Enhancing Computational Efficiency in Vision Transformers: A Survey on Token Compression and Model Optimization

Abstract

Vision Transformers (ViTs) have emerged as a pivotal advancement in machine learning, fundamentally altering the processing of visual data through token compression, multimodal learning, and model optimization. This survey explores the critical role of computational efficiency in ViTs, emphasizing the need for advanced token compression strategies to mitigate the quadratic complexity of self-attention mechanisms. Techniques such as Joint Token Pruning Squeezing and Content-aware Token Sharing exemplify efforts to reduce computational demands while maintaining performance. The integration of multimodal learning enhances ViTs' capabilities by effectively combining diverse data types, improving performance across tasks. Model optimization strategies, including quantization and knowledge distillation, further refine ViT architectures, enabling deployment in resource-constrained environments. Despite these advancements, challenges remain, particularly in achieving efficient multimodal integration and addressing computational intensity. Future research directions include exploring mixed quantization strategies, dynamic token sharing, and optimizing attention mechanisms. By focusing on these areas, the potential of ViTs can be maximized, ensuring their applicability across diverse applications. This survey underscores the importance of continuous innovation in enhancing the efficiency and effectiveness of Vision Transformers, positioning them at the forefront of machine learning advancements.

1 Introduction

1.1 Significance of Computational Efficiency in ViTs

The computational efficiency of Vision Transformers (ViTs) is essential for their deployment in applications with limited resources, such as mobile and edge computing platforms. Despite outperforming Convolutional Neural Networks (CNNs) in various computer vision tasks, ViTs encounter significant computational and memory overheads that restrict their broader applicability. The quadratic complexity of the Multi-head Self-Attention (MSA) mechanism poses a notable bottleneck, particularly with high-resolution images. This complexity, coupled with advanced text encoding methods used in Transformer models, escalates computational demands, making real-time applications impractical on resource-constrained devices. The need for efficient multimodal representations further complicates this, as integration techniques—like late fusion, early fusion, or sketch—impact performance and resource utilization in Natural Language Processing tasks. Therefore, optimizing these processes is imperative for deploying effective models in resource-limited environments [1, 2].

To mitigate inefficiencies in token processing, advanced token compression strategies are crucial for maintaining performance while reducing computational demands [3]. The high computational costs associated with ViTs not only limit their application scope but also hinder performance in detail-oriented tasks, such as multi-view 3D detection and semantic segmentation. Addressing these

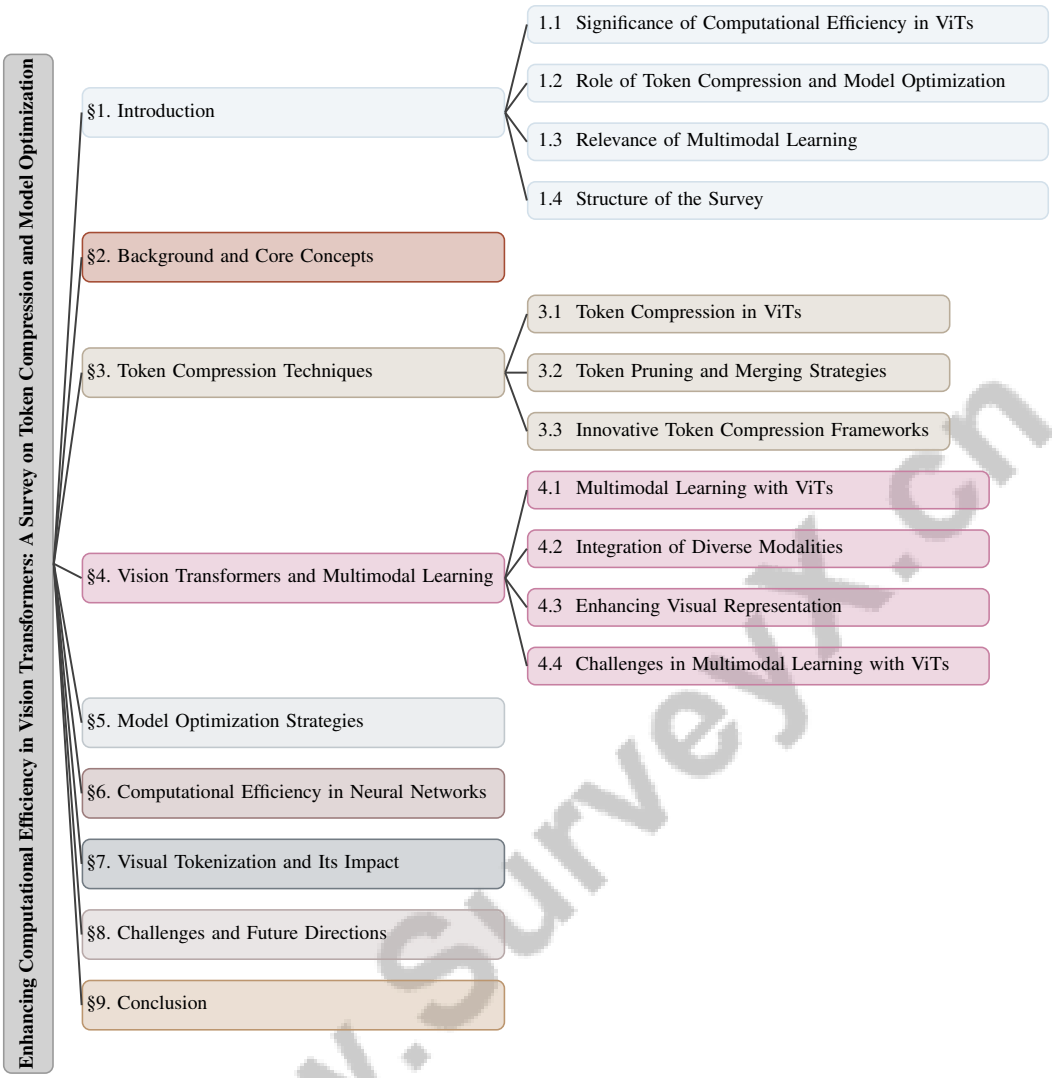


Figure 1: chapter structure

inefficiencies is vital for enhancing the applicability of ViTs across diverse domains, including those traditionally dominated by CNNs [4].

Optimizing computational efficiency is key to ensuring the robust performance and adaptability of ViTs, facilitating their widespread adoption in real-world scenarios with constrained resources [5]. This involves developing efficient training schemes and transformer compression methods that reduce memory and computational costs without sacrificing performance [6]. By enhancing computational efficiency, ViTs can better leverage their potential across various applications, broadening their practical utility and impact.

1.2 Role of Token Compression and Model Optimization

Token compression and model optimization are critical for improving the computational efficiency of Vision Transformers (ViTs), which struggle with intensive token processing demands. Techniques such as token pruning and merging are vital for reducing token counts, thereby alleviating computational load while preserving model performance. The TokenCompression3D (ToC3D) method exemplifies this by employing history object queries as high-quality foreground priors for efficient token compression and resource allocation [7]. The Joint Token Pruning Squeezing (TPS) method further compresses vision transformers aggressively while retaining essential information by consolidating features of pruned tokens into reserved ones [8].

AI-generated, for reference only.

Model optimization strategies refine ViT architectures to enhance efficiency. For instance, the Convolutional Additive Self-attention Vision Transformers (CAS-ViT) introduces a novel additive similarity function and an efficient Convolutional Additive Token Mixer (CATM), improving performance while reducing computational overhead [9]. Additionally, M2-ViT employs a two-level mixed quantization strategy, incorporating both mixed precision and mixed quantization schemes to significantly enhance hybrid ViTs' efficiency [10]. The Less-Attention Vision Transformer (LaViT) optimizes efficiency by computing attention operations selectively, leveraging previously calculated scores [11].

Innovative frameworks like CAIT integrate asymmetric token merging and consistent dynamic channel pruning to achieve high accuracy, fast inference speed, and favorable transferability [12]. The Local InFormation Enhancer (LIFE) module integrates local information to refine token embeddings, contributing to model optimization and computational efficiency [13].

Moreover, adapting pre-training techniques aligns with token compression and model optimization efforts to enhance ViT efficiency. The Self-Supervised Auxiliary Task (SSAT) method involves training ViTs alongside an auxiliary task that reconstructs missing pixels, enabling the model to learn more robust features and improve efficiency [14]. The survey by [4] emphasizes the significance of these strategies in overcoming computational constraints.

These advancements in token compression and model optimization reflect a concerted effort to refine token management and model architecture, ensuring ViTs operate efficiently across various platforms and applications. The integration of attention mechanisms into machine vision tasks, despite challenges such as data requirements and training times, underscores the importance of these strategies in enhancing performance [15].

1.3 Relevance of Multimodal Learning

Multimodal learning is crucial in the context of Vision Transformers (ViTs), enhancing the integration and representation of diverse data types to improve model performance across various tasks. The ability to combine modalities, such as visual and linguistic data, enables ViTs to address gaps in commonsense, factual, and temporal knowledge, which are essential for visiolinguistic tasks [16]. This integration is particularly significant in applications requiring the understanding and processing of complex, heterogeneous data sources, such as computer vision models incorporating various visual cues [17].

The evolution of multimodal representation learning has been marked by advancements in pretraining methodologies and applications across diverse domains, highlighting the transformative impact of multimodal approaches [18]. Since the 1980s, the advantages of combining different data types, like acoustic and visual information, have been well-documented, demonstrating multimodal learning's potential to enhance feature representation and model performance [1].

Incorporating multimodal data from sources like LiDAR offers significant opportunities to improve classification performance through enriched feature representation, critical for tasks involving complex environmental data [19]. Furthermore, the integration of vision and language modalities is essential for enhancing performance in vision-language tasks, illustrating multimodal learning's broader relevance in improving ViTs' efficacy [20].

Multimodal learning frameworks also contribute to a deeper understanding of consumer preferences by integrating diverse data types, enabling more accurate and efficient decision-making processes [21]. Collectively, these advancements underscore multimodal learning's critical role in expanding Vision Transformers' capabilities, facilitating their application in fields requiring sophisticated data integration and processing.

1.4 Structure of the Survey

This survey provides a comprehensive analysis of methodologies and advancements aimed at enhancing computational efficiency in Vision Transformers (ViTs) through token compression and model optimization. It begins with an introduction emphasizing computational efficiency's critical importance in ViTs, highlighting the pivotal roles of advancements in token compression and model optimization techniques—such as quantization, low-rank approximation, knowledge distillation, and pruning—in enhancing model performance while addressing high computational and memory require-

ments that hinder deployment in resource-constrained environments [12, 22, 23]. The introduction also underscores the relevance of multimodal learning in the context of ViTs.

The second section elaborates on core concepts, providing an overview of Vision Transformers, efficient neural networks, and the broader context of computational efficiency in machine learning frameworks, establishing a foundational understanding for subsequent discussions.

The third section explores various token compression techniques utilized in ViTs, focusing on methods that enhance computational efficiency, including specific strategies like token pruning and merging, as well as recent innovative frameworks developed for token compression.

The fourth section delves into the role of Vision Transformers in multimodal learning, analyzing how they facilitate the integration of diverse data modalities and enhance visual representation, while addressing challenges encountered in implementing multimodal learning with ViTs.

The fifth section discusses model optimization strategies, examining quantization techniques, knowledge distillation, and dynamic and adaptive methods aimed at improving neural network efficiency.

The sixth section emphasizes the importance of computational efficiency in neural networks, particularly concerning ViTs, analyzing real-world applications and trade-offs, and exploring future directions in efficient neural network design.

The seventh section analyzes the process of visual tokenization, highlighting its impact on model performance and recent innovations in this area.

The survey concludes with a discussion on challenges and future directions in enhancing computational efficiency in ViTs and related models, recognizing current obstacles in multimodal representation learning while investigating promising research avenues that could advance technique selection, data integration, and the application of knowledge graphs to improve model performance and interpretability [16, 1, 2].

This structured approach facilitates a comprehensive examination of critical themes and advancements in optimizing Vision Transformers (ViTs), focusing on various model compression techniques, such as quantization, low-rank approximation, knowledge distillation, and pruning. By evaluating these methods, the study provides valuable insights into enhancing computational efficiency while maintaining model accuracy, ultimately addressing the challenges posed by high computational and memory demands of ViTs and paving the way for their effective deployment in resource-constrained environments, including edge computing devices. Additionally, analyzing ViTs under different supervision methods reveals their flexibility in learning diverse behaviors and processing information, further contributing to their optimization for various applications [24, 23].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Vision Transformers (ViTs) Overview

Vision Transformers (ViTs) have transformed visual recognition by effectively capturing long-range dependencies, often missed by traditional Convolutional Neural Networks (CNNs) [25]. Unlike CNNs, which emphasize local feature extraction, ViTs utilize a self-attention mechanism treating images as sequences of patches, enabling the capture of global contextual information vital for image classification and segmentation [5]. However, the quadratic complexity of their self-attention mechanism, scaling as $O(n^2) with token sequence length n, poses challenges for efficient handling of long sequences and limits deployment on resource-constrained devices [26]. Hybrid models combining convolutional operations with transformer architectures have been p$

ViTs hold promise in applications like 3D point cloud processing and multi-view 3D detection, where token compression techniques enhance efficiency [5]. This progression underscores the limitations of existing models, often specialized and restricted to single modalities and tasks, highlighting the need for versatile and scalable machine learning solutions. The architecture of ViTs, with multiple self-attention layers, reduces inductive biases and enhances parameter efficiency, achieving state-of-the-art accuracy in image classification and object detection. Studies show their robustness against common corruptions and distribution shifts, often surpassing CNNs in rigorous evaluations. Research into model compression techniques, such as quantization and pruning, continues to optimize ViTs for resource-constrained environments, broadening their practical applicability in areas like

edge computing and healthcare [27, 23]. Efforts persist to address ViTs' computational challenges, maximizing their potential in computer vision applications.

2.2 Efficient Neural Networks

Efficient neural networks aim to optimize computational resources while maintaining performance, a principle crucial for Vision Transformers (ViTs) due to their computational demands. Reducing or eliminating non-linear operations, such as the Softmax function, is key to achieving efficiency. The TriO-ViT framework exemplifies this by removing such operations to facilitate post-training quantization, enhancing computational efficiency [28]. The development of specialized quantization frameworks like P2-ViT highlights the importance of efficient network design, introducing a novel post-training quantization and acceleration framework for fully quantized ViTs, which integrates a dedicated quantization scheme with a customized accelerator, improving efficiency without sacrificing performance [29].

Enhancing neural network efficiency also involves simplifying architectures by replacing complex layers with computationally efficient alternatives. For instance, substituting the attention layer with a shift operation offers a zero-parameter and zero-computation method that maintains model effectiveness while reducing computational burden [30]. Principles of efficient neural networks advance ViTs by enabling innovative training methodologies, such as automated progressive learning, accelerating training processes without compromising performance, and promoting architectural diversity to reduce redundancy within models. These advancements not only cut computational costs but also enhance the models' capacity to capture distinct image features, improving performance in various vision tasks [31, 22, 32, 33]. By focusing on quantization, architectural simplification, and eliminating computational bottlenecks, researchers can expand the applicability of ViTs in resource-constrained environments.

2.3 Computational Efficiency in Machine Learning

Computational efficiency is a critical concern in machine learning as models grow increasingly complex and data-intensive. Vision Transformers (ViTs) face inherent computational challenges due to their intricate architectures and extensive data requirements for effective training and deployment [4]. The scalability of ViTs across different model sizes and training data volumes is essential, as benchmarks assess representation quality and performance relative to model and dataset size [34]. ViTs' data-hungry nature necessitates large-scale datasets for optimal performance, impacting efficiency and applicability in domains with limited data [35]. Training inefficiencies often lead to suboptimal use of examples, affecting model efficiency [36].

Integrating multimodal data into ViTs adds computational complexities, particularly in achieving effective cross-modal alignment and fusion, crucial for enhancing computational efficiency [37]. Inefficiencies in processing multimodal data streams increase computational costs and energy consumption, challenging low-resource settings [38]. Adversarial robustness is another dimension of computational efficiency; the vulnerability of ViTs to adversarial attacks can undermine effectiveness. Strategies like Self-Ensemble and Token Refinement enhance token discriminative capacity and adversarial sample transferability, fortifying models against attacks [39]. Understanding ViTs' Lipschitz continuity provides a theoretical framework linking adversarial robustness to the Cauchy Problem, informing strategies for enhancing computational efficiency [40].

Grid-like noise artifacts in ViTs' feature maps degrade performance in tasks like semantic segmentation and object detection [41]. Addressing these artifacts is crucial for improving computational efficiency and accuracy in complex visual tasks. Pursuing computational efficiency in machine learning involves model scalability, multimodal data integration, adversarial robustness, and efficient training strategies. These initiatives are essential for deploying advanced models like ViTs across various applications and environments, enhancing interpretability, optimizing performance, and improving efficiency. Frameworks decomposing ViT representations elucidate how model components capture specific image features, enabling applications like image retrieval and spurious correlation mitigation. Innovations like the Next-ViT model address latency and accuracy trade-offs, making ViTs competitive with CNNs in industrial settings. Moreover, interpretability-aware training methods enhance understanding of model predictions, while self-supervised learning techniques improve

performance with limited data, collectively maximizing ViTs' potential in real-world scenarios [42, 14, 22, 43].

In recent years, the exploration of token compression techniques in Vision Transformers (ViTs) has gained significant attention due to the increasing demand for efficient computational methods in machine learning. As illustrated in Figure 2, the hierarchical structure of these techniques reveals a comprehensive categorization that encompasses various methods, challenges, and advancements. The figure delineates three main areas: token compression in ViTs, token pruning and merging strategies, and innovative token compression frameworks. Each category not only highlights specific techniques and mechanisms but also underscores their contributions to enhanced computational efficiency and performance, particularly in resource-constrained environments. This structured overview facilitates a better understanding of the landscape of token compression, providing a foundation for further exploration and development in this vital area of research.

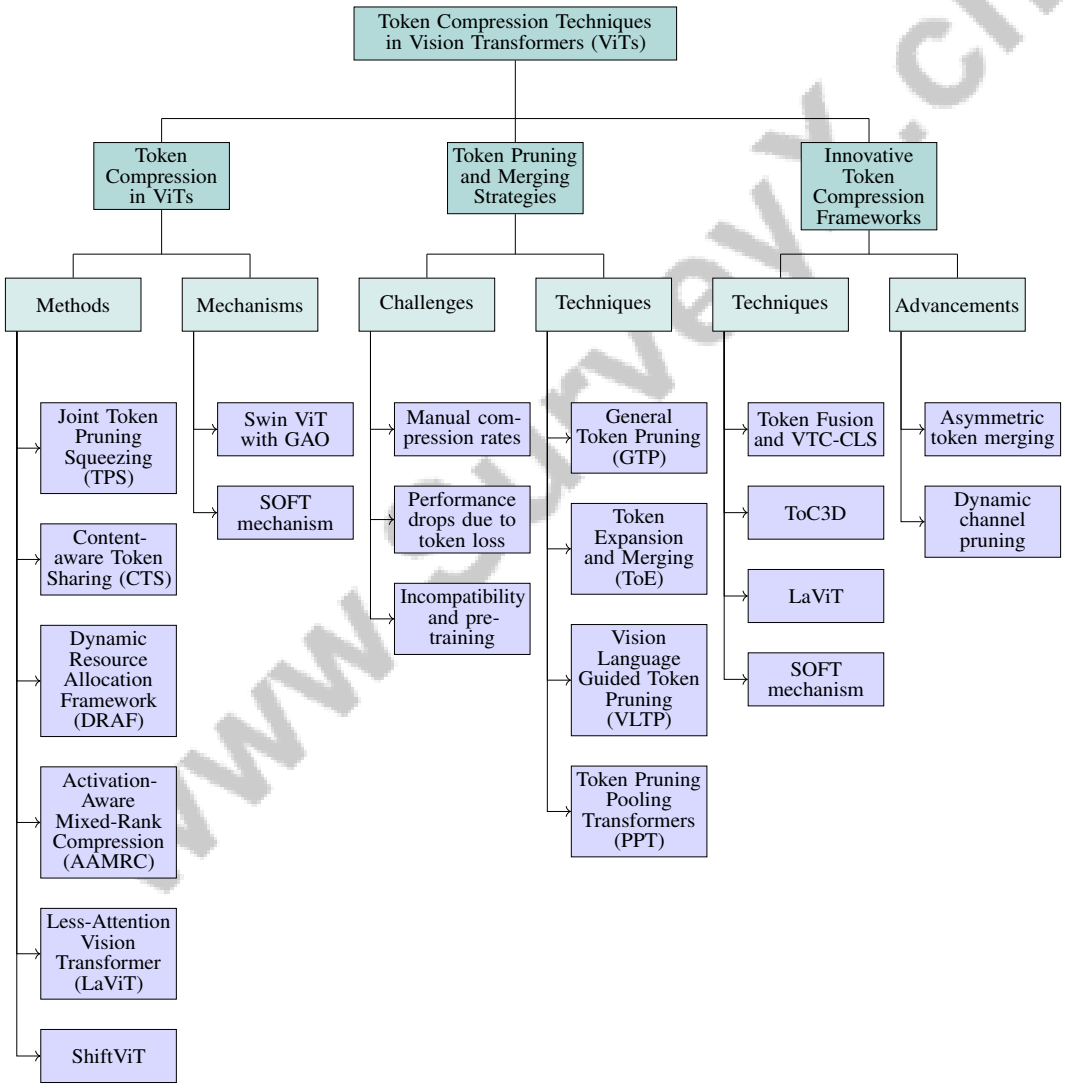


Figure 2: This figure illustrates the hierarchical structure of token compression techniques in Vision Transformers (ViTs), categorizing methods, challenges, and advancements across three main areas: token compression in ViTs, token pruning and merging strategies, and innovative token compression frameworks. Each category highlights specific techniques and mechanisms contributing to enhanced computational efficiency and performance in resource-constrained environments.

3 Token Compression Techniques

3.1 Token Compression in ViTs

Method Name	Computational Efficiency	Token Management Techniques	Performance Preservation
TPS[8]	Aggressive Token Pruning	Pruning And Squeezing	Minimal Performance Drop
CTS[3]	Content-aware Token	Token Sharing	Segmentation Quality
DRAF[6]	Processing Efficiency	Token Pruning	Enhance Performance
AAMRC[44]	Parameter Count Reduction	Token Pruning	Performance Retention
LaViT[11]	Computational Efficiency	Token Pruning	State-of-the-art
ShiftViT[30]	Zero-computation Method	-	Comparable Performance
SOFT[26]	Linear Complexity	Low-rank Decomposition	Improving Accuracy

Table 1: Comparison of various token compression methods in Vision Transformers (ViTs), highlighting their computational efficiency, token management techniques, and performance preservation capabilities. The table provides a succinct overview of innovative strategies such as token pruning, content-aware token sharing, and low-rank decomposition, which collectively enhance the efficiency and effectiveness of ViTs in computer vision tasks.

Token compression in Vision Transformers (ViTs) is crucial for improving computational efficiency by reducing token count while preserving performance. ViTs' ability to capture long-range dependencies is pivotal for achieving state-of-the-art results in computer vision tasks [45]. Innovative methods like Joint Token Pruning Squeezing (TPS) integrate token pruning with information squeezing, maintaining performance while reducing token count [8]. Content-aware Token Sharing (CTS) enhances efficiency in semantic segmentation by allowing similar patches to share tokens, optimizing processing through semantic similarity [3].

As illustrated in Figure 3, which depicts key methodologies for token compression in ViTs, various techniques such as token pruning and squeezing, low-rank approximations, and attention optimization strategies are highlighted. These methodologies collectively enhance computational efficiency while ensuring that performance remains intact. Table 1 presents a comprehensive comparison of diverse token compression methodologies used in Vision Transformers, emphasizing their impact on computational efficiency and performance preservation.

The Dynamic Resource Allocation Framework (DRAF) uses real-time data analysis for dynamic resource allocation, enhancing token processing in real-time environments [6]. Activation-Aware Mixed-Rank Compression (AAMRC) employs selective low-rank weight tensor approximations to reduce parameter counts with minimal reconstruction error [44]. Less-Attention Vision Transformer (LaViT) optimizes token processing by transforming stored attention scores from earlier layers, capturing dependencies with fewer computations [11]. ShiftViT replaces traditional attention with a shift operation, simplifying the model while maintaining performance comparable to the Swin Transformer [30].

The Swin ViT model with GAO offers a benchmark for optimization effectiveness [25]. The SOFT mechanism, using a Gaussian kernel instead of softmax normalization, facilitates low-rank matrix decomposition, further optimizing token processing [26]. These advancements highlight the importance of innovative token management approaches in ViTs, enhancing computational efficiency and enabling their application in resource-constrained environments with significant speedups and minimal performance degradation across vision tasks [46, 47, 48, 22, 49].

3.2 Token Pruning and Merging Strategies

Token pruning and merging strategies are vital for enhancing ViTs' computational efficiency by reducing token processing without sacrificing performance. Techniques dynamically drop image tokens, focusing resources on relevant patches [50]. A challenge is the reliance on manually set compression rates, risking loss of important tokens and degrading performance [51]. The TPS method addresses this by pruning tokens and squeezing information into reserved tokens through a matching and fusing process, ensuring information retention [8]. Table 2 presents a comprehensive comparison of token pruning and merging strategies that enhance the computational efficiency of Vision Transformers (ViTs) without compromising performance.

Structured spatial compression and unstructured token pruning face limitations like incompatibility and time-consuming pre-training [52]. General Token Pruning (GTP) achieves an efficient trade-off

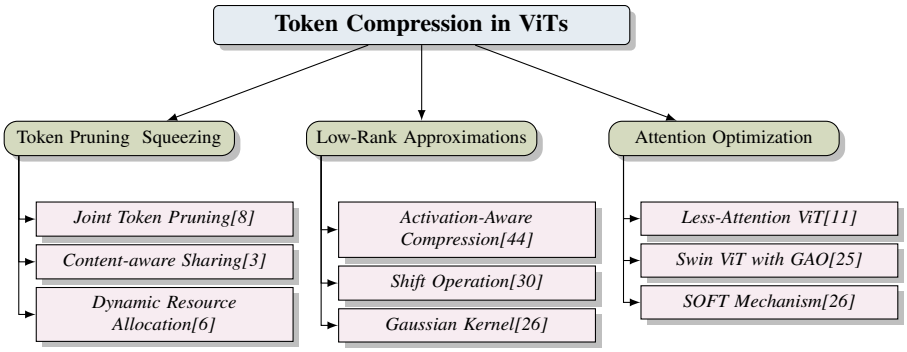


Figure 3: This figure illustrates key methodologies for token compression in Vision Transformers (ViTs), highlighting token pruning and squeezing techniques, low-rank approximations, and attention optimization strategies to enhance computational efficiency while maintaining performance.

Method Name	Efficiency Techniques	Information Retention	Performance Trade-offs
IVT[50]	Dynamic Selection Tokens	All Tokens Retained	Balance Efficiency Accuracy
DiffRate[51]	Automatic Compression Rates	Important Tokens Preserved	Minimal Accuracy Drop
TPS[8]	Token Pruning	Information Squeezing	Minimal Performance Drop
Evo-ViT[52]	Slow-fast Updating	Spatial Structure	Maintaining Performance
GTP[53]	Token Propagation	Information Integrity	Minimal Accuracy Loss
ToE[54]	Token Selection	Retain Information	Training Consistency
R-MeeTo[55]	Token Merging	Merging Tokens	Recover Lost Performance
SPViT[56]	Dynamic Attention-based	Token Packaging Technique	Latency-aware Training
CAIT[12]	Asymmetric Token Merging	Spatial Integrity	Accuracy Loss
H-ViT[57]	Multi-stage Architecture	Context-aware Embeddings	Improved Performance Grading

Table 2: Comparison of various token pruning and merging strategies employed in Vision Transformers (ViTs), detailing the efficiency techniques, information retention capabilities, and performance trade-offs. The table highlights methods such as IVT, DiffRate, and TPS, showcasing their unique approaches to optimizing computational efficiency while maintaining model accuracy.

between model efficiency and performance, allowing faster inference with minimal accuracy loss [53]. Token Expansion and Merging (ToE) accelerates training by selectively expanding and merging tokens, preserving feature distributions [54].

Despite advancements, token reduction methods often result in performance drops due to the loss of informative tokens [55]. This issue is exacerbated by dynamic image features, inadequately addressed by current pruning methods, leading to suboptimal pruning rates and accuracy deterioration [56]. Compression methods dropping redundant tokens or channels separately face challenges in transferring compressed models to tasks requiring spatial structure [12].

Hierarchical vision transformers' three-stage process, aggregating information from patch-level, region-level, and slide-level transformers, illustrates comprehensive token management enhancing slide-level representation [57]. These advanced techniques significantly enhance ViTs' computational efficiency, making them viable for resource-constrained environments, achieving substantial speedups across vision tasks while maintaining competitive performance levels. For example, the Vision Language Guided Token Pruning (VLTP) framework reduces computational costs by approximately 25

3.3 Innovative Token Compression Frameworks

Recent advancements in token compression frameworks enhance ViTs' computational efficiency by optimizing token management. These frameworks use strategies like token pruning and merging to minimize tokens processed during inference while maintaining performance. Techniques such as Token Fusion and VTC-CLS leverage input tokens' unique characteristics, achieving speed improvements and reduced computational costs across vision and language tasks [48, 59, 46, 60, 2].

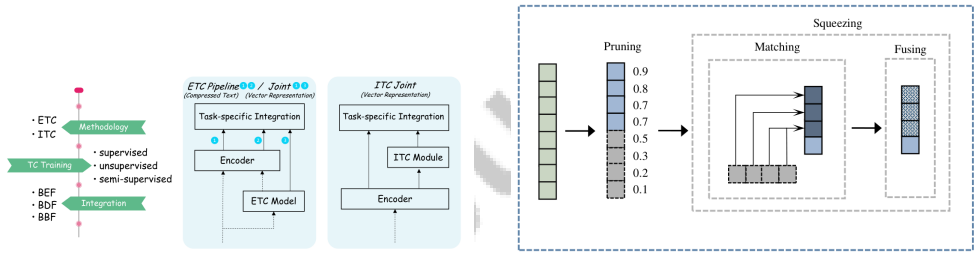
ToC3D integrates 3D motion information into token compression, outperforming methods lacking 3D awareness [7]. LaViT reduces costs by re-parameterizing attention scores from earlier layers, mitigating attention saturation compared to traditional architectures [11]. The SOFT mechanism

Method Name	Token Management Strategies	Efficiency and Performance	Framework Adaptability
ToC3D[7]	Token Compression Process	Significant Speedups	Real-time Applications
LaViT[11]	Token Pruning, Merging	Computational Efficiency, Performance	Diverse Applications, Environments
SOFT[26]	Gaussian Kernel	Linear Complexity	Additional Applications
ETC/ITC[2]	-	Improve Language Representation	Various Nlp Tasks
TPS[58]	Token Pruning & Squeezing	Superior Performance	Hybrid Vits

Table 3: Comparative Analysis of Token Compression Methods in Vision Transformers: This table presents a detailed overview of various token management strategies employed by different methods, highlighting their efficiency, performance, and adaptability across frameworks. The methods include ToC3D, LaViT, SOFT, ETC/ITC, and TPS, each demonstrating unique approaches to optimize computational efficiency and applicability in diverse environments.

replaces softmax with a Gaussian kernel, allowing low-rank approximation of the self-attention matrix, achieving linear complexity [26].

These innovative frameworks represent substantial advancements in ViTs' token compression techniques. Methodologies like asymmetric token merging and dynamic channel pruning enhance efficiency while maintaining performance and transferability across vision tasks. Proposed approaches demonstrate impressive speedups in inference time while achieving comparable accuracy on benchmarks like ImageNet and ADE20k, addressing the challenges of ViTs' computational costs in resource-constrained environments [61, 62, 46, 12, 22]. By focusing on strategic token management and leveraging advanced methodologies, these approaches ensure ViTs operate efficiently across diverse applications and environments. Table 3 provides a comprehensive comparison of innovative token compression frameworks, elucidating their respective strategies, efficiency, performance, and adaptability in enhancing Vision Transformers.



(a) Joint Task-Specific Integration for Text Compression and Representation[2]

(b) Squeezing and Fusing: A Deep Learning Approach to Reduce Model Size[58]

Figure 4: Examples of Innovative Token Compression Frameworks

As shown in Figure 4, innovative frameworks have emerged to enhance text representation and reduce model size, exemplified by two distinct approaches. The "Joint Task-Specific Integration for Text Compression and Representation" method integrates text compression with task-specific requirements through a structured pipeline encompassing End-to-End Text Compression (ETC), Integrating Text Compression (ITC), and Task-Independent Text Compression (TC Training). Conversely, the "Squeezing and Fusing" approach highlights a deep learning strategy aimed at reducing model size through "Squeezing" and "Fusing" stages, which selectively removes less critical layers while maintaining efficacy. Both frameworks illustrate cutting-edge advancements in token compression, offering solutions that balance efficiency and performance in text processing and model optimization [2, 58].

4 Vision Transformers and Multimodal Learning

4.1 Multimodal Learning with ViTs

Vision Transformers (ViTs) have advanced multimodal learning by leveraging attention mechanisms to integrate diverse data types, capturing long-range dependencies crucial for cross-modal interactions and complex data management [63, 5]. The STViT method focuses on preserving high-level semantic representation, essential for efficient multimodal learning [64]. The LIFE module demonstrates

adaptability by enhancing ViT performance on small datasets across tasks like classification and object detection [13]. The 4M framework exemplifies ViTs' flexibility and improved performance in generative tasks by managing multiple modalities within a single model [5].

Various supervision methods influence ViTs' attention mechanisms and overall performance, with certain strategies like contrastive self-supervised training yielding competitive features for part-level recognition. Consistent behaviors across different techniques, such as Offset Local Attention Heads, highlight ViTs' flexibility in processing information based on training paradigms [24, 22, 65]. This adaptability is vital for tailoring ViT architectures to specific multimodal tasks. Recent studies show ViTs' ability to learn diverse behaviors under various supervision methods, enhancing their interpretability and decision-making processes [24, 22, 42, 65].

4.2 Integration of Diverse Modalities

Integrating diverse data modalities in Vision Transformers (ViTs) is crucial for processing complex information across domains. ViTs use attention mechanisms to combine visual, auditory, and textual inputs into coherent representations, benefiting visiolinguistic tasks and complex scene understanding [17]. Research categorizes integration strategies into data representation, fusion methods, multitask learning, alignment, transfer learning, and zero-shot learning, each offering insights into how ViTs process diverse modalities. Data representation encodes different data types for effective cross-modal interactions, while fusion methods combine representations to enhance model performance.

Multitask learning and alignment are vital for scenarios requiring multiple tasks or data alignment from different modalities, essential in applications like autonomous driving, where integrating visual, auditory, and sensor data is critical for decision-making. Transfer and zero-shot learning enhance ViTs' adaptability by leveraging knowledge from one task to improve performance on others, reducing retraining needs. Techniques like Self-Supervised Auxiliary Tasks (SSAT) optimize ViTs by enabling simultaneous learning from primary and auxiliary tasks, enhancing performance and reducing environmental impact [66, 22, 14].

To illustrate these concepts, Figure 5 provides a visual representation of the integration and enhancement strategies for diverse modalities in ViTs. This figure highlights key approaches such as data representation, fusion methods, and learning enhancements like transfer and zero-shot learning, along with their applications in fields like image retrieval and visual reasoning. Integrating diverse modalities marks a significant advancement in machine learning, enabling models to process complex, heterogeneous data while leveraging insights from various modalities. Recent research highlights decomposing and interpreting image representations through text, enhancing ViTs' interpretability and applications like image retrieval and visual reasoning. Incorporating knowledge graphs into multimodal frameworks addresses commonsense understanding gaps, improving decision-making across domains [28, 14, 42, 22, 16]. By employing advanced attention mechanisms and fusion strategies, ViTs drive progress in multimodal learning, providing robust solutions for diverse applications.

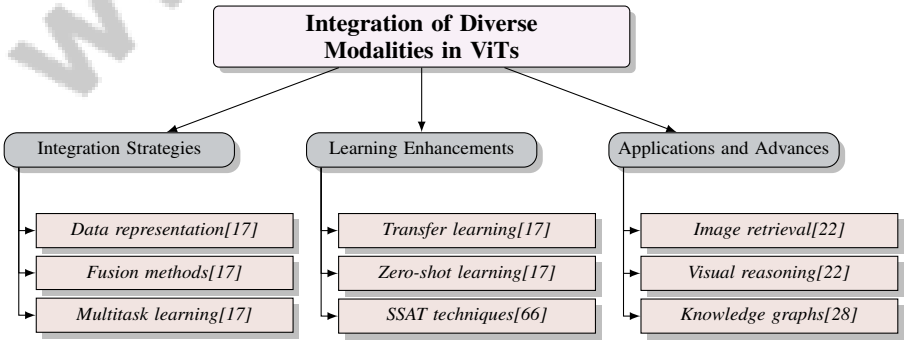


Figure 5: This figure illustrates the integration and enhancement strategies for diverse modalities in Vision Transformers (ViTs), highlighting key approaches such as data representation, fusion methods, and learning enhancements like transfer and zero-shot learning, along with their applications in fields like image retrieval and visual reasoning.

4.3 Enhancing Visual Representation

Enhancing visual representation in Vision Transformers (ViTs) involves leveraging their architecture to improve visual data processing. ViTs capture global contextual information through self-attention mechanisms, effectively modeling long-range dependencies within image data [63]. This enables ViTs to surpass traditional approaches in tasks requiring comprehensive image understanding, such as semantic segmentation and object detection [5].

Key strategies include integrating local information with global context. The Local InFormation Enhancer (LIFE) module enhances ViT performance by combining local feature details with global representations, improving the model's ability to capture fine-grained visual details [13]. This is particularly beneficial in applications requiring detailed visual analysis, such as medical imaging. Innovative token processing techniques, like Content-aware Token Sharing (CTS), refine visual representation by reducing redundancy and emphasizing semantically relevant features [3]. Hybrid models combining convolutional operations with transformer architectures utilize convolutional layers for local feature extraction and transformers for global context modeling, enhancing performance in visual recognition tasks [9].

Recent advancements in ViTs have expanded their capabilities in diverse visual tasks, including texture recognition and image retrieval. By employing techniques like self-supervised learning and novel frameworks for decomposing model components, ViTs capture distinct image features and optimize performance, even with limited data. This versatility broadens their applicability across domains, showcasing superior efficiency compared to traditional CNNs, particularly in tasks involving complex textures and dynamic visual content [22, 14, 67].

4.4 Challenges in Multimodal Learning with ViTs

Multimodal learning with Vision Transformers (ViTs) presents challenges due to the complexity of integrating diverse data modalities and associated computational demands. Aligning and fusing heterogeneous data types, such as visual, auditory, and textual information, is essential for coherent representations but increases computational overhead, affecting efficiency and scalability [17]. Data requirements for training ViTs in multimodal contexts are challenging, as large-scale, annotated datasets encompassing multiple modalities are scarce, limiting performance and generalization [35].

The computational intensity of ViTs, exacerbated by integrating multiple modalities, poses a significant hurdle. The quadratic complexity of the self-attention mechanism leads to high computational costs, especially when processing high-dimensional multimodal data [26]. This restricts ViTs' deployment in real-time applications and resource-constrained environments, necessitating more efficient architectures and token compression techniques. Interpretability in multimodal learning remains a challenge, as understanding how different modalities contribute to decision-making is critical for trust in applications like healthcare [5]. Ensuring ViTs' robustness against adversarial attacks in multimodal settings is crucial for reliability in safety-critical applications [39].

As illustrated in Figure 6, the primary challenges in multimodal learning with ViTs can be categorized into three main areas: data integration, computational complexity, and model robustness. Each category highlights specific issues such as the fusion of heterogeneous data, the high computational costs of self-attention, and the need for robust and interpretable models. Addressing these challenges requires continuous research and innovation in developing advanced multimodal learning frameworks, implementing data augmentation techniques, and designing efficient architectures to optimize computational resources. These efforts aim to improve ViTs' scalability, efficiency, and robustness, enabling successful application across domains like computer vision, healthcare, and multimodal sentiment analysis [27, 23, 68, 18].

5 Model Optimization Strategies

5.1 Quantization Techniques

Quantization techniques are pivotal for improving the computational efficiency of Vision Transformers (ViTs) by reducing the precision of model parameters and activations. This approach decreases computational and memory demands, facilitating ViT deployment in resource-constrained settings while maintaining accuracy. Advanced methods like post-training quantization and linear attention

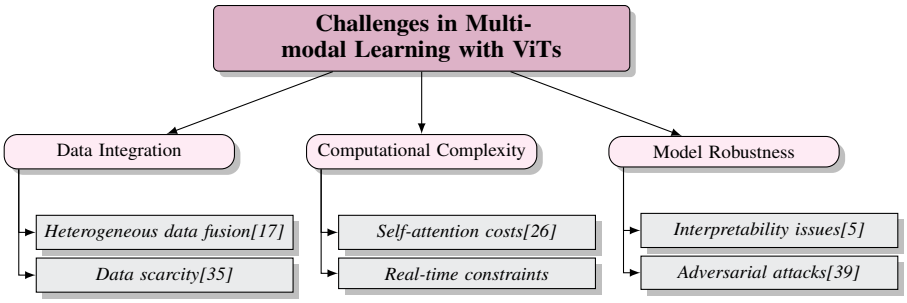
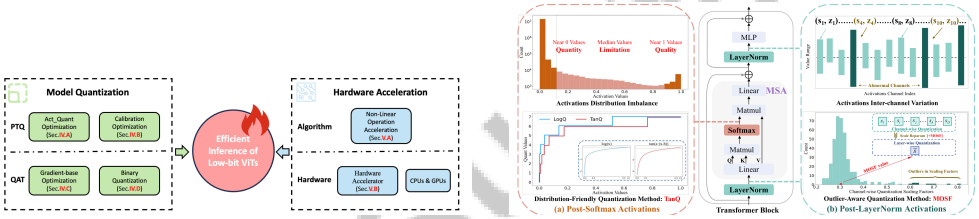


Figure 6: This figure illustrates the primary challenges in multimodal learning with Vision Transformers (ViTs), focusing on data integration, computational complexity, and model robustness. Each category highlights specific issues such as the fusion of heterogeneous data, the high computational costs of self-attention, and the need for robust and interpretable models.

integration enhance performance and efficiency [22, 14, 28]. The P2-ViT framework employs Power-of-Two scaling factors to optimize post-training quantization, while M2-ViT achieves significant energy-delay product savings [29, 10]. DiffRate uses gradient back-propagation to optimize token compression rates, enhancing efficiency by focusing on informative tokens [51]. CAS-ViT balances accuracy with low computational overhead, suitable for resource-limited environments [9]. Content-aware Token Sharing (CTS) improves efficiency by leveraging redundancy in image patches [3]. These quantization strategies significantly reduce computational and memory requirements, broadening ViTs' applicability in edge computing and real-time applications [69, 70, 23, 28].



(a) Efficient Inference of Low-bit ViTs[69]

(b) Improving Transformer Block Quantization for Efficient and Accurate Neural Network Training[70]

Figure 7: Examples of Quantization Techniques

As illustrated in Figure 7, quantization techniques are essential for enhancing neural networks' efficiency and performance, particularly in large-scale models like ViTs. The first example focuses on optimizing low-bit Vision Transformers, while the second explores activation quantization within transformer blocks, crucial for NLP tasks. These examples highlight quantization's critical role in optimizing modern neural network architectures for inference and training.

5.2 Knowledge Distillation

Knowledge distillation (KD) is a key strategy for optimizing Vision Transformers (ViTs), enabling knowledge transfer from a larger model (teacher) to a smaller model (student), thus addressing computational demands and resource constraints without significant performance loss [61]. KD enhances model optimization in frameworks like Unified Visual Transformer Compression (UVC), improving DeiT models' performance in multi-label classification by transferring critical features from teacher to student models [71]. Techniques like decoupled distillation, combining Masked Image Modeling (MIM) with KD, optimize learning in higher layers [72]. Masked Knowledge Distillation (MaskedKD) emphasizes selective knowledge transfer, reducing computational costs while maintaining performance [73]. The CAIT framework demonstrates KD's efficacy in balancing computational load reduction with spatial structure preservation [12]. KD is crucial for deploying ViTs in resource-limited scenarios, preserving essential features and facilitating effective training with reduced data requirements, valuable for tasks like image classification and medical imaging artifact detection [74, 43, 12, 75, 22].

5.3 Dynamic and Adaptive Methods

Dynamic and adaptive methods are increasingly essential for optimizing Vision Transformers (ViTs), enhancing computational efficiency and adaptability to task requirements and data characteristics. These methods enable ViTs to dynamically adjust operations, optimizing resource allocation and performance. The Token Propagation Controller (TPC) exemplifies this by variably adjusting token usage across layers [76]. Future research aims to integrate CNNs and ViTs, reducing computational burdens while maintaining high performance [4]. Hybrid models and ensemble methods promise improved scalability and robustness across applications [77]. Adjusting training example difficulty optimizes performance by adapting to training data complexity [35]. Empirical evidence shows Multi-head Self-Attentions (MSAs) enhance ViTs' adaptive capacity by flattening loss landscapes and reducing Hessian eigenvalues [63]. Integrating dynamic and adaptive methods with traditional training recipes, like LightRecipE, improves robust accuracy while maintaining clean accuracy [78]. Adaptive Layer Selection Fine-Tuning (ALaST) and Dynamic Tuning (DyT) enhance efficiency by allocating resources during fine-tuning, reducing training time, FLOPs, and memory usage. DyT improves parameter and inference efficiency by minimizing redundant computations [79, 23, 80]. These strategies broaden ViTs' applicability in resource-constrained environments, ensuring models adapt to specific task demands.

6 Computational Efficiency in Neural Networks

6.1 Real-World Applications and Trade-offs

Deploying Vision Transformers (ViTs) in practical scenarios necessitates understanding the interplay between model complexity and performance, especially under resource constraints. EfficientFormerV2 exemplifies this balance by achieving a 3.5% higher top-1 accuracy than MobileNetV2, maintaining similar latency and parameter counts, thereby demonstrating that optimized ViTs can rival lightweight CNNs in efficiency while surpassing them in performance [81]. Similarly, SPViT enhances computational efficiency by reducing latency without sacrificing accuracy on mobile devices and FPGA platforms [56].

The SSAT method further optimizes ViTs' performance on small datasets through joint task optimization, improving feature representation and reducing the carbon footprint compared to traditional training methods [14]. This underscores the potential of ViTs to excel with limited data through strategic optimization.

AFIDAF consistently outperforms existing lightweight networks in classification and downstream tasks, showcasing the effectiveness of innovative approaches in balancing model complexity and efficiency [82]. Additionally, the CTS technique reduces processed tokens by up to 44% without affecting segmentation quality, highlighting the importance of token management in enhancing computational efficiency [3].

Memory-efficient methods can decrease parameter counts in ViTs by up to 60

Trade-offs between model complexity and performance in ViTs are effectively managed through advancements in architecture, token management, and efficiency strategies. Techniques like quantization, low-rank approximation, knowledge distillation, and pruning are crucial for optimizing ViTs in resource-constrained environments, achieving a balance between computational efficiency and accuracy. The integration of robustness adapters and gated fusion modules within the TORA-ViTs framework enhances resilience to input perturbations while maintaining competitive accuracy, broadening applicability across domains such as edge computing and healthcare [23, 83]. Evaluating metrics like throughput, energy efficiency, and frame rate (FPS) further underscores the critical role of computational efficiency in real-world applications.

6.2 Future Directions in Efficient Neural Network Design

The future of efficient neural network design is poised for significant advancements, driven by ongoing research focused on optimizing computational efficiency and performance. Scaled ReLU mechanisms present a promising direction, showing potential in enhancing training efficiency and warranting further refinement for various architectures and tasks [84].

In Vision Transformers (ViTs), refining pruning strategies remains vital. Future research may explore frequency domain characteristics to inform more effective pruning implementations, reducing computational load while preserving performance [85]. Such advancements are expected to contribute to developing more efficient ViTs capable of operating in resource-limited environments without sacrificing accuracy.

Integrating dynamic and adaptive methods into neural network design is anticipated to play a crucial role in future developments. Advanced techniques that allow models to adjust operations based on specific task requirements and data characteristics are expected to enhance adaptability and scalability, broadening applications across fields like natural language processing, computer vision, and healthcare [37, 23].

The evolution of hybrid models combining convolutional operations with transformer architectures is also expected to drive progress in efficient neural network design. These models aim to leverage the synergistic strengths of both frameworks—through early and late fusion techniques—optimizing resource utilization and significantly enhancing performance in complex tasks, as evidenced by improved accuracy and AUC scores in multimodal learning scenarios [37, 20, 1, 2].

The trajectory of efficient neural network design centers on innovative strategies that optimize computational efficiency, enhance adaptability, and improve performance across diverse applications. By refining existing methodologies and exploring new approaches, the next generation of neural networks, particularly transformer-based models, is poised to achieve unprecedented levels of efficiency and effectiveness. These advancements will be propelled by sophisticated model compression techniques—such as pruning, quantization, and knowledge distillation—essential for deploying large language and vision models on practical devices while ensuring scalability and performance for state-of-the-art AI applications [37, 86].

7 Visual Tokenization and Its Impact

7.1 Understanding Visual Tokenization

Visual tokenization is pivotal in Vision Transformers (ViTs), converting visual inputs into token sequences for enhanced processing. This transformation enables ViTs to utilize self-attention mechanisms, capturing complex patterns and relationships within data [4]. By dividing images into patches treated as tokens, ViTs can model long-range dependencies and capture global context, which traditional convolutional methods often miss [63]. Beyond facilitating self-attention, visual tokenization boosts computational efficiency by enabling selective token compression and processing, thereby reducing computational demands without sacrificing performance [3]. Techniques such as token pruning and merging streamline processing, ensuring retention of only the most pertinent information [8]. This adaptability is crucial for deploying ViTs across various tasks, including image classification, object detection, and multimodal learning, underscoring visual tokenization's foundational role in achieving state-of-the-art performance [5].

Advancements in managing token redundancy, such as token labeling and pooling, significantly enhance ViT efficiency. These innovations accelerate inference speeds and maintain or improve predictive accuracy, highlighting the importance of optimizing token management for practical applications [87, 88, 89, 90, 48]. Visual tokenization's ability to transform visual data into a transformer-friendly format while optimizing computational resources underscores its significance in advancing machine learning and AI.

7.2 Impact on Model Performance

Visual tokenization significantly influences ViT performance and efficiency by shaping visual data representation and processing. By segmenting images into smaller tokens, transformers effectively utilize self-attention mechanisms, emphasizing relevant features while discarding or merging less informative tokens, enhancing computational efficiency and accuracy across various vision tasks [8, 46, 86, 54, 90]. The efficiency gains from visual tokenization arise from its capacity to reduce the computational burden of processing high-resolution images, allowing ViTs to focus resources on the most informative parts and optimize computational use [3]. This selective attention enhances the model's ability to capture global context and reduces overall computational complexity, making ViTs suitable for resource-constrained environments [8].

Visual tokenization supports advanced token compression techniques, such as pruning and merging, maintaining model performance while minimizing computational demands. These methods enable ViTs to dynamically adjust the number of tokens processed based on task complexity, ensuring retention of only the most relevant information [7]. This adaptability is crucial for improving scalability and efficiency across applications, from real-time video processing to large-scale image analysis.

Visual tokenization is vital for optimizing visual data representation and processing, enhancing ViT computational efficiency and accuracy. Recent research highlights the contributions of ViT components, including attention heads and multi-layer perceptrons (MLPs), to image representation and feature extraction related to shape, color, and texture. Innovative training methodologies, such as interpretability-aware training and automated progressive learning, optimize ViT performance and interpretability, significantly reducing training time without sacrificing accuracy. These advancements facilitate applications like image retrieval and token importance visualization, underscoring their transformative impact on the field [32, 22, 42].

7.3 Recent Innovations in Visual Tokenization

Recent advancements in visual tokenization, including token reorganization and pruning, significantly enhance ViT efficiency and performance. These innovations optimize visual data representation by selectively preserving attentive image tokens while eliminating or merging inattentive ones. For instance, the EViT method boosts inference speed by up to 50

The ShiftViT network replaces the traditional attention mechanism with a shift operation, achieving performance comparable to the Swin Transformer while simplifying the process, highlighting the potential of alternative operations in visual tokenization [30]. By streamlining tokenization, ShiftViT reduces computational overhead while retaining essential visual feature capture.

The SOFT mechanism employs dynamic token processing strategies to eliminate the need for softmax normalization by utilizing a Gaussian kernel, facilitating low-rank matrix decomposition, achieving linear complexity, and reducing computational demands [26]. Such innovative mathematical operations underscore the potential for enhancing visual tokenization efficiency.

Content-aware Token Sharing (CTS) exemplifies token processing optimization by leveraging semantic similarities, allowing semantically similar neighboring patches to share a token, significantly reducing the number of tokens processed in ViT-based semantic segmentation networks [3]. This strategy optimizes computational efficiency while maintaining high segmentation quality.

The Less-Attention Vision Transformer (LaViT) optimizes token processing through strategic manipulation of attention scores, capturing long-range dependencies with fewer computations, reducing costs while maintaining performance [11]. This method illustrates the potential of re-parameterizing attention mechanisms to enhance visual tokenization efficiency.

Collectively, these advancements in visual tokenization techniques reflect ongoing efforts to refine token management in ViTs. By implementing innovative strategies that optimize token processing through methods such as pruning, merging, and stylization, these approaches significantly enhance ViT efficiency across diverse applications. This optimization is particularly crucial for resource-constrained environments, allowing ViTs to maintain high performance while reducing computational complexity. For instance, the Token Pruning Pooling Transformers (PPT) framework minimizes redundant tokens, achieving over 45

8 Challenges and Future Directions

8.1 Challenges in Computational Efficiency

Vision Transformers (ViTs) face significant challenges in achieving computational efficiency due to their complex architectures and high computational demands. The quadratic complexity of the self-attention mechanism results in substantial GPU memory usage, impeding the widespread adoption of ViTs [4]. Additionally, attention saturation in deeper layers limits the model's ability to capture diverse semantic information [11]. While token compression strategies offer some relief, they often come with performance trade-offs, such as the loss of crucial subject and contextual information due to aggressive token pruning [8]. The effectiveness of methods like Content-aware Token Sharing

(CTS) heavily depends on the accuracy of the policy network in predicting token-sharing patches, which can impact performance if inaccurate [3].

Integrating multimodal learning within ViTs poses further challenges, especially in efficiently merging vision and language representations, which can increase computational costs [20]. Training on RGB images with a single objective has not yet achieved the multitask capabilities seen in NLP [5]. Quantization methods, crucial for computational efficiency, face obstacles such as re-quantization overheads from floating-point scaling factors, limiting integer-only inference and potential hardware speedups [29]. Furthermore, linear attention methods often fall short compared to softmax-based attention in capturing essential global and local features [65].

Scalability and complexity continue to be pressing issues, particularly when expanding features or model size. Accurate workload prediction in training-free acceleration frameworks is challenging in unpredictable environments, complicating efforts to improve computational efficiency [6]. Moreover, the specific architectural features of efficient ViTs, like those in M2-ViT, may not generalize well to other neural network architectures, limiting broader applicability [10]. Current benchmarks often overlook the impact of optimization techniques on model performance, resulting in incomplete assessments of ViT capabilities [25]. Addressing these challenges requires ongoing research into more efficient architectures, robust training techniques, and scalable solutions to enhance ViTs' computational efficiency across various applications. Optimizing token management, improving multimodal data alignment, and developing comprehensive benchmarks are crucial steps toward overcoming these limitations in deploying ViTs effectively in real-world scenarios.

8.2 Future Directions and Opportunities

Future research in Vision Transformers (ViTs) presents numerous opportunities to enhance efficiency, robustness, and applicability across diverse domains. Exploring mixed quantization strategies, as demonstrated in M2-ViT, could refine accelerator design and improve performance in resource-constrained environments [10]. Investigating the implications of scaling laws [34] could provide insights into optimizing ViT performance by applying these principles to new architectures and domains. Further analysis of components influencing ViT performance, including potential enhancements or alternatives to the attention mechanism, could refine model architectures [30].

Dynamic token sharing based on image complexity offers a novel opportunity to enhance computational efficiency. Applying the CTS framework to per-pixel prediction tasks like optical flow and depth estimation could expand ViTs' applicability in managing complex visual data [3]. Optimizations in the attention mechanism, as explored in LaViT, could enhance ViTs' effectiveness in intricate visual tasks [11]. Integrating Multi-head Self-Attentions (MSAs) with convolutional operations remains a promising area for exploration, focusing on optimizing this integration for dense prediction tasks and examining the impact of strong data augmentation on model uncertainty [63]. Enhancing the adaptability of token pruning and squeezing strategies, such as TPS, for hybrid ViTs could yield significant performance gains [8].

Moreover, incorporating additional modalities and improving tokenizer quality, as suggested by [5], could enhance generation quality and diversity. Training on larger, more curated datasets would further strengthen ViTs' capabilities in multimodal learning. Pursuing these future directions can deepen the understanding of ViTs across various learning paradigms, particularly regarding the effects of different supervision methods on performance. This includes investigating unique behaviors like Offset Local Attention Heads and the flexibility of ViTs in processing local and global information. Such insights will stimulate innovation in ViT architectures and optimize their deployment across diverse applications, ensuring their effectiveness in real-world scenarios [24, 65].

9 Conclusion

The survey highlights the transformative impact of token compression, multimodal learning, and model optimization on the computational efficiency of Vision Transformers (ViTs). By exploiting data sparsity, token compression techniques, such as the Tri-Level E-ViT framework, significantly enhance training speed while maintaining or improving accuracy, thus making ViTs more viable in environments with limited resources. Multimodal learning further strengthens data representation by integrating diverse data types, as demonstrated by the MAL-ViT model, which excels in multi-

attribute tasks by outperforming traditional CNNs and single-attribute ViTs. This adaptability is crucial for applying ViTs in complex data integration scenarios, including visiolinguistic tasks and intricate scene understanding. Moreover, model optimization methods like the D2-MAE approach offer remarkable advancements by reducing inference times and memory usage without compromising performance, thereby lowering computational costs and training durations. The IA-ViT framework exemplifies how enhanced interpretability and predictive capabilities can be achieved, reinforcing the effectiveness of these optimization strategies. The survey also underscores the importance of improving robustness and explainability in ViTs to address key challenges in AI. Techniques such as the ToE method demonstrate the ability to accelerate ViT training effectively, achieving superior results compared to traditional methods. Architectural innovations like ViTMatte illustrate the potential for optimized designs to deliver state-of-the-art outcomes in specific tasks, such as image matting, while maintaining efficiency.

References

[1] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Does a technique for building multimodal representation matter? – comparative analysis, 2022.

[2] Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3840–3857, 2021.

[3] Chenyang Lu, Daan de Geus, and Gijs Dubbelman. Content-aware token sharing for efficient semantic segmentation with vision transformers, 2023.

[4] Khawar Islam. Recent advances in vision transformer: A survey and outlook of recent work, 2023.

[5] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling, 2023.

[6] Jung Hwan Heo, Seyedarmin Azizi, Arash Fayyazi, and Massoud Pedram. Training-free acceleration of vits with delayed spatial merging, 2024.

[7] Dingyuan Zhang, Dingkang Liang, Zichang Tan, Xiaoqing Ye, Cheng Zhang, Jingdong Wang, and Xiang Bai. Make your vit-based multi-view 3d detectors faster via token compression, 2024.

[8] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023.

[9] Tianfang Zhang, Lei Li, Yang Zhou, Wentao Liu, Chen Qian, and Xiangyang Ji. Cas-vit: Convolutional additive self-attention vision transformers for efficient mobile applications, 2024.

[10] Yanbiao Liang, Huihong Shi, and Zhongfeng Wang. M^2-vit: Accelerating hybrid vision transformers with two-level mixed quantization, 2024.

[11] Shuoxi Zhang, Hanpeng Liu, Stephen Lin, and Kun He. You only need less attention at each stage in vision transformers, 2024.

[12] Ao Wang, Hui Chen, Zijia Lin, Sicheng Zhao, Jungong Han, and Guiguang Ding. Cait: Triple-win compression towards high accuracy, fast inference, and favorable transferability for vits, 2023.

[13] Ibrahim Batuhan Akkaya, Senthilkumar S. Kathiresan, Elahe Arani, and Bahram Zonooz. Enhancing performance of vision transformers on small datasets through local inductive bias incorporation, 2023.

[14] Srijan Das, Tanmay Jain, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das, and Michael S. Ryoo. Limited data, unlimited potential: A study on vits augmented by masked autoencoders, 2023.

[15] Abdul Mueed Hafiz, Shabir Ahmad Parah, and Rouf Ul Alam Bhat. Attention mechanisms and deep learning for machine vision: A survey of the state of the art, 2021.

[16] Maria Lymperaiou and Giorgos Stamou. A survey on knowledge-enhanced multimodal learning. *Artificial Intelligence Review*, 57(10):284, 2024.

[17] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.

[18] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications, 2024.

[19] Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification, 2023.

[20] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models, 2022.

[21] Junichiro Niimi. An efficient multimodal learning framework to comprehend consumer preferences using bert and cross-attention, 2024.

[22] Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. Decomposing and interpreting image representations via text in vits beyond clip, 2024.

[23] Feiyang Chen, Ziqian Luo, Lisang Zhou, Xueting Pan, and Ying Jiang. Comprehensive survey of model compression and speed up for vision transformers. *arXiv preprint arXiv:2404.10407*, 2024.

[24] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers, 2023.

[25] Sanad Aburass and Osama Dorgham. Performance evaluation of swin vision transformer model using gradient accumulation optimization technique, 2023.

[26] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity, 2022.

[27] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[28] Huihong Shi, Haikuo Shao, Wendong Mao, and Zhongfeng Wang. Trio-vit: Post-training quantization and acceleration for softmax-free efficient vision transformer, 2024.

[29] Huihong Shi, Xin Cheng, Wendong Mao, and Zhongfeng Wang. P^2-vit: Power-of-two post-training quantization and acceleration for fully quantized vision transformer, 2024.

[30] Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism, 2022.

[31] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy, 2022.

[32] Changlin Li, Bohan Zhuang, Guangrun Wang, Xiaodan Liang, Xiaojun Chang, and Yi Yang. Automated progressive learning for efficient training of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12496, 2022.

[33] Changlin Li, Bohan Zhuang, Guangrun Wang, Xiaodan Liang, Xiaojun Chang, and Yi Yang. Automated progressive learning for efficient training of vision transformers, 2022.

[34] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[35] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. Effective vision transformer training: A data-centric perspective, 2022.

[36] Sonia Bbouzidi, Ghazala Hcini, Imen Jdey, and Fadoua Drira. Convolutional neural networks and vision transformers for fashion mnist classification: A literature review, 2024.

[37] Raby Hamadi. Large language models meet computer vision: A brief survey, 2023.

[38] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

[39] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers, 2024.

[40] Zheng Wang and Wenjie Ruan. Understanding adversarial robustness of vision transformers via cauchy problem, 2022.

[41] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. Multi-scale high-resolution vision transformer for semantic segmentation, 2021.

[42] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer, 2023.

[43] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios, 2022.

[44] Seyedarmin Azizi, Mahdi Nazemi, and Massoud Pedram. Memory-efficient vision transformers: An activation-aware mixed-rank compression strategy, 2024.

[45] Sonain Jamil, Md. Jalil Piran, and Oh-Jin Kwon. A comprehensive survey of transformers for computer vision, 2022.

[46] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token cropr: Faster vits for quite a few tasks, 2024.

[47] Hanning Chen, Yang Ni, Wenjun Huang, Yezi Liu, SungHeon Jeong, Fei Wen, Nathaniel Bastian, Hugo Latapie, and Mohsen Imani. Vltp: Vision-language guided token pruning for task-oriented segmentation, 2024.

[48] Xinjian Wu, Fanhu Zeng, Xiudong Wang, and Xinghao Chen. Ppt: Token pruning and pooling for efficient vision transformers, 2024.

[49] Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Gustavo A. Vargas Hakim, David Osowiechi, Ismail Ben Ayed, and Christian Desrosiers. Tfs-vit: Token-level feature stylization for domain generalization, 2024.

[50] Xuwei Xu, Changlin Li, Yudong Chen, Xiaojun Chang, Jiajun Liu, and Sen Wang. No token left behind: Efficient vision transformer via dynamic token idling, 2023.

[51] Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. Diffrate: Differentiable compression rate for efficient vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17164–17174, 2023.

[52] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer, 2021.

[53] Xuwei Xu, Sen Wang, Yudong Chen, Yanping Zheng, Zhewei Wei, and Jiajun Liu. Gtp-vit: Efficient vision transformers via graph-based token propagation, 2024.

[54] Wenxuan Huang, Yunhang Shen, Jiao Xie, Baochang Zhang, Gaoqi He, Ke Li, Xing Sun, and Shaohui Lin. A general and efficient training for transformer via token expansion, 2024.

[55] Mingjia Shi, Yuhao Zhou, Ruiji Yu, Zekai Li, Zhiyuan Liang, Xuanlei Zhao, Xiaojiang Peng, Shanmukha Ramakrishna Vedantam, Wangbo Zhao, Kai Wang, and Yang You. Faster vision mamba is rebuilt in minutes via merged token re-training, 2025.

[56] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pages 620–640. Springer, 2022.

[57] Clément Grisi, Geert Litjens, and Jeroen van der Laak. Hierarchical vision transformers for context-aware prostate cancer grading in whole slide images, 2023.

[58] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers, 2023.

[59] Ao Wang, Fengyuan Sun, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. [cls] token tells everything needed for training-free efficient mllms, 2024.

[60] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging, 2023.

[61] Shixing Yu, Tianlong Chen, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, and Zhangyang Wang. Unified visual transformer compression. *arXiv preprint arXiv:2203.08243*, 2022.

[62] Ao Wang, Hui Chen, Zijia Lin, Sicheng Zhao, Jungong Han, and Guiguang Ding. Cait: Triple-win compression towards high accuracy, fast inference, and favorable transferability for vits. *arXiv preprint arXiv:2309.15755*, 2023.

[63] Namuk Park and Songkuk Kim. How do vision transformers work?, 2022.

[64] Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, and Mike Zheng Shou. Making vision transformers efficient from a token sparsification view, 2023.

[65] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7496, 2023.

[66] Diganta Misra, Jay Gala, and Antonio Orvieto. On the low-shot transferability of [v]-mamba, 2024.

[67] Leonardo Scabini, Andre Sacilotti, Kallil M. Zielinski, Lucas C. Ribas, Bernard De Baets, and Odemir M. Bruno. A comparative survey of vision transformers for feature extraction in texture analysis, 2024.

[68] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey, 2023.

[69] Dayou Du, Gu Gong, and Xiaowen Chu. Model quantization and hardware acceleration for vision transformers: A comprehensive survey, 2024.

[70] Lianwei Yang, Haisong Gong, and Qingyi Gu. Dopq-vit: Towards distribution-friendly and outlier-aware post-training quantization for vision transformers, 2024.

[71] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Data-efficient vision transformers for multi-label disease classification on chest radiographs, 2022.

[72] Jin Gao, Shubo Lin, Shaoru Wang, Yutong Kou, Zeming Li, Liang Li, Congxuan Zhang, Xiaoqin Zhang, Yizheng Wang, and Weiming Hu. An experimental study on exploring strong lightweight vision transformers via masked image modeling pre-training, 2024.

[73] Seungwoo Son, Jegwang Ryu, Namhoon Lee, and Jaeho Lee. The role of masking for efficient supervised knowledge distillation of vision transformers, 2024.

[74] Gousia Habib, Damandeep Singh, Ishfaq Ahmad Malik, and Brejesh Lall. Optimizing vision transformers with data-free knowledge transfer, 2024.

[75] Neel Kanwal, Trygve Eftestol, Farbod Khoraminia, Tahlita CM Zuiverloon, and Kjersti Engan. Vision transformers for small histological datasets learned through knowledge distillation, 2023.

[76] Wentao Zhu. Tpc-vit: Token propagation controller for efficient vision transformer, 2024.

[77] Asifullah Khan, Zunaira Rauf, Abdul Rehman Khan, Saima Rathore, Saddam Hussain Khan, Najmus Saher Shah, Umair Farooq, Hifsa Asif, Aqsa Asif, Umme Zahoora, Rafi Ullah Khalil, Suleman Qamar, Umme Hani Asif, Faiza Babar Khan, Abdul Majid, and Jeonghwan Gwak. A recent survey of vision transformers for medical image segmentation, 2023.

[78] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers, 2023.

[79] Alessio Devoto, Federico Alvetreti, Jary Pomponi, Paolo Di Lorenzo, Pasquale Minervini, and Simone Scardapane. Adaptive layer selection for efficient vision transformer fine-tuning, 2024.

[80] Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. Dynamic tuning towards parameter and inference efficiency for vit adaptation, 2024.

[81] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16889–16900, 2023.

[82] Yunling Zheng, Zeyi Xu, Fanghui Xue, Biao Yang, Jiancheng Lyu, Shuai Zhang, Yingyong Qi, and Jack Xin. Afidaf: Alternating fourier and image domain adaptive filters as an efficient alternative to attention in vits, 2024.

[83] Yanxi Li and Chang Xu. Trade-off between robustness and accuracy of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7568, 2023.

[84] Pichao Wang, Xue Wang, Hao Luo, Jingkai Zhou, Zhipeng Zhou, Fan Wang, Hao Li, and Rong Jin. Scaled relu matters for training vision transformers, 2022.

[85] Zhenyu Wang, Hao Luo, Pichao Wang, Feng Ding, Fan Wang, and Hao Li. Vtc-lfc: Vision transformer compression with low-frequency components. *Advances in Neural Information Processing Systems*, 35:13974–13988, 2022.

[86] Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. A survey on transformer compression. *arXiv preprint arXiv:2402.05964*, 2024.

[87] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers, 2023.

[88] Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers, 2021.

[89] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in neural information processing systems*, 34:18590–18602, 2021.

[90] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.