. The Swin Transformer, with its hierarchical structure and shifted window attention, has proven highly effective for general

LULC tasks, outperforming state-of-the-art CNNs on datasets like EuroSat and NWPU-RESISC45 ?. Building on this, ? proposed the Class-Guided Swin Transformer (CG-Swin) for semantic segmentation of remote sensing images, leveraging the Swin Transformer as an encoder and designing a class-guided Transformer block for the decoder. This architecture effectively captures long-range dependencies crucial for accurate pixel-level classification in complex scenes, achieving significant breakthroughs on datasets such as ISPRS Vaihingen and Potsdam. For multimodal LULC classification, which often involves combining hyperspectral (HS), LiDAR, and SAR data, ? presented ExViT, an Extended Vision Transformer framework. ExViT utilizes parallel ViT branches for each modality and a cross-modality attention (CMA) module to facilitate information exchange, demonstrating superior discriminative ability and classification performance compared to traditional methods.

Addressing the unique characteristics of **hyperspectral image classification (HSIC)**, ? introduced a multi-stage Vision Transformer with stacked samples. This approach tackles the challenge of ViTs sometimes ignoring local characteristics while focusing on global information, and mitigates their data hunger through an innovative data augmentation method. Similarly, for **Polarimetric SAR (PolSAR) land cover classification**, where labeled samples are often scarce, ? proposed a ViT-based method. This model leverages the ViT's powerful global feature representation and employs a Masked Autoencoder (MAE) for pre-training with unlabeled data, effectively overcoming the data scarcity limitation and demonstrating superior performance on datasets like Flevoland and Hainan. The integration of CNN-like inductive biases with ViTs also proves beneficial for remote sensing image classification (RSIC). ? developed P2FEViT, a Plug-and-Play CNN Feature Embedded Hybrid Vision Transformer. P2FEViT integrates CNN features into the ViT architecture to synchronously capture and fuse global context with discriminative local feature representations, thereby improving classification capability, reducing ViT's dependence on large-scale pre-training data, and accelerating convergence.

Beyond LULC and SAR, ViTs are also making strides in **agricultural monitoring**, where fine-grained detection and classification are essential. For instance, ? developed

SwinGD, a Swin Transformer-based model for robust grape bunch detection in complex vineyard environments, effectively handling irregularly shaped and dense objects. Similarly, ? proposed GNViT, an enhanced Vision Transformer model for groundnut pest classification, achieving near-perfect accuracy (99.95

In conclusion, Vision Transformers have rapidly become a cornerstone in remote sensing and environmental monitoring. Their unparalleled ability to model global contextual information, often enhanced by specialized attention mechanisms and integrated into efficient or hybrid frameworks, provides powerful solutions for complex tasks. From improving SAR ship detection in cluttered maritime environments to enabling highly accurate multimodal LULC classification and fine-grained agricultural monitoring, ViTs offer significant advancements in accuracy and robustness. Future research in this domain will likely focus on developing more efficient ViT architectures tailored for real-time processing on edge devices, enhancing their robustness against environmental noise and adversarial attacks, and exploring their integration into broader geospatial foundation models for more comprehensive and intelligent environmental insights.

6.4 Lightweight and Real-time Applications

The effective deployment of Vision Transformers (ViTs) in real-time applications and resource-constrained environments represents a critical frontier, demanding a delicate balance between model accuracy, computational efficiency, and minimal latency. The inherent complexity and substantial memory footprint of large ViTs often render them unsuitable for on-device processing in scenarios such as automated plant disease classification, radar-based human activity recognition (HAR), and mobile vision systems. This subsection focuses on how lightweight and efficient ViT models are specifically tailored and deployed to overcome these challenges, enabling advanced deep learning capabilities in practical, embedded contexts where computational resources are severely limited.

General-purpose lightweight ViT architectures provide foundational backbones for a wide array of mobile and edge applications. Models like MobileViT ?, which reinterpret transformers as convolutions, demonstrate superior accuracy with significantly fewer parameters than many traditional CNNs and ViTs. This efficiency makes them highly suitable for tasks requiring real-time inference on edge devices. For instance, the MobileViT-based Tracker (MVT) **?** leverages MobileViT as its core, achieving highly accurate and fast visual object tracking in real-time on resource-constrained hardware. The strength of such general-purpose models lies in their broad applicability, offering a robust starting point for various mobile vision tasks without extensive domain-specific customization, thereby extending advanced visual perception to everyday devices.

For highly specialized real-time applications, however, custom hybrid designs and optimized attention mechanisms are often paramount to meet stringent performance and efficiency requirements. Consider radar-based human activity recognition (HAR), a domain characterized by unique micro-Doppler data and the critical need for low-latency processing on embedded systems. Traditional ViTs are often computationally prohibitive in this context. To address this, huan202345b developed a Lightweight Hybrid Vision Transformer (LH-ViT). This architecture strategically integrates efficient RES-SE blocks for local feature extraction with a novel Radar-ViT module, which employs fold and unfold operations. This specialized attention mechanism drastically reduces the computational demands of multi-head attention, making it particularly adept at capturing global micro-Doppler features efficiently for low-latency HAR. This approach highlights how tailoring the attention mechanism to the specific data modality and application constraints can optimize for both performance and computational cost, a crucial aspect for embedded systems.

Similarly, in automated plant disease classification, the demand for real-time, on-site diagnosis in agricultural settings drives the need for highly efficient models. borhani2022w8x tackled the computational burden by proposing custom, simplified CNN and Transformer building blocks within novel hybrid CNN-ViT architectures. Their systematic investigation demonstrated that these custom hybrid models effectively mitigate the speed deceleration associated with attention mechanisms while maintaining high diagnostic accuracy, crucial for real-time agricultural deployments on portable devices. In contrast, tabbakh2023ao7 proposed TLMViT, combining transfer learning models with ViTs for deep feature extraction, showcasing the benefit of multi-model integration for improved performance by leveraging pre-trained knowledge. Another approach, GNViT **?**, focuses on enhancing a Vision Transformer model for groundnut pest classification through robust data augmentation and transfer learning, achieving high accuracy. While p2024nbn claims "near-perfect accuracy," a critical review notes that such claims require specific metrics (e.g., F1-score, AUC) and dataset context for proper academic assessment, and the implicit requirement for efficient inference for practical field use is a key challenge not explicitly detailed in their architectural innovations. These agricultural examples illustrate a spectrum of solutions: from fine-grained custom block design for specific data types to leveraging transfer learning and multi-model ensembles for robustness, all aimed at enabling practical, real-time field deployments.

Beyond architectural design, achieving real-time performance on extremely resource-constrained edge devices often necessitates further post-design optimizations and sophisticated hardware-software co-design. While the detailed methodologies for pruning, quantization, and hardware acceleration are discussed in Section 5.4, their application is critical for deploying lightweight ViTs in stringent real-time scenarios. For instance, models developed for mobile vision or agricultural monitoring are frequently subjected to aggressive quantization (e.g., 8-bit or even lower precision) to reduce memory footprint and computational cost, enabling their execution on dedicated edge AI accelerators. Such co-design efforts are exemplified by frameworks like EQ-ViT **?**, which, while an acceleration framework itself, demonstrates how end-to-end optimization, including quantization-aware training and heterogeneous computing, can achieve deterministic low-latency inference (e.g., sub-millisecond) and significant energy efficiency gains for real-time ViT deployment. These advancements highlight that the most impactful gains for real-world applications often stem from a holistic approach that considers the entire deployment pipeline, from architectural choice to hardware-software synergy.

In conclusion, the research landscape for Vision Transformers in lightweight and real-time applications is characterized by a multi-faceted approach driven by specific application needs. This includes the development of versatile general-purpose lightweight models like MobileViT for broad mobile vision tasks, as well as highly specialized hybrid architectures such as LH-ViT for unique data modalities like radar-based HAR. These architectural innovations are often complemented by post-design algorithmic optimizations and rigorous hardware-software co-design to meet the stringent latency, power, and memory constraints of edge devices. Despite these advancements, challenges remain in achieving optimal accuracy, robustness, and generalizability across highly diverse and unpredictable real-world environments, especially for ultra-low-power, extremely resource-constrained edge deployments. Future directions will likely involve further adaptive models that dynamically adjust complexity based on available resources and more robust, application-aware optimization techniques to ensure ViTs can reliably operate in the most demanding real-time scenarios.

7 Future Directions and Open Challenges

7.1 Towards More Efficient and Scalable Architectures

The profound capabilities of Vision Transformers (ViTs) are often accompanied by significant computational and memory demands, primarily due to the quadratic complexity of the self-attention mechanism with respect to the number of input tokens. This inherent limitation presents substantial hurdles for processing high-resolution imagery, achieving real-time inference, and deploying models on resource-constrained edge devices. While earlier sections (e.g., Section 3.1 on Swin Transformer) discussed hierarchical designs that mitigate this by introducing windowed attention, the continuous drive for efficiency pushes beyond these established paradigms, focusing on more fundamental architectural and algorithmic innovations. This subsection explores emerging research directions aimed at minimizing the memory footprint and computational overhead while maintaining or improving performance, which is crucial for making ViTs practical for widespread use in real-world applications **?**. This pursuit directly addresses the fundamental trade-offs between model capacity, speed, and resource consumption, seeking to unlock new application domains for ViTs.

One critical avenue for future efficiency improvement lies in fundamentally re-imagining the self-attention mechanism itself, moving beyond its quadratic scaling. While hierarchical window-based approaches achieve linear complexity with respect to image size, they still involve quadratic complexity within each window. *Linear attention mechanisms* represent a significant research frontier, aiming to reduce complexity by approximating the softmax operation or factorizing the attention matrix, often without explicit non-linearities. These methods typically fall into categories such as kernel-based approximations (e.g., using random Fourier features to approximate the softmax kernel), low-rank matrix factorizations, or explicit removal of non-linearities. For instance, UFO-ViT **?** proposes a Unit Force Operated Vision Transformer that achieves linear complexity by eliminating non-linearity from the original self-attention and factorizing matrix multiplication. This approach, by modifying only a few lines of code, demonstrates competitive or superior performance on image classification and dense prediction tasks across various model capacities, highlighting the potential of simplified attention computations to push efficiency boundaries. Such efforts are crucial for scaling ViTs to unprecedented input sizes, such as gigapixel images, where even windowed attention might be prohibitive.

Beyond static approximations, researchers are actively exploring *adaptive and sparse attention mechanisms* that dynamically focus computational resources on the most salient image regions. This mitigates the inefficiencies of both dense global attention (high cost, often processing irrelevant features) and fixed sparse attention (data-agnostic limitations). The *Vision Transformer with Deformable Attention* (DAT) **?** and its enhanced version DAT++ **?** exemplify this by introducing a deformable multi-head attention module. Here, the positions of key and value pairs are adaptively allocated in a data-dependent manner, allowing the model to dynamically attend to relevant regions. This flexible scheme maintains the representational power of global attention while significantly improving efficiency and performance across various vision tasks. Similarly, LF-ViT **?** addresses spatial redundancy by strategically curtailing computational demands. It processes a reduced-resolution image, and if a definitive prediction is elusive, a Neighborhood Global Class

Attention (NGCA) mechanism identifies class-discriminative regions, which are then used from the original image for enhanced recognition. This two-phase approach, with consistent parameters across phases, significantly reduces FLOPs and amplifies throughput without compromising performance, offering a path towards more intelligent resource allocation.

Another critical direction involves *architectural simplification and streamlining* to reduce overhead and improve inference speed. While Section 5.2 discusses alternative token mixers, this thread focuses on optimizing the *overall structure* of ViTs. Hiera **?** exemplifies this by arguing that many "bells-and-whistles" added to modern hierarchical vision transformers for supervised classification performance are unnecessary. By leveraging strong self-supervised pre-training (e.g., Masked Autoencoders, MAE), Hiera demonstrates that a significantly simpler hierarchical ViT can achieve higher accuracy and be substantially faster both at inference and during training. This suggests that future efficient designs might prioritize architectural minimalism, relying on robust pre-training to imbue models with necessary inductive biases, thereby reducing the need for complex, hand-engineered components that add computational overhead.

For practical deployment, especially on mobile and edge devices, *hardware-aware designs and quantization techniques* are paramount. These approaches optimize ViTs not just algorithmically but also for the specific constraints of target hardware, pushing towards ultra-low-power and real-time inference. Hardware accelerators like ViTA **?** are specifically designed for ViT inference on resource-constrained edge devices, employing head-level pipelines and inter-layer MLP optimizations to achieve high hardware utilization and reasonable frame rates with low power consumption. Furthermore, *quantization* is a crucial strategy for reducing model size and computational cost. Q-ViT **?** proposes a fully differentiable quantization method where both quantization scales and bit-widths are learnable, leveraging head-wise bit-width to squeeze model size while preserving performance. It also identifies Multi-head Self-Attention (MSA) and GELU as key aspects for ViT quantization robustness. Pushing the limits further, Bi-ViT **?** explores fully-binarized ViTs, addressing attention distortion caused by gradient vanishing and ranking disorder through learnable scaling factors and ranking-aware distillation. Such extreme quantization can yield significant theoretical acceleration in FLOPs, albeit with careful management of accuracy trade-offs. The ultimate goal is often achieved through *algorithm-hardware co-design*, as seen in EQ-ViT **?**, an end-to-end acceleration framework for real-time ViT inference on platforms like AMD Versal ACAP. This framework combines a novel spatial and heterogeneous accelerator architecture with a comprehensive quantization-aware training strategy, demonstrating significant speedups and energy efficiency gains over existing solutions. Similarly, FPGA-aware automatic acceleration frameworks with mixed-scheme quantization **?** are being developed to optimize ViTs for specific FPGA architectures, achieving substantial improvements in frame rate with minimal accuracy drops. These hardware-centric optimizations are particularly synergistic with algorithmic advancements like sparse attention, where reduced precision in less salient regions could yield further computational savings.

In summary, the continuous drive for efficient and scalable ViT architectures is a multifaceted research endeavor that extends beyond current state-of-the-art solutions. It spans from fundamental re-designs of the attention mechanism (e.g., linear attention), through the exploration of adaptive and sparse attention, to architectural streamlining and sophisticated hardware-aware co-design and aggressive quantization strategies. These efforts are critical for overcoming the inherent computational challenges of ViTs, enabling their deployment across a broader spectrum of real-world applications, from high-resolution medical imaging to real-time edge computing, by meticulously balancing model capacity, speed, and resource consumption. The future of ViT efficiency lies in the intelligent integration of these diverse strategies, pushing the boundaries of what is computationally feasible.

7.2 Beyond Vision: Multimodal and Foundation Models

The trajectory of artificial intelligence is rapidly shifting from single-modality, task-specific models towards more generalized and holistic understanding, spearheaded by the emergence of multimodal and foundation models. This paradigm aims to develop powerful

Vision Transformers (ViTs) capable of processing and integrating information from diverse modalities, such as text, audio, or 3D data, thereby moving closer to human-like comprehension. These large-scale "foundation models," pre-trained on vast and varied datasets, are designed to serve as universal backbones, learning rich, transferable representations across different data types and tasks. Their significance lies in their ability to generalize to novel tasks and data distributions with minimal or no fine-tuning, a critical step towards more generalized artificial intelligence.

A pivotal development in this shift has been the rise of Vision-Language Pre-training (VLP), which leverages the scalability of Transformer architectures to learn joint representations of images and text. Models like CLIP (Contrastive Language-Image Pre-training) **?** and ALIGN (A Large-scale ImaGe-Nosearch pre-training) **?** exemplify this approach. They are trained on massive datasets of image-text pairs (e.g., 400 million for CLIP, 1.8 billion for ALIGN) using contrastive learning objectives, where the model learns to associate corresponding image and text embeddings while distinguishing them from non-matching pairs. This pre-training enables remarkable zero-shot transfer capabilities, allowing the models to perform tasks like image classification or retrieval on unseen categories without explicit fine-tuning, simply by comparing image features to text prompts. The power of these models lies in their ability to bridge the semantic gap between visual and linguistic domains, demonstrating a foundational understanding that extends beyond raw pixel values.

Building upon these VLP successes, the concept of foundation models has expanded to encompass general-purpose visual backbones. While earlier self-supervised methods like MAE **?** and DINO **?** were crucial for learning robust visual features from unlabeled data, models like DINOv2 **?** represent a further scaling of this paradigm. DINOv2 trains large ViTs on billions of unlabeled images, yielding highly robust and generalizable visual features that can serve as strong backbones for a wide array of downstream vision tasks, often outperforming supervised pre-training. Critically, these models exhibit emergent properties, such as the ability to perform dense pixel-level tasks without explicit supervision, showcasing their deep understanding of visual semantics. Another significant example is the Segment Anything Model (SAM) ?, a foundation model specifically designed for promptable segmentation. SAM is trained on an unprecedented dataset of 11 million images and 1.1 billion masks, allowing it to segment any object in an image given various prompts (e.g., points, bounding boxes, text). This demonstrates the power of large-scale pre-training to create models with remarkable generalization and interactive capabilities, moving beyond fixed-category segmentation to a more flexible, user-driven approach.

The ultimate goal of multimodal foundation models is to integrate diverse sensory inputs for more complex reasoning and in-context learning. Models such as Flamingo ? represent a significant step in this direction, combining powerful pre-trained vision encoders (like CLIP) with large language models (LLMs) to enable few-shot learning for vision-language tasks. Flamingo achieves this by using cross-attention layers that condition the LLM on visual features, allowing it to process interleaved image and text inputs and generate coherent responses based on a few examples. This architecture enables capabilities like visual question answering, image captioning, and visual dialogue with unprecedented flexibility. Similarly, generalist agents like Gato ? demonstrate the potential for a single Transformer to perform a wide range of tasks across different modalities, from playing Atari games to controlling robotic arms, by treating diverse inputs and outputs as a unified sequence. This approach highlights the ambition to create truly general-purpose AI systems that can learn and adapt across domains, moving beyond specialized models to a more unified intelligence. Furthermore, EVA ? explores the limits of transfer learning with a unified text-and-image encoder, demonstrating a direct step towards more comprehensive multimodal understanding by leveraging both image-text and image-only data. The integration of deformable convolutions into Transformer-like architectures, as seen in InternImage ?, further exemplifies the ongoing exploration of combining strengths from different paradigms to build large-scale vision foundation models.

Despite these significant advancements, several challenges remain. The sheer scale of these models necessitates immense computational resources for training and deployment, raising concerns about their environmental impact and accessibility. Data curation for multimodal pre-training is also a complex task, as biases present in large web-scraped datasets can lead to unfair or harmful model behaviors. Furthermore, while these models demonstrate impressive emergent capabilities, the mechanisms for true cross-modal reasoning, beyond superficial integration or concatenation, are still an active area of research. Ensuring interpretability and robustness in such complex, black-box systems is also critical for their responsible deployment. The future direction points towards even larger, more data-efficient, and ethically aligned foundation models that can seamlessly integrate information from the visual world with other sensory inputs, fostering a more holistic and adaptable understanding.

7.3 Novel Architectures and Beyond Attention Mechanisms

The pervasive dominance of the Transformer architecture in visual AI, largely due to its potent self-attention mechanism, has spurred an intense search for alternative architectural paradigms that can overcome its inherent limitations, particularly concerning computational cost, memory footprint, and the explicit capture of local inductive biases. This subsection delves into emerging and entirely new architectural designs and token mixing mechanisms that either move significantly beyond or fundamentally modify the self-attention block, challenging the prevailing Transformer-centric view. The exploration of these novel approaches, such as state-space models (SSMs) and potentially biologically inspired mechanisms, aims to offer superior efficiency, distinct inductive biases, and innovative strategies for capturing long-range dependencies, thereby diversifying the architectural landscape of visual AI.

While initial Vision Transformers (ViTs) demonstrated the power of global attention, subsequent research focused on optimizing these architectures through hierarchical designs and windowed attention to mitigate quadratic complexity and improve performance on dense prediction tasks. Even efforts to replace self-attention with simpler token mixers, such as Fourier transforms or pooling operations, have shown competitive results, suggesting that the broader "meta-architecture" of Transformers might be as crucial as the specific attention mechanism itself. However, these approaches often operate within the established Transformer block structure, prompting a deeper inquiry into fundamentally different computational primitives for sequence modeling in vision.

A promising direction involves the adaptation of State-Space Models (SSMs), which have recently demonstrated remarkable capabilities in efficiently modeling long sequences, particularly in natural language processing. SSMs offer an alternative to attention by processing sequences through a hidden state that evolves over time, enabling efficient capture of long-range dependencies with linear complexity. This mechanism inherently provides a different inductive bias compared to the global pairwise interactions of self-attention, potentially leading to more efficient and effective visual representations.

A significant stride in this direction is presented by hatamizadeh2024xr6 with *MambaVision*, a novel hybrid Mamba-Transformer backbone specifically engineered for vision applications. This work directly addresses the challenge of moving beyond traditional attention by redesigning the Mamba formulation to enhance its capability for efficient modeling of visual features. MambaVision's core innovation lies in its ability to process visual information sequentially while maintaining a global context through its state-space representation, offering a compelling alternative to the quadratic complexity of full self-attention. Crucially, hatamizadeh2024xr6 demonstrate through extensive ablation studies the feasibility and benefits of integrating Vision Transformers (ViT) with Mamba. Their findings reveal that equipping the Mamba architecture with self-attention blocks, particularly in the final layers, significantly improves its capacity to capture intricate long-range spatial dependencies. This hybrid approach suggests that a synergistic combination of different token mixing mechanisms—the efficient sequential processing of Mamba and the global interaction of self-attention—can yield superior performance.

The family of MambaVision models introduced by hatamizadeh2024xr6 adopts a hierarchical architecture, akin to successful Vision Transformers like Swin, to meet various design criteria and scale effectively. For image classification on the ImageNet-1K dataset, MambaVision variants achieve state-of-the-art (SOTA) performance in terms of both Top-1 accuracy and throughput, underscoring its computational efficiency without sacrificing accuracy. Furthermore, in downstream tasks such as object detection, instance segmentation on MS COCO, and semantic segmentation on ADE20K datasets, MambaVision consistently outperforms comparably sized backbones while demonstrating favorable performance. These results highlight MambaVision's potential to offer better trade-offs between performance, computational cost, and generalizability across diverse visual tasks.

The emergence of architectures like MambaVision represents a critical juncture in visual AI, challenging the long-held assumption that self-attention is the sole or optimal mechanism for global context aggregation. By leveraging the strengths of State-Space Models, MambaVision provides a concrete example of how novel architectural paradigms can offer distinct advantages in efficiency and inductive biases, particularly for capturing long-range dependencies in a more streamlined manner. The hybrid nature of MambaVision also opens up new avenues for research into optimally combining different token mixing strategies, moving beyond a monolithic architectural design towards more modular and functionally specialized components. Unresolved issues include a deeper theoretical understanding of the inductive biases introduced by SSMs in vision, the exploration of other biologically inspired mechanisms, and the identification of optimal hybrid configurations that balance the strengths of diverse architectural primitives. Future directions will likely involve further refinement of SSMs for vision, investigating their interpretability, and exploring novel ways to integrate them with other computational blocks to unlock new breakthroughs in visual representation learning.

7.4 Ethical Considerations and Societal Impact

The remarkable ascent of Vision Transformers (ViTs) has unlocked unprecedented capabilities across diverse computer vision tasks, yet this increasing power necessitates a rigorous examination of their profound ethical implications and broader societal impact. As ViTs become more ubiquitous and their architectures more complex, critical concerns emerge regarding algorithmic bias, the potential for misuse in sensitive applications, and the overarching imperative for responsible AI development that prioritizes transparency, interpretability, and robust fairness evaluations. These are not merely technical challenges but fundamental open questions that will shape public trust and the equitable deployment of advanced AI systems.

A primary ethical challenge revolves around the amplification of algorithmic bias, particularly stemming from the vast, often uncurated datasets used for pre-training large ViT models. Unlike traditional CNNs with stronger inductive biases, ViTs' reliance on global attention and their notorious data hunger often necessitate training on massive, unfiltered web-scale datasets (e.g., LAION). These datasets inevitably reflect and encode societal biases, which ViTs can then learn and perpetuate, leading to unfair or discriminatory outcomes. For instance, in high-stakes applications like medical imaging, the vulnerability of ViTs to adversarial attacks **?** raises serious safety concerns. If a ViT model for medical diagnosis is susceptible to subtle perturbations, its reliability in clinical settings is compromised, potentially leading to misdiagnoses or disparate impacts on different patient populations if the model's robustness varies across demographic groups. Furthermore, while efforts to enhance object detection or scene classification in remote sensing **??** aim for improved accuracy, without careful curation and fairness audits of the augmented data, these techniques could inadvertently amplify existing biases, leading to discriminatory outcomes in applications like urban planning or resource allocation. The complexity of hybrid architectures, such as MambaVision **?**, further complicates the identification and mitigation of these embedded biases, as the sources of learned representations become more opaque.

Beyond bias, the enhanced capabilities and deployment efficiency of advanced ViTs raise significant concerns regarding their potential for misuse, particularly in surveillance and other harmful contexts. Architectures optimized for efficient deployment, such as Next-ViT **?** and TRT-ViT **?**, achieve impressive latency/accuracy trade-offs, making them highly attractive for real-time applications. While beneficial for many commercial uses, this efficiency also lowers the barrier for deploying powerful vision systems in surveillance infrastructure, potentially infringing on privacy and civil liberties. The application of ViTs in autonomous vehicle safety assessment, as explored by **?**, exemplifies both the promise and peril. While ViT-TA can accurately classify critical situations and identify probable causes using attention maps, thereby improving AV safety, the underlying power to analyze complex real-world scenarios in detail also highlights the immense responsibility associated with deploying such systems. Similarly, advancements in face anti-spoofing using ViTs fine-tuned with self-supervised frameworks like DINO **?** enhance biometric security but simultaneously underscore the increasing sophistication of facial recognition technologies, demanding robust ethical guidelines to prevent their weaponization or use in oppressive regimes.

These significant ethical challenges underscore the critical importance of responsible AI development, framing transparency, interpretability, and robust fairness evaluations as crucial future research directions. The "black-box" nature of deep learning is often exacerbated in complex ViT architectures, making it difficult to understand their decision-making processes. This opacity hinders the identification and rectification of biases, necessitating dedicated research into ViT interpretability. Promising avenues include explainable ViTs for medical applications, such as SleepXViT for automatic sleep staging **?** and prototype-based ViTs for COVID-19 detection **?**. SleepXViT, for instance, provides intuitive explanations by mimicking human "visual scoring" and offering high-resolution heatmaps, thereby fostering trust and facilitating synergy between AI and human experts. Similarly, the interpretable ViT by **?** uses prototype parts to explain model decisions, making the inference process transparent and meaningful for critical health applications. These efforts are crucial for moving beyond mere accuracy to ensure that models are trustworthy and accountable.

However, a critical open challenge remains the development of scalable auditing tools specifically designed for billion-parameter ViT foundation models, which are often pre-trained on vast, uncurated datasets. The sheer scale and complexity of these models make traditional fairness audits impractical. Future research must focus on creating inherently fair attention mechanisms that are robust to dataset biases and on developing standardized ethical guidelines and regulatory frameworks for ViT deployment across sensitive domains. The discussion must move beyond generic AI ethics to analyze how ViT-specific properties—such as their global attention mechanism, their notorious data hunger, and the emergent properties from self-supervised learning on vast, uncurated datasets—might introduce novel or exacerbated ethical challenges compared to CNNs. Without proactive measures to embed ethical considerations throughout the AI lifecycle, from data collection and model training to deployment and monitoring, the societal benefits of Vision Transformers could be overshadowed by their unintended negative consequences.

In conclusion, while the rapid advancements in Vision Transformers promise transformative capabilities, they also introduce profound ethical dilemmas that demand immediate and sustained attention. The pervasive risk of algorithmic bias, particularly from large-scale self-supervised pre-training, and the potential for misuse in surveillance and other high-stakes applications necessitate a concerted effort towards responsible AI development. Future research must not only focus on pushing performance boundaries but also prioritize the development of transparent, interpretable, and robust ViT systems. Fostering public trust in these powerful AI technologies hinges on our collective ability to navigate these ethical complexities, ensuring that innovation is guided by a strong commitment to societal well-being and equitable outcomes.

8 Conclusion

8.1 Summary of Key Developments

The emergence of Vision Transformers (ViTs) has profoundly reshaped the landscape of computer vision, challenging the long-standing dominance of Convolutional Neural Networks (CNNs). As introduced in Section 2.1, the foundational ViT architecture **?** demonstrated that by segmenting images into patches and processing them with a standard Transformer encoder, impressive performance could be achieved in image classification. This breakthrough underscored the power of global self-attention in capturing long-range dependencies across an entire image, a capability often limited in traditional CNNs. However, the initial ViT paradigm faced significant hurdles: its considerable data hunger, requiring massive pre-training datasets like JFT-300M, and its quadratic computational complexity with respect to image resolution, which hindered its application to high-resolution inputs and dense prediction tasks. Early research, as detailed in Section 2.2, swiftly addressed these limitations. Knowledge distillation, notably exemplified by DeiT **?**, enabled ViTs to achieve competitive performance with significantly smaller datasets by learning from pre-trained CNN teachers, effectively transferring inductive biases. Concurrently, efforts focused on enhancing architectural stability and tokenization for deeper models, as discussed in Section 2.3, paving the way for more robust and practical ViT deployments.

The inherent limitations of the original ViT for dense prediction tasks, which demand multi-scale feature representations and efficient processing of high-resolution inputs, spurred a critical wave of architectural innovation. As elaborated in Section 3, the development of hierarchical Vision Transformers became a pivotal breakthrough. The *Swin Transformer* **?** stands out as a landmark contribution, introducing a hierarchical design coupled with shifted window-based self-attention. This ingenious mechanism addressed the quadratic complexity by localizing attention within non-overlapping windows while enabling cross-window information exchange through shifting, thereby generating multi-scale feature maps that scale linearly with image size. This made Swin Transformer a versatile and efficient backbone, capable of excelling in complex tasks like object detection and semantic segmentation, effectively bridging the performance gap with CNNs in these domains. Further advancements in this direction, such as the Pyramid Vision Transformer (PVT) **?** and Multiscale Vision Transformers (MViT) **?**, continued to refine multi-scale feature extraction and efficient attention mechanisms. More recently, the Hiera architecture **?** demonstrated that, with strong self-supervised pretraining, hierarchical ViTs could achieve high accuracy and speed even with simplified designs, challenging the necessity of overly complex, vision-specific components. This evolution underscored a strategic shift towards architectural designs that balance global context with computational efficiency and multi-scale representation.

A transformative development in overcoming ViT's data dependency was the widespread adoption of self-supervised learning (SSL) paradigms, as thoroughly explored in Section 4. These strategies enabled ViTs to learn powerful visual representations from vast amounts of unlabeled data, significantly reducing the reliance on expensive human annotations.

Masked Image Modeling (MIM), inspired by BERT in NLP, emerged as a highly effective approach. Models like Masked Autoencoders (MAE) **?** demonstrated that reconstructing masked image patches forced ViTs to learn rich, semantic features, particularly with high masking ratios that encourage global understanding. Complementary to MIM, self-distillation and contrastive learning methods, such as DINO **?**, revealed remarkable emergent properties in ViT features, including the ability to perform object segmentation without explicit supervision. By training a student network to match a teacher's output, DINO showcased how ViTs could learn robust and semantically meaningful representations through unsupervised means. The success of these SSL techniques has been instrumental in scaling ViTs to unprecedented sizes, leading to the development of 'Vision Foundation Models' **??** that serve as highly robust and general-purpose visual backbones, capable of transferring effectively across a broad spectrum of downstream tasks with minimal fine-tuning.

Beyond architectural refinements and training strategies, a significant research thrust has involved critically re-evaluating the self-attention mechanism itself and exploring hybrid architectures, as detailed in Section 5. This line of inquiry sought to combine the strengths of CNNs, particularly their inductive biases for local feature extraction and computational efficiency, with the global context modeling of Transformers. Hybrid models, such as MobileViT **?** and Next-ViT **?**, strategically integrated convolutional layers or convolutional inductive biases within the Transformer framework, yielding lightweight and efficient designs suitable for deployment on resource-constrained devices. These models often achieve superior performance by leveraging the best of both worlds, demonstrating that a synergistic approach can outperform pure paradigms in certain contexts. Simultaneously, researchers explored radical alternatives to complex self-attention. Works like UFO-ViT **?** introduced linear complexity attention mechanisms, while ShiftViT **?** provocatively showed that even zero-parameter shift operations could replace attention layers while maintaining competitive performance, suggesting that the overall "MetaFormer" architectural design might be more crucial than attention alone. Other innovations, such as PLG-ViT **?** with its parallel local-global self-attention and NomMer **?** with dynamic context nomination, further refined efficient attention by adaptively combining local and global information. This convergence of ideas even led to the modernization of CNNs, exemplified by ConvNeXt **?**, which adopted ViT design principles to achieve competitive performance, effectively blurring the lines between these once distinct architectural paradigms. The recent advent of MambaVision **?**, integrating state-space models, signals a continued exploration of novel token mixing mechanisms beyond traditional attention.

Collectively, these advancements have propelled Vision Transformers from a nascent concept to a mature and versatile technology, finding widespread application across diverse computer vision domains, as showcased in Section 6. From robust object detection and semantic segmentation to specialized tasks in medical image analysis and remote sensing, ViTs, often in their hierarchical or hybrid forms, have demonstrated exceptional adaptability and performance. The continuous evolution reflects an ongoing tension within the field: balancing the expressive power and global receptive field of pure self-attention with the practical demands for computational efficiency, stronger inductive biases, and multi-scale processing. Looking ahead, as discussed in Section 7, future directions will likely involve further convergence of architectural ideas, the exploration of even more novel token mixing mechanisms, and the development of increasingly robust and generalizable multimodal and foundation models. The journey of Vision Transformers underscores a dynamic field driven by continuous innovation, where the pursuit of more powerful, efficient, and universally applicable visual intelligence remains the central objective.

8.2 Unresolved Tensions and Future Outlook

The remarkable ascent of Vision Transformers (ViTs) has profoundly reshaped computer vision, establishing them as formidable contenders, and often successors, to traditional Convolutional Neural Networks (CNNs). However, the field is far from settled, grappling with several fundamental, unresolved tensions that define the frontier of future research. These debates center on achieving an optimal balance between architectural expressivity and computational efficiency, the quest for truly universal visual representations, the practicalities of model deployment, and the seamless integration of ViTs into broader multimodal and foundation models. Navigating these challenges is crucial for unlocking the next generation of visual intelligence.

One of the most persistent architectural tensions lies in reconciling the global context modeling prowess of self-attention with the imperative for computational efficiency and robust local inductive biases. While early ViTs **?** showcased the power of global attention, their quadratic complexity and significant data requirements highlighted the need for more efficient designs. This led to a diverse array of innovations, from hierarchical architectures like Swin Transformer **?** that localize attention to windows, to more adaptive variants such as Deformable Attention Transformer **?** and linear attention mechanisms like UFO-ViT **?**, all striving to reduce computational overhead while preserving long-range dependencies. However, a deeper meta-question emerged: is complex self-attention truly indispensable? Provocative works like PoolFormer **?** and ShiftViT **?** demonstrated that even remarkably simple token mixers, or zero-parameter shift operations, within a Transformer-like "MetaFormer" structure could yield competitive results. This suggests that the overall architectural design and the strategic integration of inductive biases, rather than solely the attention mechanism, are significant drivers of ViT's success (as explored in Section 5.2). This perspective is further reinforced by the modernization of CNNs (e.g., ConvNeXt **?**) using ViT-inspired principles, blurring the lines between the two paradigms (as discussed in Section 5.3). The future likely involves a continued exploration of hybrid models (e.g., MobileViT **?**, Next-ViT **?**) that strategically combine convolutional locality with global attention, and the emergence of entirely new architectural primitives, such as State Space Models in MambaVision **?**. This ongoing architectural dialogue, potentially guided by neural architecture search **?**, underscores a continuous quest for token mixers that optimally balance expressivity, efficiency, and inherent inductive biases, moving towards a more unified theory of visual processing.

The pursuit of truly universal visual representations, capable of transferring across an expansive range of tasks and domains with minimal fine-tuning, remains a paramount challenge. Self-supervised learning (SSL) methods, particularly Masked Autoencoders (MAE) **?** and DINO **?**, have been pivotal in mitigating ViT's data hunger and fostering robust representations (as detailed in Section 4). While systematic studies **?** have demonstrated ViTs' superior transfer learning capabilities compared to ConvNets, the ideal pre-training strategy and architectural design for maximal universality are still under active investigation. Challenges persist in adapting SSL to complex hierarchical architectures, as seen with the need for innovations like Uniform Masking (UM-MAE) **?**. Recent efforts, such as ViTDet **?** and Hiera **?**, suggest that strong self-supervised pre-training can simplify hierarchical ViT designs, leading to more efficient and accurate models. However, the reliance on massive pre-training datasets remains a bottleneck, with research exploring the feasibility of training ViT-based object detectors from scratch **?** highlighting the need for fundamental architectural changes and extended training. A significant step towards this universality is exemplified by GiT **?**, which proposes a "Generalist Vision Transformer" framework capable of handling diverse vision tasks—from image captioning to detection and segmentation—through a universal language interface, without task-specific modules. This paradigm shift aims to foster mutual enhancement across tasks and reduce the architectural gap between vision and language, suggesting a future where visual backbones are inherently more robust and generalizable, reducing the dependency on colossal datasets or enabling efficient learning from scratch.

Beyond theoretical architectures, the practical deployment of ViTs, especially on resource-constrained edge devices, presents critical challenges related to model complexity and efficiency. While lightweight hybrid designs (e.g., MobileViT **?**, as discussed in Section 5.4) offer promising avenues, a holistic approach is essential. This involves not only efficient architectural design but also advanced quantization techniques, such as Q-ViT **?** which employs differentiable head-wise bit-width allocation, and knowledge distillation methods like AttnDistill **?** to transfer knowledge from large teachers to smaller student models. Crucially, algorithm-hardware co-design is emerging as a vital strategy, with specialized accelerators like ViTA **?** and EQ-ViT **?** demonstrating how tailoring hardware to ViT inference patterns can achieve real-time performance and energy efficiency on edge platforms. Further architectural innovations like LF-ViT **?**, which reduces spatial redundancy by focusing computation on class-discriminative regions, also contribute to deployment efficiency. The future of ViT deployment hinges on a synergistic integration of these efforts, pushing beyond theoretical FLOPs towards real-world latency, energy consumption, and robustness in diverse operational environments.

Finally, the integration of ViTs into broader multimodal and foundation models represents perhaps the most ambitious frontier. The token-based nature of Transformers inherently positions them as strong candidates for processing and integrating diverse data types, including text, audio, and 3D data. Early multimodal explorations, such as the Visual Saliency Transformer (VST) **?** for RGB-D data and cross-attention mechanisms in models like CrossViT **?** and SwinCross **?** for fusing features, have laid the groundwork. In specialized domains like remote sensing, models such as ExViT **?** and PolSAR-MPIformer **?** extend ViTs to handle complex hyperspectral, LiDAR, and SAR imagery, enabling more comprehensive environmental understanding. The ultimate vision is the development of unified vision-language foundation models, exemplified by TransVG++ **?** for visual grounding, which aim to achieve a more holistic understanding of the world by learning rich, transferable representations across an extremely broad range of data types and tasks. As noted by **?**, scaling these models to unprecedented sizes and effectively aligning diverse modalities poses significant challenges in data curation, architectural complexity, and computational cost. Nevertheless, this trajectory represents a profound shift towards truly generalized artificial intelligence, where ViTs serve as a core component for intelligent systems capable of perceiving, reasoning, and interacting with a complex, multi-sensory world.

In conclusion, the Vision Transformer landscape is defined by a dynamic interplay of innovation and persistent challenges. The core unresolved tensions—balancing attention's power with efficiency, achieving truly universal representations, enabling practical deployment, and advancing multimodal integration—are actively shaping the research agenda. Future work will undoubtedly continue to explore novel architectural paradigms, push the boundaries of self-supervised learning, embrace algorithm-hardware co-design, and advance multimodal integration. This continuous pursuit will ultimately contribute to the development of more intelligent, adaptable, and responsible AI systems for visual perception, moving beyond current limitations towards a more comprehensive understanding of visual and multimodal data.

References

References

Ze Liu, Yutong Lin, Yue Cao, et al. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.* IEEE International Conference on Computer Vision.

Jingyun Liang, Jie Cao, Guolei Sun, et al. (2021). *SwinIR: Image Restoration Using Swin Transformer.* 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

Kai Han, Yunhe Wang, Hanting Chen, et al. (2020). *A Survey on Vision Transformer.* IEEE Transactions on Pattern Analysis and Machine Intelligence.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda (2021). *CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification.* IEEE International Conference on Computer Vision.

Sachin Mehta, and Mohammad Rastegari (2021). *MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer.* International Conference on Learning Representations.

Yanghao Li, Hanzi Mao, Ross B. Girshick, et al. (2022). *Exploring Plain Vision Transformer Backbones for Object Detection.* European Conference on Computer Vision.

Shoufa Chen, Chongjian Ge, Zhan Tong, et al. (2022). *AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition.* Neural Information Processing Systems.

Zhuofan Xia, Xuran Pan, S. Song, et al. (2022). *Vision Transformer with Deformable Attention.* Computer Vision and Pattern Recognition.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, et al. (2021). *DeepViT: Towards Deeper Vision Transformer.* arXiv.org.

Nian Liu, Ni Zhang, Kaiyuan Wan, et al. (2021). *Visual Saliency Transformer.* IEEE International Conference on Computer Vision.

Seung Hoon Lee, Seunghyun Lee, and B. Song (2021). *Vision Transformer for Small-Size Datasets.* arXiv.org.

Bo Zhang, Shuyang Gu, Bo Zhang, et al. (2021). *StyleSwin: Transformer-based GAN for High-resolution Image Generation.* Computer Vision and Pattern Recognition.

Yun Jiang, Yuan Zhang, Xinyi Lin, et al. (2022). *SwinBTS: A Method for 3D Multimodal Brain Tumor Segmentation Using Swin Transformer.* Brain Science.

Md. Nazmul Islam, Madina Hasan, Md. Kabir Hossain, et al. (2022). *Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography.* Scientific Reports.

Jiashi Li, Xin Xia, W. Li, et al. (2022). *Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios.* arXiv.org.

Ting Yao, Yingwei Pan, Yehao Li, et al. (2022). *Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning.* European Conference on Computer Vision.

Y. Borhani, Javad Khoramdel, and E. Najafi (2022). *A deep learning based approach for automated plant disease classification using vision transformer.* Scientific Reports.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, et al. (2021). *Towards Robust Vision Transformer.* Computer Vision and Pattern Recognition.

Junyu Chen, Yufan He, E. Frey, et al. (2021). *ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration.* arXiv.org.

Shibo Jie, and Zhi-Hong Deng (2022). *Convolutional Bypasses Are Better Vision Transformer Adapters.* European Conference on Artificial Intelligence.

Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu (2022). *SUNet: Swin Transformer UNet for Image Denoising.* International Symposium on Circuits and Systems.

Hezheng Lin, Xingyi Cheng, Xiangyu Wu, et al. (2021). *CAT: Cross Attention in Vision Transformer.* IEEE International Conference on Multimedia and Expo.

Yang Lin, Tianyu Zhang, Peiqin Sun, et al. (2021). *FQ-ViT: Post-Training Quantization for Fully Quantized Vision Transformer.* International Joint Conference on Artificial Intelligence.

Zhikai Li, and Qingyi Gu (2022). *I-ViT: Integer-only Quantization for Efficient Vision Transformer Inference.* IEEE International Conference on Computer Vision.

Yanjing Li, Sheng Xu, Baochang Zhang, et al. (2022). *Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer.* Neural Information Processing Systems.

Jinyu Yang, Jingjing Liu, N. Xu, et al. (2021). *TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation.* IEEE Workshop/Winter Conference on Applications of Computer Vision.

Shixing Yu, Tianlong Chen, Jiayi Shen, et al. (2022). *Unified Visual Transformer Compression.* International Conference on Learning Representations.

Wanyi Zhuang, Qi Chu, Zhentao Tan, et al. (2022). *UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection.* European Conference on Computer Vision.

Peifang Deng, Kejie Xu, and Hong Huang (2021). *When CNNs Meet Vision Transformer: A Joint Framework for Remote Sensing Scene Classification.* IEEE Geoscience and Remote Sensing Letters.

Jun Wang, Xiaohan Yu, and Yongsheng Gao (2021). *Feature Fusion Vision Transformer for Fine-Grained Visual Categorization.* British Machine Vision Conference.

Teng Wang, Lei Gong, Chao Wang, et al. (2022). *ViA: A Novel Vision-Transformer Accelerator Based on FPGA.* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

Xiang Li, Wenhai Wang, Lingfeng Yang, et al. (2022). *Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality.* arXiv.org.

Guangting Wang, Yucheng Zhao, Chuanxin Tang, et al. (2022). *When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism.* AAAI Conference on Artificial Intelligence.

Jiajun Deng, Zhengyuan Yang, Daqing Liu, et al. (2022). *TransVG++: End-to-End Visual Grounding With Language Conditioned Vision Transformer.* IEEE Transactions on Pattern Analysis and Machine Intelligence.

Shusheng Yang, Xinggang Wang, Yu Li, et al. (2022). *Temporally Efficient Vision Transformer for Video Instance Segmentation.* Computer Vision and Pattern Recognition.

Behnaz Gheflati, and H. Rivaz (2021). *Vision Transformer for Classification of Breast Ultrasound Images.* arXiv.org.

Xinyu Tang, Zengbing Xu, and Zhigang Wang (2022). *A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model.* Italian National Conference on Sensors.

Xiaohan Yu, Jun Wang, Yang Zhao, et al. (2022). *Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization.* Pattern Recognition.

Hanting Li, Ming-Fa Sui, Feng Zhao, et al. (2021). *MViT: Mask Vision Transformer for Facial Expression Recognition in the wild.* arXiv.org.

Xiaoliang Meng, Yuechi Yang, Libo Wang, et al. (2022). *Class-Guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery.* IEEE Geoscience and Remote Sensing Letters.

Z. Li, Mengshu Sun, Alec Lu, et al. (2022). *Auto-ViT-Acc: An FPGA-Aware Automatic Acceleration Framework for Vision Transformer with Mixed-Scheme Quantization.* International Conference on Field-Programmable Logic and Applications.

Y. Bazi, Mohamad Mahmoud Al Rahhal, M. L. Mekhalfi, et al. (2022). *Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery.* IEEE Transactions on Geoscience and Remote Sensing.

Hao Zheng, Guohui Wang, and Xuchen Li (2022). *Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron.* Journal of Food Measurement Characterization.

Xiaohong W. Gao, Y. Qian, and Alice Gao (2021). *COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models.* arXiv.org.

Zangwei Zheng, Xiangyu Yue, Kai Wang, et al. (2022). *Prompt Vision Transformer for Domain Generalization.* arXiv.org.

Chunguang Bi, Nan Hu, Yiqiang Zou, et al. (2022). *Development of Deep Learning Methodology for Maize Seed Variety Recognition Based on Improved Swin Transformer.* Agronomy.

Yihan Chen, Xingyu Gu, Zhen Liu, et al. (2022). *A Fast Inference Vision Transformer for Automatic Pavement Image Classification and Its Visual Interpretation Method.* Remote Sensing.

Zhuoran Song, Yihong Xu, Zhezhi He, et al. (2022). *CP-ViT: Cascade Vision Transformer Pruning via Progressive Sparsity Prediction.* arXiv.org.

Tao Li, Zheng Zhang, Lishen Pei, et al. (2022). *HashFormer: Vision Transformer Based Deep Hashing for Image Retrieval.* IEEE Signal Processing Letters.

James Wensel, Hayat Ullah, and Arslan Munir (2022). *ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos.* IEEE Access.

Wenxiao Wang, Lu-yuan Yao, Long Chen, et al. (2021). *CrossFormer: A Versatile Vision Transformer Based on Cross-scale Attention.* arXiv.org.

Usman Naseem, Matloob Khushi, and Jinman Kim (2022). *Vision-Language Transformer for Interpretable Pathology Visual Question Answering.* IEEE journal of biomedical and health informatics.

Yanan Wu, Shouliang Qi, Yu Sun, et al. (2021). *A vision transformer for emphysema classification using CT images.* Physics in Medicine and Biology.

Yanjun Lyu, Xiao-Wen Yu, Dajiang Zhu, et al. (2022). *Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer*. Petra.

Koushik Sivarama Krishnan, and Karthik Sivarama Krishnan (2021). *Vision Transformer based COVID-19 Detection using Chest X-rays*. 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC).

Huanrui Yang, Hongxu Yin, Maying Shen, et al. (2021). *Global Vision Transformer Pruning with Hessian-Aware Saliency*. Computer Vision and Pattern Recognition.

Zhexin Li, Tong Yang, Peisong Wang, et al. (2022). *Q-ViT: Fully Differentiable Quantization for Vision Transformer*. arXiv.org.

Hongmiao Wang, Cheng Xing, Junjun Yin, et al. (2022). *Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer*. Remote Sensing.

Minghao Chen, Kan Wu, Bolin Ni, et al. (2021). *Searching the Search Space of Vision Transformer*. Neural Information Processing Systems.

Teerapong Panboonyuen, Kulsawasd Jitkajornwanich, S. Lawawirojwong, et al. (2021). *Transformer-Based Decoder Designs for Semantic Segmentation on Remotely Sensed Images*. Remote Sensing.

Junjie Liang, Cihui Yang, Jingting Zhong, et al. (2022). *BTSwin-Unet: 3D U-shaped Symmetrical Swin Transformer-based Network for Brain Tumor Segmentation with Self-supervised Pre-training*. Neural Processing Letters.

Hong-Yu Zhou, Chi-Ken Lu, Sibei Yang, et al. (2021). *ConvNets vs. Transformers: Whose Visual Representations are More Transferable?*. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

S. Dubey, S. Singh, and Wei Chu (2021). *Vision Transformer Hashing for Image Retrieval*. IEEE International Conference on Multimedia and Expo.

Selen Ayas, and Esra Tunc-Gormus (2022). *SpectralSWIN: a spectral-swin transformer network for hyperspectral image classification.* International Journal of Remote Sensing.

Jialin Tian, Xing Xu, Fumin Shen, et al. (2022). *TVT: Three-Way Vision Transformer through Multi-Modal Hypersphere Learning for Zero-Shot Sketch-Based Image Retrieval.* AAAI Conference on Artificial Intelligence.

Xingyu Liu, Yuehua Wu, Wenkai Liang, et al. (2022). *High Resolution SAR Image Classification Using Global-Local Network Structure Based on Vision Transformer and CNN.* IEEE Geoscience and Remote Sensing Letters.

Yuan Zhang, Jian Cao, Ling Zhang, et al. (2021). *A free lunch from ViT: adaptive attention multi-scale fusion Transformer for fine-grained visual recognition.* IEEE International Conference on Acoustics, Speech, and Signal Processing.

Qi Han, Zejia Fan, Qi Dai, et al. (2021). *Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight.* arXiv.org.

Sangwon Kim, J. Nam, and ByoungChul Ko (2022). *Facial Expression Recognition Based on Squeeze Vision Transformer.* Italian National Conference on Sensors.

Xiaoli Zhou, Chaowei Tang, Pan Huang, et al. (2022). *ASI-DBNet: An Adaptive Sparse Interactive ResNet-Vision Transformer Dual-Branch Network for the Grading of Brain Cancer Histopathological Images.* Interdisciplinary Sciences Computational Life Sciences.

Zhongxu Hu, Yiran Zhang, Yang Xing, et al. (2022). *Toward Human-Centered Automated Driving: A Novel Spatiotemporal Vision Transformer-Enabled Head Tracker.* IEEE Vehicular Technology Magazine.

Haoran You, Yunyang Xiong, Xiaoliang Dai, et al. (2022). *Castling-ViT: Compressing Self-Attention via Switching Towards Linear-Angular Attention at Vision Transformer Inference.* Computer Vision and Pattern Recognition.

Pengzhen Ren, Changlin Li, Guangrun Wang, et al. (2022). *Beyond Fixation: Dynamic Window Visual Transformer*. Computer Vision and Pattern Recognition.

Jinhai Wang, Zongyin Zhang, Lufeng Luo, et al. (2021). *SwinGD: A Robust Grape Bunch Detection Model Based on Swin Transformer in Complex Vineyard Environment*. Horticulturae.

Xiao Xiao, Wenliang Guo, Rui Chen, et al. (2022). *A Swin Transformer-Based Encoding Booster Integrated in U-Shaped Network for Building Extraction*. Remote Sensing.

Sonain Jamil, Muhammad Sohail Abbas, and Anisha Roy (2022). *Distinguishing Malicious Drones Using Vision Transformer*. Applied Informatics.

Long Bai, Liangyu Wang, Tong Chen, et al. (2022). *Transformer-Based Disease Identification for Small-Scale Imbalanced Capsule Endoscopy Dataset*. Electronics.

Kuoyang Li, Min Zhang, Maiping Xu, et al. (2022). *Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion*. Remote Sensing.

Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar (2022). *Self-Ensembling Vision Transformer (SEViT) for Robust Medical Image Classification*. International Conference on Medical Image Computing and Computer-Assisted Intervention.

Z. Sha, and Jianfeng Li (2022). *MITformer: A Multiinstance Vision Transformer for Remote Sensing Scene Classification*. IEEE Geoscience and Remote Sensing Letters.

Xiaosong Zhang, Yunjie Tian, Wei Huang, et al. (2022). *HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling*. arXiv.org.

Nils Hütten, R. Meyes, and Tobias Meisen (2022). *Vision Transformer in Industrial Visual Inspection*. Applied Sciences.

Ali Hatamizadeh, Ziyue Xu, Dong Yang, et al. (2022). *UNetFormer: A Unified Vision Transformer Model and Pre-Training Framework for 3D Medical Image Segmentation*. arXiv.org.

Mansooreh Montazerin, Soheil Zabihi, E. Rahimian, et al. (2022). *ViT-HGR: Vision Transformer-based Hand Gesture Recognition from High Density Surface EMG Signals.* Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa (2022). *Robustifying Vision Transformer without Retraining from Scratch by Test-Time Class-Conditional Feature Alignment.* International Joint Conference on Artificial Intelligence.

Minhee Kang, Wooseop Lee, Keeyeon Hwang, et al. (2022). *Vision Transformer for Detecting Critical Situations And Extracting Functional Scenario for Automated Vehicle Safety Assessment.* Social Science Research Network.

Geng Tian, Ziwei Wang, Chang Wang, et al. (2022). *A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt.* Frontiers in Microbiology.

Lihong Peng, Chang Wang, Geng Tian, et al. (2022). *Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet.* Frontiers in Microbiology.

Chi M. K. Ho, K. Yow, Zhongwen Zhu, et al. (2022). *Network Intrusion Detection via Flow-to-Image Conversion and Vision Transformer Classification.* IEEE Access.

Xin Xia, Jiashi Li, Jie Wu, et al. (2022). *TRT-ViT: TensorRT-oriented Vision Transformer.* arXiv.org.

Zhenmin Wang, Haoyu Chen, Q. Zhong, et al. (2022). *Recognition of penetration state in GTAW based on vision transformer using weld pool image.* The International Journal of Advanced Manufacturing Technology.

Jashila Nair Mogan, C. Lee, K. Lim, et al. (2022). *Gait-ViT: Gait Recognition with Vision Transformer.* Italian National Conference on Sensors.

Yuguang Yang, Hong-Mei Fu, Shang Gao, et al. (2022). *Intrusion detection: A model based on the improved vision transformer.* Transactions on Emerging Telecommunications Technologies.

Nannan Li, Yaran Chen, Weifan Li, et al. (2022). *BViT: Broad Attention-Based Vision Transformer.* IEEE Transactions on Neural Networks and Learning Systems.

Tan Yu, Gangming Zhao, Ping Li, et al. (2022). *BOAT: Bilateral Local Attention Vision Transformer.* British Machine Vision Conference.

Jiacheng Li, Menglin Wang, and Xiaojin Gong (2022). *Transformer Based Multi-Grained Features for Unsupervised Person Re-Identification.* 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW).

Jiahao Huang, Xiaodan Xing, Zhifan Gao, et al. (2022). *Swin Deformable Attention U-Net Transformer (SDAUT) for Explainable Fast MRI.* International Conference on Medical Image Computing and Computer-Assisted Intervention.

Mengxue Qu, Yu Wu, Wu Liu, et al. (2022). *SiRi: A Simple Selective Retraining Mechanism for Transformer-based Visual Grounding.* European Conference on Computer Vision.

Wenyuan Zeng, Meng Li, Wenjie Xiong, et al. (2022). *MPCViT: Searching for Accurate and Efficient MPC-Friendly Vision Transformer with Heterogeneous Attention.* IEEE International Conference on Computer Vision.

Ji Lin, Haifeng Lin, and Fang Wang (2022). $STPM_SAHI : ASmall -$ $TargetForestFireDetectionModelBasedonSwinTransformerandSlicingAidedHyperInference.F$

L. Reghunath, and R. Rajan (2022). *Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music.* EURASIP Journal on Audio, Speech, and Music Processing.

Dipanjali Kundu, Umme Raihan Siddiqi, and Md. Mahbubur Rahman (2022). *Vision*

Transformer based Deep Learning Model for Monkeypox Detection. 2022 25th International Conference on Computer and Information Technology (ICCIT).

Fan Sun, Wujie Zhou, Lv Ye, et al. (2022). *Hierarchical Decoding Network Based on Swin Transformer for Detecting Salient Objects in RGB-T Images.* IEEE Signal Processing Letters.

Yupeng Li, Huimin Lu, Yifan Wang, et al. (2022). *ViT-Cap: A Novel Vision Transformer-Based Capsule Network Model for Finger Vein Recognition.* Applied Sciences.

Bangwei Guo, Xingyu Li, Miao Yang, et al. (2022). *Predicting microsatellite instability and key biomarkers in colorectal cancer from HE-stained images: achieving state-of-the-art predictive performance with fewer data using Swin Transformer.* The Journal of Pathology: Clinical Research.

Ao Li, Yaqin Zhao, and Zhaoxiang Zheng (2022). *Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection.* Forests.

Xiaoben Jiang, Yu Zhu, Gan Cai, et al. (2022). *MXT: A New Variant of Pyramid Vision Transformer for Multi-label Chest X-ray Image Classification.* Cognitive Computation.

Yang Lin, Tianyu Zhang, Peiqin Sun, et al. (2021). *FQ-ViT: Fully Quantized Vision Transformer without Retraining.* arXiv.org.

Jing Wang, Haotian Fa, X. Hou, et al. (2022). *MSTRIQ: No Reference Image Quality Assessment Based on Swin Transformer with Multi-Stage Fusion.* 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Han Li, Sufang Li, Jiguo Yu, et al. (2022). *Plant disease and insect pest identification based on vision transformer.* Other Conferences.

Sangjoon Park, and Jong-Chul Ye (2022). *Multi-Task Distributed Learning Using Vision Transformer With Random Patch Permutation.* IEEE Transactions on Medical Imaging.

Yifan Shen, Li Liu, Zhihao Tang, et al. (2022). *Explainable Survival Analysis with Convolution-Involved Vision Transformer.* AAAI Conference on Artificial Intelligence.

Aili Wang, Shuang Xing, Yan Zhao, et al. (2022). *A Hyperspectral Image Classification Method Based on Adaptive Spectral Spatial Kernel Combined with Improved Vision Transformer*. Remote Sensing.

Tianxin Tao, Daniele Reda, and M. V. D. Panne (2022). *Evaluating Vision Transformer Methods for Deep Reinforcement Learning from Pixels*. arXiv.org.

Shupei Wu, Youqiang Sun, and He Huang (2021). *Multi-granularity Feature Extraction Based on Vision Transformer for Tomato Leaf Disease Recognition*. 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST).

Jinwei Liu, Yan Li, Guitao Cao, et al. (2022). *Feature Pyramid Vision Transformer for MedMNIST Classification Decathlon*. IEEE International Joint Conference on Neural Network.

Yangtao Wang, Yanzhao Xie, Lisheng Fan, et al. (2022). *STMG: Swin transformer for multi-label image recognition with graph convolution network*. Neural computing applications (Print).

Zinan Xiong, Chenxi Wang, Ying Li, et al. (2022). *Swin-Pose: Swin Transformer Based Human Pose Estimation*. Conference on Multimedia Information Processing and Retrieval.

Ruinan Sun, and Yu Pang (2022). *Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on Improved Swin Transformer*. arXiv.org.

Zheng Qi, AprilPyone Maungmaung, Yuma Kinoshita, et al. (2022). *Privacy-Preserving Image Classification Using Vision Transformer*. European Signal Processing Conference.

Xiaojian Ma, Weili Nie, Zhiding Yu, et al. (2022). *RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning*. International Conference on Learning Representations.

Ziyang Wang, Will Zhao, Zixuan Ni, et al. (2022). *Adversarial Vision Transformer for Medical Image Semantic Segmentation with Limited Annotations*. British Machine Vision Conference.

Kai Wang, Fei Yang, and Joost van de Weijer (2022). *Attention Distillation: self-supervised vision transformer students need more guidance*. British Machine Vision Conference.

Fatema-E- Jannat, and A. Willis (2022). *Improving Classification of Remotely Sensed Images with the Swin Transformer*. SoutheastCon.

Yuzhong Chen, Zhe Xiao, Lin Zhao, et al. (2022). *Mask-guided Vision Transformer (MG-ViT) for Few-Shot Learning*. arXiv.org.

Haoran Wang, Yanju Ji, Kaiwen Song, et al. (2021). *ViT-P: Classification of Genitourinary Syndrome of Menopause From OCT Images Based on Vision Transformer Models*. IEEE Transactions on Instrumentation and Measurement.

Usman Sajid, Xiangyu Chen, Hasan Sajid, et al. (2021). *Audio-Visual Transformer Based Crowd Counting*. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

W. Xing, and K. Egiazarian (2022). *Residual Swin Transformer Channel Attention Network for Image Demosaicing*. European Workshop on Visual Information Processing.

A. Garaiman, F. Nooralahzadeh, C. Mihai, et al. (2022). *Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: an artificial intelligence model*. Rheumatology.

Ziyang Wang, Nanqing Dong, and I. Voiculescu (2022). *Computationally-Efficient Vision Transformer for Medical Image Semantic Segmentation Via Dual Pseudo-Label Supervision*. International Conference on Information Photonics.

Zejiang Hou, and S. Kung (2022). *Multi-Dimensional Vision Transformer Compression* *via Dependency Guided Gaussian Process Search*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

L. Agilandeeswari, and S. D. Meena (2022). *SWIN transformer based contrastive self-supervised learning for animal detection and classification*. Multimedia tools and applications.

Haonan Qin, Weiying Xie, Yunsong Li, et al. (2022). *HTD-VIT: Spectral-Spatial Joint Hyperspectral Target Detection with Vision Transformer*. IEEE International Geoscience and Remote Sensing Symposium.

Boyuan Wang (2022). *Automatic Mushroom Species Classification Model for Foodborne Disease Prevention Based on Vision Transformer*. Journal of Food Quality.

Hyunwoo Yu, J. Shim, Jaeho Kwak, et al. (2022). *Vision Transformer-Based Retina Vessel Segmentation with Deep Adaptive Gamma Correction*. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Rayene Amina Boukabouya, A. Moussaoui, and Mohamed Berrimi (2022). *Vision Transformer Based Models for Plant Disease Detection and Diagnosis*. International Symposium on Information and Automation.

Nan Wang, Xiangjun Meng, Xiangchao Meng, et al. (2022). *Convolution-Embedded Vision Transformer With Elastic Positional Encoding for Pansharpening*. IEEE Transactions on Geoscience and Remote Sensing.

Jeonggeun Song (2021). *UFO-ViT: High Performance Linear Vision Transformer without Softmax*. arXiv.org.

Jiangtao Xie, Rui Zeng, Qilong Wang, et al. (2021). *So-ViT: Mind Visual Tokens for Vision Transformer*. arXiv.org.

Yu-shan Sun, Hao Zheng, Guo-cheng Zhang, et al. (2022). *DP-ViT: A Dual-Path Vision Transformer for Real-Time Sonar Target Detection*. Remote Sensing.

Yanhao Jing, and Feng Wang (2022). *TP-VIT: A Two-Pathway Vision Transformer for Video Action Recognition*. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Rui Li, Weihua Li, Yi Yang, et al. (2022). *Swinv2-Imagen: Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation*. Neural computing applications (Print).

Hwanjun Song, Deqing Sun, Sanghyuk Chun, et al. (2022). *An Extendable, Efficient and Effective Transformer-based Object Detector*. arXiv.org.

Neha Shukla, Anand Pandey, A. P. Shukla, et al. (2022). *ECG-ViT: A Transformer-Based ECG Classifier for Energy-Constraint Wearable Devices*. J. Sensors.

Nguyen H. Tran, Ta Duc Huy, S. T. Duong, et al. (2022). *Improving Local Features with Relevant Spatial Information by Vision Transformer for Crowd Counting*. British Machine Vision Conference.

Weixiang Hong, Jiangwei Lao, Wang Ren, et al. (2022). *Training Object Detectors from Scratch: An Empirical Study in the Era of Vision Transformer*. Computer Vision and Pattern Recognition.

Teerapong Panboonyuen, Sittinun Thongbai, W. Wongweeranimit, et al. (2021). *Object Detection of Road Assets Using Transformer-Based YOLOX with Feature Pyramid Decoder on Thai Highway Panorama*. Inf..

Hong Zhao, Zhiwen Chen, Lan Guo, et al. (2022). *Video captioning based on vision transformer and reinforcement learning*. PeerJ Computer Science.

Yuchen Wang, L. Qing, Zhengyong Wang, et al. (2022). *Multi-Level Transformer-Based Social Relation Recognition*. Italian National Conference on Sensors.

Hao Liu, Xinghua Jiang, Xin Li, et al. (2021). *NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition*. Computer Vision and Pattern Recognition.

A. Gul, Ozdemir Cetin, Christoph Reich, et al. (2022). *Histopathological image classification based on self-supervised vision transformer and weak labels.* Medical Imaging.

Youpeng Zhao, Huadong Tang, Yingying Jiang, et al. (2022). *Lightweight Vision Transformer with Cross Feature Attention.* arXiv.org.

Yali Yang, Yuanping Xu, Chaolong Zhang, et al. (2022). *Hierarchical Vision Transformer with Channel Attention for RGB-D Image Segmentation.* International Symposium on Signal Processing Systems.

M. S. Al-Quraishi, I. Elamvazuthi, T. Tang, et al. (2022). *Decoding the User's Movements Preparation From EEG Signals Using Vision Transformer Architecture.* IEEE Access.

Weiqiang Jin, Hang Yu, and Xiangfeng Luo (2021). *CvT-ASSD: Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector.* IEEE International Conference on Tools with Artificial Intelligence.

K. Lee, Bhavin Jawade, D. Mohan, et al. (2022). *Attribute De-biased Vision Transformer (AD-ViT) for Long-Term Person Re-identification.* Advanced Video and Signal Based Surveillance.

Yongtao Shi, Xiaodong Zhao, Fan Zhang, et al. (2022). *Non-Intrusive Load Monitoring Based on Swin-Transformer with Adaptive Scaling Recurrence Plot.* Energies.

Huaqi Zhang, Huang Chen, Jin Qin, et al. (2022). *MC-ViT: Multi-path cross-scale vision transformer for thymoma histopathology whole slide image typing.* Frontiers in Oncology.

Abid Hasan Zim, Aeyan Ashraf, Aquib Iqbal, et al. (2022). *A Vision Transformer-Based Approach to Bearing Fault Classification via Vibration Signals.* Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.

Shuai Bao, Jiping Liu, Liang Wang, et al. (2022). *Landslide Susceptibility Mapping by Fusing Convolutional Neural Networks and Vision Transformer.* Italian National Conference on Sensors.

Mengshu Sun, Z. Li, Alec Lu, et al. (2022). *FPGA-aware automatic acceleration framework for vision transformer with mixed-scheme quantization: late breaking results*. Design Automation Conference.

Travis J. E. Munyer, D. Brinkman, Xin Zhong, et al. (2022). *Foreign Object Debris Detection for Airport Pavement Images Based on Self-Supervised Localization and Vision Transformer*. 2022 International Conference on Computational Science and Computational Intelligence (CSCI).

Hong-wei Fan, Ningge Ma, Xu-hui Zhang, et al. (2022). *New intelligent fault diagnosis approach of rolling bearing based on improved vibration gray texture image and vision transformer*. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science.

Yi Wang, Zhiwen Fan, Tianlong Chen, et al. (2022). *Can We Solve 3D Vision Tasks Starting from A 2D Vision Transformer?*. arXiv.org.

Luqman Ali, Hamad Al Jassmi, Wasif Khan, et al. (2022). *Crack45K: Integration of Vision Transformer with Tubularity Flow Field (TuFF) and Sliding-Window Approach for Crack-Segmentation in Pavement Structures*. Buildings.

Abdeldjalil Chougui, Achraf Moussaoui, and A. Moussaoui (2022). *Plant-Leaf Diseases Classification using CNN, CBAM and Vision Transformer*. International Symposium on Information and Automation.

Li Zhuang (2021). *Deep-Learning-Based Diagnosis of Cassava Leaf Diseases Using Vision Transformer*. Artificial Intelligence and Cloud Computing Conference.

Xiaoyue Chen, Sei-ichiro Kamata, and Weilian Zhou (2021). *Hyperspectral Image Classification Based on Multi-stage Vision Transformer with Stacked Samples*. IEEE Region 10 Conference.

Ali Hatamizadeh, and Jan Kautz (2024). *MambaVision: A Hybrid Mamba-Transformer Vision Backbone*. Computer Vision and Pattern Recognition.

Chaitanya K. Ryali, Yuan-Ting Hu, Daniel Bolya, et al. (2023). *Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles.* International Conference on Machine Learning.

Jing Yao, Bing Zhang, Chenyu Li, et al. (2023). *Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework.* IEEE Transactions on Geoscience and Remote Sensing.

Xiangtai Li, Henghui Ding, Wenwei Zhang, et al. (2023). *Transformer-Based Visual Segmentation: A Survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zhuoyi Zhao, Xiang Xu, Shutao Li, et al. (2024). *Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network.* IEEE Transactions on Geoscience and Remote Sensing.

Mostafa Dehghani, Basil Mustafa, J. Djolonga, et al. (2023). *Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution.* Neural Information Processing Systems.

Yuchen Duan, Weiyun Wang, Zhe Chen, et al. (2024). *Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures.* International Conference on Learning Representations.

Utpal Barman, Parismita Sarma, Mirzanur Rahman, et al. (2024). *ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture.* Agronomy.

Sonain Jamil, and Anisha Roy (2023). *An efficient and robust Phonocardiography (PCG)-based Valvular Heart Diseases (VHD) detection framework using Vision Transformer (ViT).* Comput. Biol. Medicine.

Ishak Paçal, Melek Alaftekin, and F. Zengul (2024). *Enhancing Skin Cancer Diagnosis Using Swin Transformer with Hybrid Shifted Window-Based Multi-head Self-attention and SwiGLU-Based MLP.* Journal of imaging informatics in medicine.

Xiaosong Zhang, Yunjie Tian, Lingxi Xie, et al. (2023). *HiViT: A Simpler and More Efficient Design of Hierarchical Vision Transformer*. International Conference on Learning Representations.

Galib Muhammad Shahriar Himel, Md. Masudul Islam, Kh Abdullah Al-Aff, et al. (2024). *Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System*. International Journal of Biomedical Imaging.

Qin Xu, Jiahui Wang, Bo Jiang, et al. (2023). *Fine-Grained Visual Classification via Internal Ensemble Learning Transformer*. IEEE transactions on multimedia.

Kaichen Chi, Yuan Yuan, and Qi Wang (2023). *Trinity-Net: Gradient-Guided Swin Transformer-Based Remote Sensing Image Dehazing and Beyond*. IEEE Transactions on Geoscience and Remote Sensing.

Badri N. Patro, Vinay P. Namboodiri, and Vijay Srinivas Agneeswaran (2023). *Spect-Former: Frequency and Attention is what you need in a Vision Transformer*. IEEE Workshop/Winter Conference on Applications of Computer Vision.

Xuran Pan, Tianzhu Ye, Zhuofan Xia, et al. (2023). *Slide-Transformer: Hierarchical Vision Transformer with Local Self-Attention*. Computer Vision and Pattern Recognition.

Ziyang Wang, and Chao Ma (2024). *Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation*. arXiv.org.

A. Tabbakh, and Soubhagya Sankar Barpanda (2023). *A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer "TLMViT" for Plant Disease Classification*. IEEE Access.

Pramit Dutta, Khaleda Akhter Sathi, Md.Azad Hossain, et al. (2023). *Conv-ViT: A Convolution and Vision Transformer-Based Hybrid Feature Extraction Method for Retinal Disease Detection*. Journal of Imaging.

Yuhang Qiu, Honghui Chen, Xingbo Dong, et al. (2024). *IFViT: Interpretable Fixed-Length Representation for Fingerprint Matching via Vision Transformer.* IEEE Transactions on Information Forensics and Security.

Guoqiang Li, Yuchao Wang, Qing Zhao, et al. (2023). *PMVT: a lightweight vision transformer for plant disease identification on mobile devices.* Frontiers in Plant Science.

Hu Zhao, Keyan Ren, Tianyi Yue, et al. (2024). *TransFG: A Cross-View Geo-Localization of Satellite and UAVs Imagery Pipeline Using Transformer-Based Feature Aggregation and Gradient Guidance.* IEEE Transactions on Geoscience and Remote Sensing.

Huaxiang Song, Yuxuan Yuan, Zhiwei Ouyang, et al. (2024). *Quantitative regularization in robust vision transformer for remote sensing image classification.* Photogrammetric Record.

Yimin Cai, Yuqing Long, Zhenggong Han, et al. (2023). *Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution.* BMC Medical Informatics and Decision Making.

S. Akinpelu, Serestina Viriri, and A. Adegun (2024). *An enhanced speech emotion recognition using vision transformer.* Scientific Reports.

Mansoor Hayat, Nouman Ahmad, Anam Nasir, et al. (2024). *Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification.* IEEE Access.

Yongxin Li, Mengyuan Liu, You Wu, et al. (2024). *Learning Adaptive and View-Invariant Vision Transformer for Real-Time UAV Tracking.* International Conference on Machine Learning.

Muhammad Asad Arshed, Shahzad Mumtaz, Muhammad Ibrahim, et al. (2023). *Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models.* Inf..

S. Qin, Taiyue Qi, Tang Deng, et al. (2024). *Image segmentation using Vision Transformer for tunnel defect assessment*. Comput. Aided Civ. Infrastructure Eng..

C. Lee, K. Lim, Yu Xuan Song, et al. (2023). *Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer*. Plants.

Jaouad Tagnamas, Hiba Ramadan, Ali Yahyaouy, et al. (2024). *Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images*. Visual Computing for Industry, Biomedicine, and Art.

Shuiwang Li, Yangxiang Yang, Dan Zeng, et al. (2023). *Adaptive and Background-Aware Vision Transformer for Real-Time UAV Tracking*. IEEE International Conference on Computer Vision.

Bofan Song, D. Kc, Rubin Yuchan Yang, et al. (2024). *Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer*. Cancers.

Serra Aksoy, P. Demircioğlu, and I. Bogrekci (2024). *Enhancing Melanoma Diagnosis with Advanced Deep Learning Models Focusing on Vision Transformer, Swin Transformer, and ConvNeXt*. Dermatopathology.

Saebom Leem, and Hyunseok Seo (2024). *Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention*. AAAI Conference on Artificial Intelligence.

Shiming Chen, W. Hou, Salman H. Khan, et al. (2024). *Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning*. Computer Vision and Pattern Recognition.

Fudong Lin, Summer Crawford, Kaleb Guillot, et al. (2023). *MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer*. IEEE International Conference on Computer Vision.

Morteza Ghahremani, Mohammad Khateri, Bailiang Jian, et al. (2024). *H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration*. Computer Vision and Pattern Recognition.

Haiyang Wang, Hao Tang, Li Jiang, et al. (2024). *GiT: Towards Generalist Vision Transformer through Universal Language Interface*. European Conference on Computer Vision.

Mohammad Shahin, F. F. Chen, Mazdak Maghanaki, et al. (2024). *Improving the Concrete Crack Detection Process via a Hybrid Visual Transformer Algorithm*. Italian National Conference on Sensors.

Liang Zhu, Yingyue Li, Jiemin Fang, et al. (2023). *WeakTr: Exploring Plain Vision Transformer for Weakly-supervised Semantic Segmentation*. arXiv.org.

Zitong Yu, Rizhao Cai, Yawen Cui, et al. (2023). *Rethinking Vision Transformer and Masked Autoencoder in Multimodal Face Anti-Spoofing*. International Journal of Computer Vision.

Jinsol Ko, Soyeon Park, and H. G. Woo (2024). *Optimization of vision transformer-based detection of lung diseases from chest X-ray images*. BMC Medical Informatics Decis. Mak..

Qiying Yang, and Rongzuo Guo (2024). *An Unsupervised Method for Industrial Image Anomaly Detection with Vision Transformer-Based Autoencoder*. Italian National Conference on Sensors.

Waleed Nazih, Ahmad O. Aseeri, Osama Youssef Atallah, et al. (2023). *Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images*. IEEE Access.

Zhuofan Xia, Xuran Pan, Shiji Song, et al. (2023). *DAT++: Spatially Dynamic Vision Transformer with Deformable Attention*. arXiv.org.

Shashank Nag, G. Datta, Souvik Kundu, et al. (2023). *ViTA: A Vision Transformer Inference Accelerator for Edge Applications*. International Symposium on Circuits and Systems.

Abdul Haluk Batur Gezici, and Emre Sefer (2024). *Deep Transformer-Based Asset Price and Direction Prediction.* IEEE Access.

Fethi Ghazouani, Pierre Vera, and Su Ruan (2023). *Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention.* International Journal of Computer Assisted Radiology and Surgery.

Guanqun Wang, He Chen, Liang Chen, et al. (2023). *P2FEViT: Plug-and-Play CNN Feature Embedded Hybrid Vision Transformer for Remote Sensing Image Classification.* Remote Sensing.

Yu Guo, Zhi Zhang, and Yuzhen Huang (2024). *Dual Class Token Vision Transformer for Direction of Arrival Estimation in Low SNR.* IEEE Signal Processing Letters.

Wei Wang, Xin Yang, and Jinhui Tang (2023). *Vision Transformer With Hybrid Shifted Windows for Gastrointestinal Endoscopy Image Classification.* IEEE transactions on circuits and systems for video technology (Print).

Fujian Zheng, Shuai Lin, Wei Zhou, et al. (2023). *A Lightweight Dual-Branch Swin Transformer for Remote Sensing Scene Classification.* Remote Sensing.

Jashila Nair Mogan, C. Lee, K. Lim, et al. (2023). *Gait-CNN-ViT: Multi-Model Gait Recognition with Convolutional Neural Networks and Vision Transformer.* Italian National Conference on Sensors.

Nikolas Ebert, D. Stricker, and Oliver Wasenmüller (2023). *PLG-ViT: Vision Transformer with Parallel Local and Global Self-Attention.* Italian National Conference on Sensors.

Yong Wang, Cheng Lu, Hailun Lian, et al. (2024). *Speech Swin-Transformer: Exploring a Hierarchical Transformer with Shifted Windows for Speech Emotion Recognition.* IEEE International Conference on Acoustics, Speech, and Signal Processing.

Jie Cao, Tingting Xu, Yu-he Deng, et al. (2024). *Galaxy morphology classification based on Convolutional vision Transformer (CvT).* Astronomy amp; Astrophysics.

Yaoming Yang, Zhili Cai, Shuxia Qiu, et al. (2024). *Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image*. PLoS ONE.

Tahir Hussain, Hayaru Shouno, Abid Hussain, et al. (2025). *EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification*. IEEE Access.

D. Shim, and H. J. Kim (2023). *SwinDepth: Unsupervised Depth Estimation using Monocular Sequences via Swin Transformer and Densely Cascaded Network*. IEEE International Conference on Robotics and Automation.

Taukir Alam, Wei-Cheng Yeh, Fang Rong Hsu, et al. (2024). *An Integrated Approach using YOLOv8 and ResNet, SeResNet Vision Transformer (ViT) Algorithms based on ROI Fracture Prediction in X-ray Images of the Elbow.*. Current medical imaging.

Ruiping Yang, Liu Kun, Shaohua Xu, et al. (2024). *ViT-UperNet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation*. Complex amp; Intelligent Systems.

Dong Wang, Jian Lian, and Wanzhen Jiao (2024). *Multi-label classification of retinal disease via a novel vision transformer model*. Frontiers in Neuroscience.

Huaxiang Song, Hanjun Xia, Wenhui Wang, et al. (2024). *QAGA-Net: enhanced vision transformer-based object detection for remote sensing images*. International Journal of Intelligent Computing and Cybernetics.

Xiaoye Li, and Bin-Bin Zhang (2023). *FV-ViT: Vision Transformer for Finger Vein Recognition*. IEEE Access.

Xiaochen Ma, Bo Du, Xianggen Liu, et al. (2023). *IML-ViT: Image Manipulation Localization by Vision Transformer*. arXiv.org.

Huiyan Han, H. Zeng, Liqun Kuang, et al. (2024). *A human activity recognition method based on Vision Transformer*. Scientific Reports.

Oğuzhan Katar, and Ozal Yildirim (2023). *An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization.* Diagnostics.

S. Hemalatha, and Jayachandiran Jai Jaganath Babu (2024). *A Multitask Learning-Based Vision Transformer for Plant Disease Localization and Classification.* International Journal of Computational Intelligence Systems.

Chiyu Ma, Jon Donnelly, Wenjun Liu, et al. (2024). *Interpretable Image Classification with Adaptive Prototype-based Vision Transformers.* Neural Information Processing Systems.

D. K. Lai, Zi-Han Yu, Tommy Yau-Nam Leung, et al. (2023). *Vision Transformers (ViT) for Blanket-Penetrating Sleep Posture Recognition Using a Triple Ultra-Wideband (UWB) Radar System.* Italian National Conference on Sensors.

Zhikan Wang, Zhongyao Cheng, Jiajie Xiong, et al. (2024). *A Timely Survey on Vision Transformer for Deepfake Detection.* arXiv.org.

Zhixin Ling, Zhen Xing, Xiangdong Zhou, et al. (2023). *PanoSwin: a Pano-style Swin Transformer for Panorama Understanding.* Computer Vision and Pattern Recognition.

Zhifeng Wang, Jialong Yao, Chunyan Zeng, et al. (2023). *Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network.* Syst..

Miao Yin, Burak Uzkent, Yilin Shen, et al. (2023). *GOHSP: A Unified Framework of Graph and Optimization-based Heterogeneous Structured Pruning for Vision Transformer.* AAAI Conference on Artificial Intelligence.

Swapneel Mishra, Saumya Seth, Shrishti Jain, et al. (2024). *Image Caption Generation using Vision Transformer and GPT Architecture.* 2024 2nd International Conference on Advancement in Computation Computer Technologies (InCACCT).

Moein Heidari, Reza Azad, Sina Ghorbani Kolahi, et al. (2024). *Enhancing Efficiency in Vision Transformer Networks: Design Techniques and Insights.* arXiv.org.

Sheng Yu, Dihua Zhai, and Yuanqing Xia (2023). *A Novel Robotic Pushing and Grasping Method Based on Vision Transformer and Convolution.* IEEE Transactions on Neural Networks and Learning Systems.

Qihao Zhao, Yangyu Huang, Wei Hu, et al. (2023). *MixPro: Data Augmentation with MaskMix and Progressive Attention Labeling for Vision Transformer.* International Conference on Learning Representations.

C. Pan, Junran Peng, and Zhaoxiang Zhang (2024). *Depth-Guided Vision Transformer With Normalizing Flows for Monocular 3D Object Detection.* IEEE/CAA Journal of Automatica Sinica.

Sha Huan, Zhaoyue Wang, Xiaoqiang Wang, et al. (2023). *A lightweight hybrid vision transformer network for radar-based human activity recognition.* Scientific Reports.

Mohamad Mulham Belal, and Dr. Divya Meena Sundaram (2023). *Global-Local Attention-Based Butterfly Vision Transformer for Visualization-Based Malware Classification.* IEEE Access.

Yanjing Li, Sheng Xu, Mingbao Lin, et al. (2023). *Bi-ViT: Pushing the Limit of Vision Transformer Quantization.* AAAI Conference on Artificial Intelligence.

Yingzi Huo, Kai Jin, Jiahong Cai, et al. (2023). *Vision Transformer (ViT)-based Applications in Image Classification.* 2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS).

Jiseob Kim, Kyuhong Shim, Junhan Kim, et al. (2023). *Vision Transformer-Based Feature Extraction for Generalized Zero-Shot Learning.* IEEE International Conference on Acoustics, Speech, and Signal Processing.

Chunyu Fan, Q. Su, Zhifeng Xiao, et al. (2023). *ViT-FRD: A Vision Transformer Model for Cardiac MRI Image Segmentation Based on Feature Recombination Distillation.* IEEE Access.

Kai Zhao, Ruitao Lu, Siyu Wang, et al. (2023). *ST-YOLOA: a Swin-transformer-based YOLO model with an attention mechanism for SAR ship detection under complex background.* Frontiers in Neurorobotics.

Tao Xie, Kun Dai, Zhiqiang Jiang, et al. (2023). *ViT-MVT: A Unified Vision Transformer Network for Multiple Vision Tasks.* IEEE Transactions on Neural Networks and Learning Systems.

Gary Y. Li, Junyu Chen, Se-In Jang, et al. (2023). *SwinCross: Cross-modal Swin Transformer for Head-and-Neck Tumor Segmentation in PET/CT Images.* Medical Physics (Lancaster).

Xiaochen Ma, Bo Du, Zhuohang Jiang, et al. (2023). *IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer.* Unpublished manuscript.

Oluwatosin Tanimola, Olamilekan Shobayo, O. Popoola, et al. (2024). *Breast Cancer Classification Using Fine-Tuned SWIN Transformer Model on Mammographic Images.* Analytics.

Tiansheng Chen, and L. Mo (2023). *Swin-Fusion: Swin-Transformer with Feature Fusion for Human Action Recognition.* Neural Processing Letters.

Navin Ranjan, and Andreas E. Savakis (2024). *LRP-QViT: Mixed-Precision Vision Transformer Quantization via Layer-wise Relevance Propagation.* arXiv.org.

Xiangqu Fu, Qirui Ren, Hao Wu, et al. (2023). *P3 ViT: A CIM-Based High-Utilization Architecture With Dynamic Pruning and Two-Way Ping-Pong Macro for Vision Transformer.* IEEE Transactions on Circuits and Systems Part 1: Regular Papers.

Chaojun Shi, Shiwei Zhao, Kecheng Zhang, et al. (2023). *Face-based age estimation using improved Swin Transformer with attention-based convolution.* Frontiers in Neuroscience.

Deressa Wodajo Deressa, Hannes Mareen, Peter Lambert, et al. (2023). *GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer.* Applied Sciences.

Sanad Aburass, and O. Dorgham (2023). *Performance Evaluation of Swin Vision Transformer Model using Gradient Accumulation Optimization Technique.* arXiv.org.

Vikas Hassija, Balamurugan Palanisamy, Arpita Chatterjee, et al. (2025). *Transformers for Vision: A Survey on Innovative Methods for Computer Vision.* IEEE Access.

Yihang Huang, and Wan Li (2023). *Resizer Swin Transformer-Based Classification Using sMRI for Alzheimer's Disease.* Applied Sciences.

Jianwei Liu, Shirui Lyu, Denis Hadjivelichkov, et al. (2023). _ViT-A*: Legged Robot Path Planning using Vision Transformer A*._ IEEE-RAS International Conference on Humanoid Robots.

Ru He, Xiaomin Wang, Huazhen Chen, et al. (2023). *VHR-BirdPose: Vision Transformer-Based HRNet for Bird Pose Estimation with Attention Mechanism.* Electronics.

Yangyang Guo, Wenhao Hong, Jiaxin Wu, et al. (2023). *Vision-Based Cow Tracking and Feeding Monitoring for Autonomous Livestock Farming: The YOLOv5s-CA+DeepSORT-Vision Transformer.* IEEE robotics automation magazine.

Yun Wang, Shuai Shi, and Jie Chen (2023). *Efficient Blind Hyperspectral Unmixing with Non-Local Spatial Information Based on Swin Transformer.* IEEE International Geoscience and Remote Sensing Symposium.

Goutam Yelluru Gopal, and Maria A. Amer (2023). *Mobile Vision Transformer-based Visual Object Tracking.* British Machine Vision Conference.

Zhiyang Liu, Pengyu Yin, and Zhenhua Ren (2023). *An Efficient FPGA-Based Accelerator for Swin Transformer.* arXiv.org.

Zujun Fu (2022). *Vision Transformer: Vit and its Derivatives.* arXiv.org.

P. Sahoo, S. Saha, S. Mondal, et al. (2022). *Vision Transformer Based COVID-19 Detection Using Chest CT-scan images.* 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).

Roy Ganz, Yair Kittenplon, Aviad Aberdam, et al. (2024). *Question Aware Vision Transformer for Multimodal Reasoning.* Computer Vision and Pattern Recognition.

Ishak Paçal, and Ismail Kunduracioglu (2024). *Data-Efficient Vision Transformer Models for Robust Classification of Sugarcane.* Journal of Soft Computing and Decision Analytics.

Nada M. Hassan, Safwat Hamad, and Khaled Mahar (2024). *YOLO-based CAD framework with ViT transformer for breast mass detection and classification in CESM and FFDM images.* Neural computing applications (Print).

Abinaya K, and S. B (2024). *A Deep Learning-Based Approach for Cervical Cancer Classification Using 3D CNN and Vision Transformer..* Journal of imaging informatics in medicine.

Xuan-Bac Nguyen, Hoang-Quan Nguyen, Samuel Yen-Chi Chen, et al. (2024). *QClusformer: A Quantum Transformer-based Framework for Unsupervised Visual Clustering.* International Conference on Quantum Computing and Engineering.

Abdulaziz Almohimeed, Mohamed Shehata, Nora El-Rashidy, et al. (2024). *ViT-PSO-SVM: Cervical Cancer Predication Based on Integrating Vision Transformer with Particle Swarm Optimization and Support Vector Machine.* Bioengineering.

Chao Hao, Zitong Yu, Xin Liu, et al. (2024). *A Simple Yet Effective Network Based on Vision Transformer for Camouflaged Object and Salient Object Detection.* IEEE Transactions on Image Processing.

Haiming Yao, Wei Luo, Jianan Lou, et al. (2024). *Scalable Industrial Visual Anomaly Detection With Partial Semantics Aggregation Vision Transformer.* IEEE Transactions on Instrumentation and Measurement.

Peiyan Dong, Jinming Zhuang, Zhuoping Yang, et al. (2024). *EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP Architecture.* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

Haoyu Zhang, Raghavendra Ramachandra, Kiran B. Raja, et al. (2024). *Generalized Single-Image-Based Morphing Attack Detection Using Deep Representations from Vision Transformer.* 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

D. E. Boukhari (2024). *Facial Beauty Prediction Based on Vision Transformer.* International Journal of Electrical and Electronic Engineering amp; Telecommunications.

Huaxiang Song, Junping Xie, Yunyang Wang, et al. (2025). *Optimized Data Distribution Learning for Enhancing Vision Transformer-Based Object Detection in Remote Sensing Images.* Photogrammetric Record.

Heng Zhou, Jingmin Yang, Shanghui Deng, et al. (2024). *VTIL: A multi-layer indoor location algorithm for RSSI images based on vision transformer.* Engineering Research Express.

Wafae Abbaoui, Sara Retal, Soumia Ziti, et al. (2024). *Automated Ischemic Stroke Classification from MRI Scans: Using a Vision Transformer Approach.* Journal of Clinical Medicine.

Xiangyang Yang, Dan Zeng, Xucheng Wang, et al. (2024). *Adaptively Bypassing Vision Transformer Blocks for Efficient Visual Tracking.* Pattern Recognition.

Zhiding Yang, and Weimin Huang (2024). *SWHFormer: A Vision Transformer for Significant Wave Height Estimation From Nautical Radar Images.* IEEE Transactions on Geoscience and Remote Sensing.

Youbing Hu, Yun Cheng, Anqi Lu, et al. (2024). *LF-ViT: Reducing Spatial Redundancy in Vision Transformer for Efficient Image Recognition.* AAAI Conference on Artificial Intelligence.

Bin Yang, Binghan Zhang, Yilong Han, et al. (2024). *Vision transformer-based visual language understanding of the construction process.* Alexandria Engineering Journal.

Arman Keresh, and Pakizar Shamoi (2024). *Liveness Detection in Computer Vision: Transformer-Based Self-Supervised Learning for Face Anti-Spoofing.* IEEE Access.

Yuxin Hu, Han Zhou, Ning Cao, et al. (2024). *Synthetic CT generation based on CBCT using improved vision transformer CycleGAN.* Scientific Reports.

Abdulkream A Alsulami, Aishah Albarakati, A. A. Al-Ghamdi, et al. (2024). *Identification of Anomalies in Lung and Colon Cancer Using Computer Vision-Based Swin Transformer with Ensemble Model on Histopathological Images.* Bioengineering.

Lu Yang, Songtao Guo, Defang Liu, et al. (2024). *ConViTML: A Convolutional Vision Transformer-Based Meta-Learning Framework for Real-Time Edge Network Traffic Classification.* IEEE Transactions on Network and Service Management.

Venkatasaichandrakanth P, and I. M (2024). *GNViT- An enhanced image-based groundnut pest classification using Vision Transformer (ViT) model.* PLoS ONE.

Xinhao Wu, Sirui Xu, Ming-Yu Gao, et al. (2024). *A new ECT image reconstruction algorithm based on Vision transformer (ViT).* Flow Measurement and Instrumentation.

Qiwei Dong, Xiaoru Xie, and Zhongfeng Wang (2024). *SWAT: An Efficient Swin Transformer Accelerator Based on FPGA.* Asia and South Pacific Design Automation Conference.

S. M. M. Swapno, S. N. Nobel, Md Babul Islam, et al. (2025). *ViT-SENet-Tom: machine learning-based novel hybrid squeeze-excitation network and vision transformer framework for tomato fruits classification.* Neural computing applications (Print).

Dayeon Yoo, Jeesu Kim, and Jinwoo Yoo (2024). *FSwin Transformer: Feature-Space Window Attention Vision Transformer for Image Classification.* IEEE Access.

Kan He, Wei Zhang, Xuejun Zong, et al. (2024). *Network Intrusion Detection Based on Feature Image and Deformable Vision Transformer Classification.* IEEE Access.

Zichen Zhang, Jing Li, C. Cai, et al. (2024). *Bearing Fault Diagnosis Based on Image Information Fusion and Vision Transformer Transfer Learning Model.* Applied Sciences.

Yueqi Zhang, Lichen Feng, Hongwei Shan, et al. (2024). *A 109-GOPs/W FPGA-Based Vision Transformer Accelerator With Weight-Loop Dataflow Featuring Data Reusing and Resource Saving.* IEEE transactions on circuits and systems for video technology (Print).

Xinlong Dong, Peicheng Shi, Yueyue Tang, et al. (2024). *Vehicle Classification Algorithm Based on Improved Vision Transformer.* World Electric Vehicle Journal.

Yavuz Emre Kayacan, and I. Erer (2024). *A Vision-Transformer-Based Approach to Clutter Removal in GPR: DC-ViT.* IEEE Geoscience and Remote Sensing Letters.

Jintao Liu, Alfredo Tolón Becerra, José Fernando Bienvenido-Barcena, et al. (2024). *CFFI-Vit: Enhanced Vision Transformer for the Accurate Classification of Fish Feeding Intensity in Aquaculture.* Journal of Marine Science and Engineering.

Huihong Shi, Xin Cheng, Wendong Mao, et al. (2024). *P2-ViT: Power-of-Two Post-Training Quantization and Acceleration for Fully Quantized Vision Transformer.* IEEE Transactions on Very Large Scale Integration (VLSI) Systems.

Xinyue Xin, Ming Li, Yan Wu, et al. (2024). *PolSAR-MPIformer: A Vision Transformer Based on Mixed Patch Interaction for Dual-Frequency PolSAR Image Adaptive Fusion Classification.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Jian Zhou, Guochuan Zhao, and Yonglong Li (2024). *Vison Transformer-Based Automatic Crack Detection on Dam Surface.* Water.

Ehsan Monjezi, G. Akbarizadeh, and Karim Ansari-Asl (2024). *RI-ViT: A Multi-Scale Hybrid Method Based on Vision Transformer for Breast Cancer Detection in Histopathological Images.* IEEE Access.

Eu-tteum Baek (2025). *Attention Score-Based Multi-Vision Transformer Technique for Plant Disease Classification.* Italian National Conference on Sensors.

David L. Payne, Xuan Xu, Farshid Faraji, et al. (2024). *Automated Detection of Cervical Spinal Stenosis and Cord Compression via Vision Transformer and Rules-Based Classification.* American Journal of Neuroradiology.

Nan Qi, Yan Piao, Hao Zhang, et al. (2024). *Seizure prediction based on improved vision transformer model for EEG channel optimization.* Computer Methods in Biomechanics and Biomedical Engineering.

J. Mercier, O. Ertz, and E. Bocher (2024). *Quantifying Dwell Time With Location-based Augmented Reality: Dynamic AOI Analysis on Mobile Eye Tracking Data With Vision Transformer.* Journal of Eye Movement Research.

Mohamed Yacin Sikkandar, S. Sundaram, Ahmad Alassaf, et al. (2024). *Utilizing adaptive deformable convolution and position embedding for colon polyp segmentation with a visual transformer.* Scientific Reports.

Mingyang Hou, Zhiyong Huang, Zhi Yu, et al. (2024). *CSwT-SR: Conv-Swin Transformer for Blind Remote Sensing Image Super-Resolution With Amplitude-Phase Learning and Structural Detail Alternating Learning.* IEEE Transactions on Geoscience and Remote Sensing.

Kintoh Allen Nfor, Tagne Poupi Theodore Armand, Kenesbaeva Periyzat Ismaylovna, et al. (2025). *An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition.* Nutrients.

Changcheng Xiang, Duofen Yin, Fei Song, et al. (2024). *A Fast and Robust Safety Helmet Network Based on a Mutilscale Swin Transformer.* Buildings.

Yuan Tian, Jingxuan Zhu, Huang Yao, et al. (2024). *Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention.* Applied Sciences.

Nan Zhou, Mingming Xu, Biaoqun Shen, et al. (2024). *ViT-UNet: A Vision Transformer Based UNet Model for Coastal Wetland Classification Based on High Spatial Resolution Imagery.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Gizatie Desalegn Taye, Zewdie Habtie Sisay, Genet Worku Gebeyhu, et al. (2024). *Thoracic computed tomography (CT) image-based identification and severity classification of COVID-19 cases using vision transformer (ViT).* Discover Applied Sciences.

Manal Abdullah Alohali, Nora El-Rashidy, Saad Alaklabi, et al. (2024). *Swin-GA-RF: genetic algorithm-based Swin Transformer and random forest for enhancing cervical cancer classification.* Frontiers in Oncology.

Zhenchang Gao, Shanshan Chen, Jinxian Huang, et al. (2024). *Real-time quantitative detection of hydrocolloid adulteration in meat based on Swin Transformer and smartphone..* Journal of Food Science.

Yufeng Du, Rongyun Zhang, Peicheng Shi, et al. (2024). *ST-LaneNet: Lane Line Detection Method Based on Swin Transformer and LaneNet.* Chinese Journal of Mechanical Engineering.

R. Tiwari, Himani Maheshwari, Vinay Gautam, et al. (2024). *CurrencyNet: A Vision Transformer-Based Approach for Indian Currency Note Classification with Optimizer Exploration.* 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT).

Chunlai Du, Yanhui Guo, and Yuhang Zhang (2024). *A Deep Learning-Based Intrusion Detection Model Integrating Convolutional Neural Network and Vision Transformer for Network Traffic Attack in the Internet of Things.* Electronics.

A. Chaurasia, H. C. Harris, P. W. Toohey, et al. (2024). *A generalised vision transformer-based self-supervised model for diagnosing and grading prostate cancer using histological images.* medRxiv.

Meryem Altin Karagöz, Özkan U. Nalbantoglu, and Geoffrey C. Fox (2024). *Residual Vision Transformer (ResViT) Based Self-Supervised Learning Model for Brain Tumor Classification.* arXiv.org.

Hyojin Lee, You Rim Choi, Hyun Kyung Lee, et al. (2025). *Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals.* npj Digit. Medicine.

Sezer Dümen, Esra Kavalcı Yılmaz, Kemal Adem, et al. (2024). *Performance of vision transformer and swin transformer models for lemon quality classification in fruit juice factories*. European Food Research and Technology.

Gazi Jannatul Ferdous, Khaleda Akhter Sathi, Md. Azad Hossain, et al. (2024). *SPT-Swin: A Shifted Patch Tokenization Swin Transformer for Image Classification*. IEEE Access.

Sara Akan, Songül Varli, and Mohammad Alfrad Nobel Bhuiyan (2024). *An enhanced Swin Transformer for soccer player reidentification*. Scientific Reports.

Pradeep Nahak, D. K. Pratihar, and A. K. Deb (2024). *Tomato maturity stage prediction based on vision transformer and deep convolution neural networks*. International Journal of Hybrid Intelligent Systems.

Yufei Han, Haoyuan Chen, Linwei Yao, et al. (2024). *MAT-VIT:A Vision Transformer with MAE-Based Self-Supervised Auxiliary Task for Medical Image Classification*. International Conference on Computer Supported Cooperative Work in Design.

Xiaoping Zhao, Jingjing Xu, Zhichen Lin, et al. (2024). *BiCFormer: Swin Transformer based model for classification of benign and malignant pulmonary nodules*. Measurement science and technology.

Tao Li, and Yi Zhang (2024). *A Contour-Aware Monocular Depth Estimation Network Using Swin Transformer and Cascaded Multiscale Fusion*. IEEE Sensors Journal.

Yancheng Wang, and Yingzhen Yang (2024). *Efficient Visual Transformer by Learnable Token Merging*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Haochen Qi, Xiangwei Kong, Zhibo Jin, et al. (2024). *A Vision-Transformer-Based Convex Variational Network for Bridge Pavement Defect Segmentation*. IEEE transactions on intelligent transportation systems (Print).

Shaojun Zhu, Guotao Chen, Hongguang Chen, et al. (2024). *Squeeze-and-excitation-attention-based mobile vision transformer for grading recognition of bladder prolapse in pelvic MRI images.*. Medical Physics (Lancaster).

Barsha Roy, Md. Farukuzzaman Faruk, Md Nazmul Islam, et al. (2024). *A Cutting-Edge Ensemble of Vision Transformer and ResNet101v2 Based Transfer Learning for the Precise Classification of Leukemia Sub-types from Peripheral Blood Smear Images.* International Conference on Electrical Engineering and Information Communication Technology.

Chunbao Wang, Xiangyu Wang, Zeyu Gao, et al. (2024). *Multiple serous cavity effusion screening based on smear images using vision transformer.* Scientific Reports.

Guangliang Pan, Qihui Wu, Bo Zhou, et al. (2024). *Spectrum Prediction With Deep 3D Pyramid Vision Transformer Learning.* IEEE Transactions on Wireless Communications.

Haiying Du, Jie Shen, Jing Wang, et al. (2024). *Vision transformer-based electronic nose for enhanced mixed gases classification.* Measurement science and technology.

Kevin Luo, and Ie-bin Lian (2024). *Building a Vision Transformer-Based Damage Severity Classifier with Ground-Level Imagery of Homes Affected by California Wildfires.* Fire.

Samy Abd El-Nabi, Ahmed F. Ibrahim, El-Sayed M. El-Rabaie, et al. (2025). *Driver Drowsiness Detection Using Swin Transformer and Diffusion Models for Robust Image Denoising.* IEEE Access.

Ebru Ergün (2025). *High precision banana variety identification using vision transformer based feature extraction and support vector machine.* Scientific Reports.

Muhammad Ahmed Mohsin, Muhammad Jazib, Zeeshan Alam, et al. (2025). *Vision Transformer Based Semantic Communications for Next Generation Wireless Networks.* 2025 IEEE International Conference on Communications Workshops (ICC Workshops).

Luella Marcos, Paul S. Babyn, and J. Alirezaie (2024). *Pure Vision Transformer (CT-ViT) with Noise2Neighbors Interpolation for Low-Dose CT Image Denoising.*. Journal of imaging informatics in medicine.

Xianhui Peng, Chenchen Xu, Peng Zhang, et al. (2024). *Computer vision classification detection of chicken parts based on optimized Swin-Transformer.* CyTA - Journal of Food.

Claudio Urrea, and Maximiliano Vélez (2024). *Enhancing Autonomous Visual Perception in Challenging Environments: Bilateral Models with Vision Transformer and Multilayer Perceptron for Traversable Area Detection.* Technologies.

Jinnian Zhang, Weijie Chen, Tanmayee Joshi, et al. (2024). *BAE-ViT: An Efficient Multimodal Vision Transformer for Bone Age Estimation.* Tomography.

Hira Saleem, Flora Salim, and Cormac Purcell (2024). *STC-ViT: Spatio Temporal Continuous Vision Transformer for Weather Forecasting.* Unpublished manuscript.

Yang Zhou, Cai Yang, Ping Wang, et al. (2024). *ViT-FuseNet: Multimodal Fusion of Vision Transformer for Vehicle-Infrastructure Cooperative Perception.* IEEE Access.

P. Lijin, M. Ullah, Anuja Vats, et al. (2024). *PolySegNet: improving polyp segmentation through swin transformer and vision transformer fusion.*. Biomedical Engineering Letters.

Lan Huang, Jiong Ma, Hui Yang, et al. (2024). *Research and implementation of multi-disease diagnosis on chest X-ray based on vision transformer.* Quantitative Imaging in Medicine and Surgery.

Chuanyu Chen, Yi Luo, Qiuyang Hou, et al. (2024). *A vision transformer-based deep transfer learning nomogram for predicting lymph node metastasis in lung adenocarcinoma.*. Medical Physics (Lancaster).

Mohammed Shahin, and Mohamed Deriche (2024). *A Novel Framework based on a Hybrid*

Vision Transformer and Deep Neural Network for Deepfake Detection. International Multi-Conference on Systems, Signals Devices.

Yang Xu, and Zuqiang Meng (2024). *Interpretable vision transformer based on prototype parts for COVID-19 detection.* IET Image Processing.

Joohyuk Park, Yong-Nam Oh, Yongjune Kim, et al. (2024). *Vision Transformer-Based Semantic Communications With Importance-Aware Quantization.* IEEE Internet of Things Journal.

O. Elharrouss, Y. Akbari, Noor Almaadeed, et al. (2025). *PDC-ViT : Source Camera Identification using Pixel Difference Convolution and Vision Transformer.* Neural computing applications (Print).

Yongqiang Du, Haoran Liu, Shengjie He, et al. (2024). *InViT: GAN Inversion-Based Vision Transformer for Blind Image Inpainting.* IEEE Access.

Qianyu Guo, Ziqing Yu, Jiaming Fu, et al. (2024). *Force-EvT: A Closer Look at Robotic Gripper Force Measurement with Event-Based Vision Transformer.* 2024 6th International Conference on Reconfigurable Mechanisms and Robots (ReMAR).

Kunpeng Zhang, Mengyan Lyu, Xinxin Guo, et al. (2024). *Temporal Shift Module-Based Vision Transformer Network for Action Recognition.* IEEE Access.

Lu Xu, Rui Shi, and Yijia Zhang (2025). *A Radio Frequency Sensor-Based UAV Detection and Identification System Using Improved Vision Transformer-Based Model.* IEEE Sensors Journal.

Yang Li, Doudou Zhang, and Jianli Xiao (2024). *A New Method for Vehicle Logo Recognition Based on Swin Transformer.* arXiv.org.

_to_visual_transformers13Introduction to Visual Transformerssection.1 _1_the_rise_of_transfor
Rise of Transformers in Deep Learningsubsection.1.1 _2_motivation_for_vision_transformers:_
for Vision Transformers: Beyond CNNssubsection.1.2 _3_scope__and__structure_of_the_review1
and Structure of the Reviewsubsection.1.3 _vision_transformer_architectures__and__early_optim
Vision Transformer Architectures and Early Optimizationssection.2 _1_the_original_vision_transform
Original Vision Transformer (ViT) Paradigmsubsection.2.1 _2_data-efficient_training__and__di
Efficient Training and Distillationsubsection.2.2 _3_enhancing_stability__and__tokenization_
Stability and Tokenization for Deeper Modelssubsection.2.3 __and__efficient_vision_transforme
and Efficient Vision Transformer Architecturessection.3 _1_shifted_window-based_attention_for_hie
Window-Based Attention for Hierarchical Processingsubsection.3.1 _2_pyramid_structures
scale_feature_representation3.219Pyramid Structures for Multi-Scale Feature Representationsubsection.3.2 _3_multi-scale__and__efficient_attention_mechanisms3.322Multi-Scale and Efficient Attention Mechanismssubsection.3.3 _learning_paradigms_for_vision_tran
Supervised Learning Paradigms for Vision Transformerssection.4 _1_masked_image_modeling_(mim)
Image Modeling (MIM) for Representation Learningsubsection.4.1 _2_self-distillation__and__contrastive_learning_without_labels4.227Self-Distillation and Contrastive Learning without Labelssubsection.4.2 _3_scaling_self-supervised_vits_to_foundation_
Self-Supervised ViTs to Foundation Modelssubsection.4.3 _architectures__and__beyon
attention534Hybrid Architectures and Beyond Self-Attentionsection.5 _1_integrating_convolutional_i
Convolutional Inductive Biases into Transformerssubsection.5.1 _2_rethinking_token_mixing:_a
attention5.237Rethinking Token Mixing: Alternatives to Self-Attentionsubsection.5.2 _3_modernizing_cnns_with_vision_transformer_design_principles5.339Modernizing CNNs with Vision Transformer Design Principlessubsection.5.3 _4_efficient__and__lightweight_hybrid_des
and Lightweight Hybrid Designs for Deploymentsubsection.5.4 __and__domain-specific_adaptations
and Domain-Specific Adaptations of Visual Transformerssection.6 _1_object_detection__and__sema
Detection and Semantic Segmentationsubsection.6.1 _2_medical_image_analysis__and__3d_s
Image Analysis and 3D Segmentationsubsection.6.2 _3_remote_sensing__and__environmental_
Sensing and Environmental Monitoringsubsection.6.3 _4_lightweight__and__real-time_applications6.449Lightweight and Real-time Applicationssubsection.6.4 _di-