**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Comparing Vision Transformers and Convolutional Neural Networks: A Systematic Analysis

**Chandrasekar Adhithya Harsha Pasumarthi**
*Independent Researcher, USA*
**Corresponding Author**: Chandrasekar Adhithya Harsha Pasumarthi, **E-mail**: ca.pasumarthi@gmail.com

**| ABSTRACT**

Vision Transformers have emerged as powerful alternatives to Convolutional Neural Networks for image classification tasks. Systematic comparisons under controlled settings remain limited despite growing adoption of transformer-based vision models. The present article conducts comprehensive evaluation of ViTs and CNNs across identical datasets, training conditions, and computational budgets. Multiple architectures including ResNet, EfficientNet, ViT-Base, and DeiT undergo training on benchmark datasets such as CIFAR-10, CIFAR-100, and customized real-world datasets. Performance evaluation encompasses accuracy, F1-score, training stability, adversarial robustness, and inference latency metrics. Results demonstrate that ViTs outperform CNNs on larger datasets while exhibiting superior robustness to noise and perturbations. CNNs maintain advantages for small datasets due to strong inductive biases embedded within convolutional architectures. The effective receptive field in deep convolutional networks exhibits Gaussian distribution patterns centered on each spatial location. Vision transformers learn spatial relationships entirely from data through global self-attention mechanisms. Dataset scale fundamentally determines relative performance characteristics between architectural families. Transformer architectures require substantial training data to discover optimal attention patterns. Convolutional networks converge efficiently on smaller datasets through built-in spatial priors. The article identifies specific conditions under which each architecture demonstrates clear advantages. Findings contribute to understanding of transformer-based vision models while offering practical guidance for architecture selection in applied machine learning systems.

**| KEYWORDS**

Vision Transformers, Convolutional Neural Networks, Image Classification, Deep Learning Architectures, Adversarial Robustness, Transfer Learning

**| ARTICLE INFORMATION**

**ACCEPTED:** 01 January 2026 **PUBLISHED:** 28 January 2026 **DOI:** 10.32996/jcsts.2026.8.2.3

*Introduction*

Convolutional neural networks have ruled pc vision for over a decade. The architecture leverages spatial locality and translation equivariance as fundamental inductive biases. Deep residual learning architectures revolutionized the field by introducing skip connections that enable training of extremely deep networks. These residual connections address the degradation problem where adding more layers paradoxically decreases accuracy in plain networks [1]. The skip connections allow gradients to flow directly through the network during backpropagation. This design enables the construction of networks with hundreds of layers while maintaining stable training dynamics.

CNNs extract hierarchical features through localized receptive fields. Early layers detect edges and textures. Deeper layers capture complex semantic concepts and object parts. The translation equivariance property allows pattern recognition regardless of spatial position. Parameter sharing across image locations reduces model complexity while maintaining representational capacity. Convolutional operations create strong inductive biases that facilitate learning from limited data. Residual networks demonstrated that depth remains crucial for achieving superior performance on challenging recognition tasks [1]. The

architecture enables learning of residual functions with reference to layer inputs rather than learning unreferenced functions. This formulation proves easier to optimize in practice.

Vision Transformers introduced a paradigm shift by eliminating explicit convolutional operations entirely. The architecture treats images as sequences of patches processed through self-attention mechanisms. Each image divides into fixed-size patches that undergo linear embedding. Standard transformer encoder blocks then process the resulting sequence. ViTs lack inherent spatial inductive biases unlike their convolutional counterparts. The model learns spatial relationships entirely from data through global self-attention mechanisms. Training vision transformers requires careful consideration of architectural choices and training strategies [2]. The self-attention mechanism computes pairwise interactions between all image patches simultaneously.

Initial ViT implementations demonstrated competitive performance on large-scale datasets. However, questions persist regarding behavior under resource constraints and data scarcity. Preliminary investigations suggested substantial data requirements compared to CNNs. Significant performance degradation occurs when training exclusively on medium-sized datasets. The quadratic computational complexity of self-attention raises concerns about inference efficiency. Deployment feasibility in resource-constrained environments remains uncertain. Adversarial robustness characteristics require deeper investigation. Conflicting reports exist regarding whether global attention provides inherent resistance to localized perturbations. Training strategies significantly impact the final model performance and generalization capabilities [2].

Existing comparative studies often evaluate architectures under different training regimes. Many investigations compare pre-trained models fine-tuned on downstream tasks. This approach makes isolating architectural effects impossible. Performance differences may stem from pre-training dataset characteristics rather than architectural properties. Other studies employ inconsistent training protocols across architectures. Variations in optimization schedules introduce confounding variables. Data augmentation strategies and regularization techniques often differ between experiments. This lack of controlled experimental design prevents definitive conclusions. Observed differences may result from implementation details rather than genuine architectural advantages. Vision transformer training benefits from specific augmentation techniques and optimization configurations tailored to the architecture [2].

This research addresses these limitations through controlled experiments that isolate architectural effects. Representative CNN and ViT architectures undergo systematic comparison under identical conditions. Training protocols, optimization strategies, and evaluation metrics remain consistent across all architectures. The investigation examines behavior across datasets of varying scales. Different perturbation types and computational constraints receive thorough analysis. This approach identifies specific conditions where each architecture family demonstrates clear advantages.

*Related Work / Methodology*

Previous comparative evaluations of Vision Transformers and Convolutional Neural Networks often employed inconsistent experimental protocols. Different training regimes, optimization schedules, and data augmentation strategies confounded architectural differences with implementation variations. Many investigations compared pre-trained models fine-tuned on downstream tasks, making isolation of architectural effects impossible. Performance differences stemmed from pre-training dataset characteristics rather than inherent architectural properties. The present article addresses these limitations through controlled experimental design maintaining identical conditions across all architectures.

The methodology employs representative architectures from both families including ResNet-based CNNs and standard ViT implementations. Training occurs on benchmark datasets spanning different scales from thousands to hundreds of thousands of images. Identical optimization strategies apply across all models including learning rate schedules, warmup periods, and regularization techniques. Data augmentation remains consistent preventing confounding variables from affecting results. Evaluation metrics encompass classification accuracy, training stability, adversarial robustness under perturbations, and computational efficiency during inference.

The framework systematically varies dataset scale while holding other factors constant. Small-scale experiments reveal advantages of convolutional inductive biases. Large-scale experiments demonstrate transformer superiority given sufficient training data. Adversarial robustness testing employs gradient-based attacks evaluating architectural vulnerabilities. Transfer learning experiments assess generalization across distribution shifts. Computational profiling quantifies inference latency and memory requirements across hardware platforms. The controlled methodology isolates genuine architectural differences from experimental artifacts providing definitive performance boundaries.

*Architectural Foundations and Training Dynamics*

*Fundamental Design Principles*

CNNs exploit spatial structure through localized receptive fields that expand progressively through network depth. The theoretical receptive field differs significantly from the effective receptive field in practice. Research demonstrates that not all pixels within the theoretical receptive field contribute equally to network outputs [3]. Central pixels exert disproportionate influence compared to peripheral regions. The effective receptive field exhibits a Gaussian distribution pattern centered on each

location. This concentration occurs because gradient magnitudes decrease with distance from the center during backpropagation.

Modern architectures construct hierarchical feature representations through stacked convolutional layers. Initial layers employ small kernels that capture local patterns including edges and texture elements. Deeper layers aggregate information from progressively larger spatial regions. The effective receptive field grows with network depth but maintains its concentrated Gaussian characteristic [3]. Translation equivariance emerges from applying identical learned filters across all spatial locations. Networks recognize patterns regardless of position within the image. Pooling operations provide scale invariance while reducing computational demands through spatial downsampling.

Vision Transformers adopt fundamentally different architectural principles. The standard approach divides input images into fixed-size patches without overlap. Each patch undergoes flattening and linear projection to create embeddings. Multi-head self-attention mechanisms process these sequences to capture global relationships. The architecture lacks inherent spatial bias present in convolutional designs. Spatial relationships require learning entirely from training data through attention patterns. Self-attention computes pairwise interactions between all patches simultaneously. This design enables global receptive fields from the first layer. Positional encodings inject spatial information into the otherwise permutation-invariant architecture.

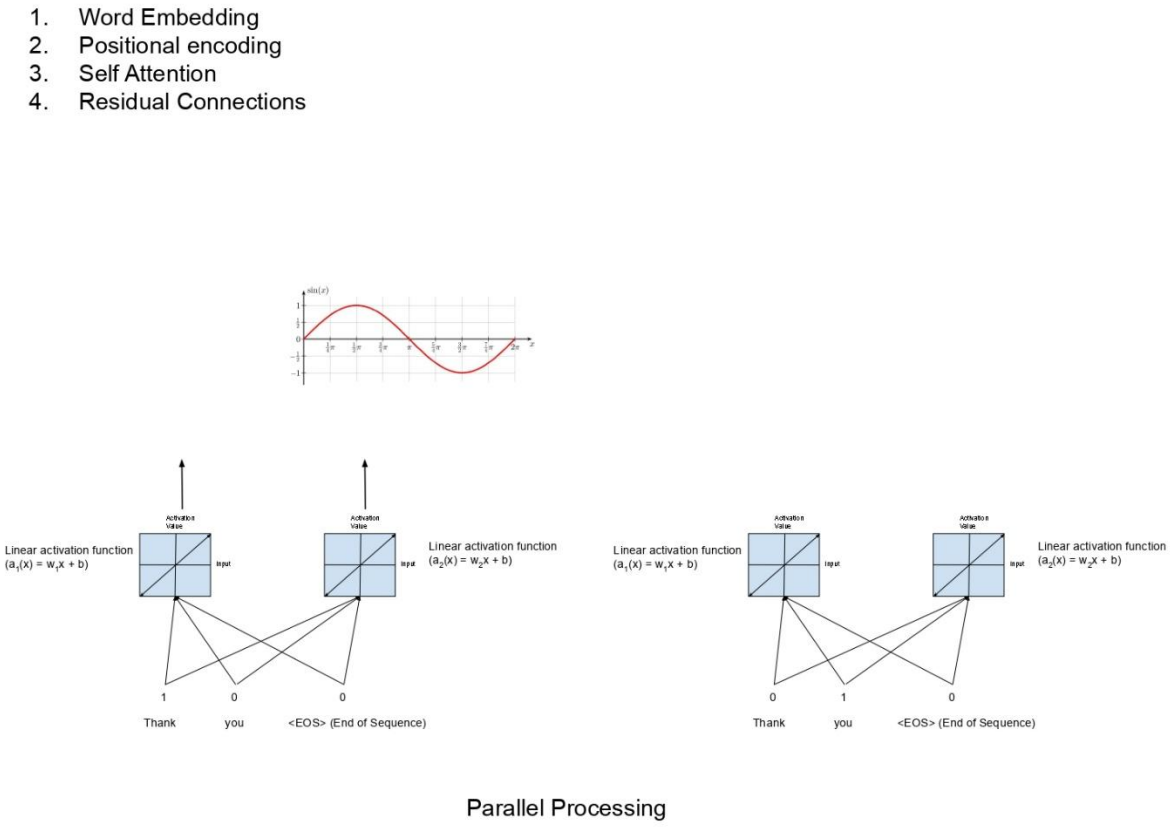


Fig 1. Core Components of Vision Transformer Architecture [3]

[Note: The architecture comprises four fundamental components: word (patch) embeddings that convert image patches into vector representations, positional encodings using sinusoidal functions to inject spatial information, self-attention mechanisms enabling parallel processing of all sequence elements, and residual connections facilitating gradient flow through deep networks.]

*Training Behavior and Optimization Characteristics*

Architectural differences manifest distinctly during training procedures. CNNs converge efficiently on smaller datasets due to strong inductive biases. The concentrated effective receptive field means that nearby pixels dominate gradient flow during backpropagation [3]. This property accelerates learning of local feature detectors. Training curves show smooth, monotonic improvement across epochs. Gradient stability remains consistent throughout optimization. Strong spatial priors enable rapid learning of feature hierarchies.

ViTs exhibit markedly different training dynamics. Without explicit spatial priors, extensive training becomes necessary for discovering spatial relationships. Early training phases often show unstable loss curves. The model requires substantial iteration

counts before stable convergence emerges. However, sufficient data availability transforms training characteristics dramatically. Large-scale datasets enable rapid convergence to superior performance levels.

Knowledge distillation techniques offer pathways to improve training efficiency. Recent advances demonstrate that attention mechanisms can bridge architectural differences during distillation [4]. Teacher models transfer learned representations to student architectures through attention-based knowledge transfer. This approach reduces data requirements compared to training from scratch. The distillation process preserves critical feature relationships while adapting to different architectural constraints.

The self-attention mechanism's global connectivity introduces unique optimization challenges. Gradient flow through attention layers requires careful management. Learning rate scheduling becomes critical for successful training. Warmup strategies gradually increase learning rates during initial phases. This technique establishes stable parameter configurations before aggressive optimization. Attention-based distillation frameworks demonstrate that feature alignment across architectures benefits from explicit attention supervision [4]. The approach guides student models toward learning similar attention patterns as teacher networks.

| Architecture | Spatial Processing | Receptive Field | Parameter Efficiency | Training Convergence | Optimization Requirements |
|---|---|---|---|---|---|
| CNNs | Local connectivity through convolutional kernels | Gradually expanding with Gaussian distribution | High through weight sharing | Efficient on small datasets with smooth curves | Standard learning rates with batch normalization |
| Vision Transformers | Global self-attention across all patches | Global from first layer | Lower due to attention mechanisms | Requires extensive training on limited data | Careful scheduling with warmup strategies |

Table 1. Architectural Characteristics And Training Properties Of Cnns And Vits [3, 4]

*Dataset Scale Dependencies and Performance Characteristics*

*Small Dataset Behavior*

On datasets containing thousands to tens of thousands of images, CNNs consistently demonstrate superior performance across diverse recognition tasks. Convolutional inductive biases provide essential structural priors that compensate for limited training data availability. Applications in specialized domains illustrate these advantages clearly. Agricultural pest identification represents a challenging scenario with limited annotated samples and complex visual backgrounds. Deep residual networks achieve effective classification despite constrained dataset sizes [5]. The networks successfully identify pest species even when backgrounds contain dense foliage, soil textures, and varying illumination conditions.

Local connectivity patterns enable effective feature learning when sample diversity remains limited. Each convolutional filter processes small spatial neighborhoods before aggregating information across layers. This design reduces the hypothesis space during training. Residual connections facilitate gradient flow through very deep architectures. Skip connections allow building networks with substantial depth while maintaining training stability. Parameter efficiency through weight sharing substantially reduces overfitting risk [5]. The same learned filters apply across all spatial positions in feature maps. Networks effectively observe many more examples of each feature detector than nominal dataset size suggests.

ViTs struggle under data-constrained conditions due to architectural flexibility requiring substantial training samples. The lack of inherent spatial priors means spatial hierarchies must emerge purely from observed patterns. Without sufficient samples, models cannot effectively leverage global attention capabilities. Training often results in memorization rather than learning generalizable representations. The attention mechanism can attend to any image patch regardless of spatial distance. This flexibility becomes a liability when limited data fails to provide adequate examples of meaningful relationships. Complex background scenarios exacerbate these challenges as models must learn to distinguish relevant features from distractors [5].

*Large-Scale Dataset Performance*

Dataset size increases beyond hundreds of thousands of images fundamentally alter relative performance characteristics. ViTs increasingly outperform CNNs as training data scales up substantially. However, data-efficient training strategies enable transformer architectures to achieve competitive performance with reduced sample requirements. Detection transformers benefit from architectural innovations and training techniques specifically designed to minimize data dependence [6]. These approaches reduce the performance gap between transformers and CNNs on medium-sized datasets.

Query-based detection frameworks introduce unique challenges for data-efficient learning. The architecture requires learning complex relationships between image features and object queries. Standard training procedures demand extensive datasets to establish these mappings effectively. Recent advances focus on improving initialization strategies and training stability. Better query initialization reduces the number of training iterations required for convergence. Enhanced training protocols accelerate learning of attention patterns between queries and visual features [6].

Global receptive fields enable holistic image understanding from initial transformer layers. Every patch can potentially interact with every other patch through attention computations. The architecture's flexibility allows learning task-specific feature hierarchies without predetermined constraints. Efficient training techniques reduce reliance on massive datasets while maintaining competitive performance. Hybrid approaches combining convolutional stems with transformer bodies balance inductive biases with attention-based reasoning [6].

CNNs maintain competitive performance on large datasets despite architectural limitations. Fixed receptive field growth patterns limit capturing distant spatial relationships. Convolutional kernels aggregate information from progressively larger regions through stacked layers. This gradual expansion may miss important contextual information from distant image regions. Computational efficiency remains advantageous for deployments in resource-constrained environments. Convolutional operations map efficiently to specialized hardware accelerators. Parameter counts remain manageable through weight sharing mechanisms.

| Dataset Scale | CNN Performance | ViT Performance | Key Advantage | Architectural Benefit | Transfer Learning |
|---|---|---|---|---|---|
| Small (thousands of images) | Superior accuracy and generalization | Struggles with overfitting and memorization | Convolutional inductive bias | Built-in spatial priors compensate for limited data | Effective with frozen early layers |
| Medium (tens of thousands) | Strong performance with efficient training | Moderate performance requiring longer training | Parameter efficiency through sharing | Local connectivity enables feature learning | Requires domain similarity |
| Large (hundreds of thousands) | Competitive but plateauing performance | Superior performance with rapid convergence | Global attention mechanisms | Learns task-specific hierarchies without constraints | Strong generalization across domains |

Table 2. Performance Characteristics Across Different Dataset Scales [5, 6]

*Robustness and Generalization Properties*

*Adversarial and Noise Robustness*

ViTs demonstrate enhanced robustness to various input perturbations including adversarial attacks and natural corruptions. Systematic evaluations reveal fundamental differences in vulnerability patterns between architectures. Comparative analyses show that transformer architectures exhibit distinct responses to adversarial perturbations compared to convolutional networks [7]. The self-attention mechanism's global perspective enables more holistic image understanding. This architectural property makes models less sensitive to localized perturbations.

The architecture processes information through multiple attention heads operating independently. Each head learns different feature relationships across image patches. This redundancy provides implicit regularization against input variations. Adversarial attacks targeting specific attention mechanisms may not affect all heads equivalently. Natural image corruptions including noise, blur, and weather effects affect transformers differently than CNNs. Comprehensive robustness evaluations demonstrate that architectural choices significantly influence resilience to perturbations [7].

CNNs show greater vulnerability to certain categories of adversarial perturbations. Small, carefully crafted perturbations exploit the local processing nature of convolutional operations. These perturbations cascade through the hierarchical structure amplifying errors at each layer. Gradient-based attacks prove particularly effective against convolutional architectures due to smooth differentiable structures. The locality bias that aids learning on clean data becomes exploitable under adversarial conditions. However, architectural modifications substantially improve CNN robustness. Adversarial training incorporates perturbed examples during optimization. Defense mechanisms including input preprocessing and ensemble methods enhance resilience [7].

*1) Distribution Shift and Transfer Learning*

Data distributions different from training sets reveal generalization capabilities across architectures. ViTs generally exhibit superior performance under distribution shift scenarios. Learned representations capture abstract, transferable features

applicable across diverse visual domains. However, standard transformer architectures contain inefficiencies that limit their practical deployment. Attention mechanisms compute relationships between all patch pairs regardless of relevance. This exhaustive computation introduces unnecessary computational overhead. Recent architectural innovations address these limitations through selective attention mechanisms [8].

Skip-attention approaches improve efficiency by reducing redundant computations in attention layers. The technique identifies which attention operations contribute meaningfully to predictions. Less informative attention computations can be bypassed without significant performance degradation. This selective processing reduces computational requirements while maintaining or improving accuracy. The approach demonstrates that not all attention operations prove equally valuable for final predictions [8].

CNNs demonstrate effective transfer learning within related visual domains. Pre-trained models serve as feature extractors for diverse recognition tasks. Early convolutional layers learn generic edge and texture detectors transferring broadly across problems. Deeper layers capture more task-specific semantic information. Switch learning works well whilst supply and goal domain names share comparable characteristics. Considerable domain shifts require tremendous fine-tuning or structure modifications. Hierarchical feature extraction shows sensitivity to domain characteristics. Fixed convolutional structures impose constraints on feature learning. Attention mechanisms provide greater flexibility for adapting to novel visual domains through dynamic feature reweighting [8].

| **Property** | **CNNs** | **Vision Transformers** | **Mechanism** | **Defense Strategy** | **Domain Adaptation** |
|---|---|---|---|---|---|
| TaAdversarial Perturbations | Higher vulnerability to gradient-based attacks | Enhanced robustness through global processing | Localized perturbations cascade through layers vs. distributed attention | Adversarial training and ensemble methods | Requires extensive fine-tuning |
| Natural Corruptions | Sensitive to local degradations | Resilient through attention reweighting | Fixed receptive fields vs. adaptive attention | Input preprocessing and augmentation | Skip-attention reduces redundant computations |
| Distribution Shift | Moderate generalization within similar domains | Superior generalization across diverse domains | Hierarchical features vs. abstract representations | Careful fine-tuning of deeper layers | Dynamic feature reweighting for novel inputs |

Table 3. **Robustness And Generalization Comparison [7, 8].**

*Computational Considerations and Practical Implications*

*Inference Efficiency and Resource Requirements*

CNNs maintain significant advantages in computational efficiency for real-time applications. The local connectivity pattern reduces memory footprint substantially compared to globally connected architectures. Parameter sharing across spatial locations minimizes total parameter counts while maintaining representational capacity. However, standard static convolutions process all input channels and spatial locations uniformly. This approach introduces computational redundancy when certain regions or channels contain less informative content. Dynamic convolution addresses this limitation by adapting kernel parameters based on input characteristics [9].

Dynamic convolution mechanisms aggregate multiple convolution kernels with input-dependent attention weights. The approach learns to emphasize relevant kernel components for each input sample. This adaptability improves model expressiveness without substantially increasing parameter counts. Multiple parallel convolution kernels capture diverse feature patterns. Linear combinations of these kernels generate input-specific filters. The dynamic aggregation enables efficient feature extraction by focusing computational resources on informative patterns [9].

ViTs require substantially greater computational resources during both training and inference. The self-attention mechanism computes relationships between all input positions. Standard attention formulations exhibit quadratic complexity with respect to sequence length. Processing high-resolution images generates long patch sequences creating scalability challenges. Reminiscence requirements grow unexpectedly as image decision increases. The attention computation dominates overall computational cost in transformer architectures [10].

Recent architectural innovations partially address computational limitations. Hierarchical designs process images at multiple scales reducing sequence lengths. Efficient attention mechanisms approximate full attention through various strategies. However, fundamental computational requirements remain higher than convolutional alternatives for equivalent model capacities.

*Deployment and Hardware Considerations*

Practical deployment scenarios strongly influence architecture selection decisions. CNNs benefit from extensive hardware optimization across diverse platforms. Mobile devices and embedded systems provide specialized support for convolutional operations. Dynamic convolution extends these benefits while adding adaptive capabilities. The technique maintains compatibility with existing optimization frameworks and hardware accelerators. Implementation requires minimal modifications to standard convolution primitives [9].

Model compression techniques reduce deployment costs for CNN architectures. Quantization decreases precision from floating-point to integer representations. Pruning removes redundant parameters without significant accuracy degradation. These optimizations enable deployment on resource-constrained devices.

ViTs require more powerful hardware for maintaining acceptable inference speeds. The attention mechanism's computational pattern differs from convolutions. Memory-intensive operations strain available bandwidth on resource-constrained devices. Attention mechanisms rely on matrix multiplication as the fundamental primitive. The operation computes weighted combinations of value vectors based on learned attention distributions [10]. Emerging hardware designs increasingly optimize for transformer operations. Specialized accelerators reduce computational overhead through custom datapaths.

The attention mechanism enables modeling long-range dependencies without architectural constraints. This flexibility comes at computational cost compared to local operations. Positional encodings inject sequential information into the permutation-invariant architecture. The approach allows transformers to process sequences of arbitrary length [10]. Hardware-software co-design processes optimize transformer execution throughout implementation stacks.

| Aspect | CNNs | Vision Transformers | Complexity | Hardware Support | Optimization Techniques |
|---|---|---|---|---|---|
| Inference Speed | Fast with optimized convolution operations | Slower due to attention computations | Linear with spatial dimensions | Extensive across mobile and embedded devices | Dynamic convolution for adaptive processing |
| Memory Requirements | Lower through parameter sharing | Higher from quadratic attention complexity | Manageable footprint | Specialized accelerators widely available | Knowledge distillation and pruning |
| Scalability | Efficient across resolutions | Quadratic growth with patch count | Controlled through pooling | Maps efficiently to GPUs and ASICs | Hierarchical designs and efficient attention |
| Deployment Platforms | Mobile devices to cloud systems | Requires powerful hardware | Optimized primitives | Emerging transformer-specific accelerators | Model compression and quantization |

Table 4. Computational Efficiency And Deployment Considerations [9, 10].

*Conclusion*

The systematic evaluation establishes clear overall performance limitations between vision transformers and convolutional neural networks under managed experimental situations. Architectural selection calls for careful consideration of dataset traits, computational constraints, and deployment requirements instead of conventional choices. Cnns stay superior picks for packages regarding confined training records, stringent computational budgets, or real-time inference requirements. Built-in inductive biases enable effective learning with modest sample sizes while computational efficiency supports deployment across diverse hardware platforms. The concentrated effective receptive field accelerates learning of local feature detectors through focused gradient flow. Dynamic convolution mechanisms extend CNN capabilities by adapting kernel parameters based on input characteristics without substantial parameter increases. Vision Transformers excel when substantial training data becomes available and computational resources permit deployment. Superior robustness to perturbations and enhanced generalization capabilities make transformers valuable for applications requiring reliable performance across distribution shifts. The self-

attention mechanism's flexibility enables learning complex spatial relationships beyond capabilities of fixed convolutional structures. Skip-attention approaches improve transformer efficiency by reducing redundant computations in attention layers. Findings reveal both architectural families occupy important niches in modern computer vision rather than viewing ViTs as universal CNN replacements. Practitioners should recognize complementary strengths when selecting architectures for specific applications. Hybrid architectures combining convolutional and attention-based operations represent promising directions potentially capturing advantages of both approaches. Future developments should investigate adaptive architectures balancing inductive biases with learned spatial relationships based on task characteristics and available data. Hardware capabilities continue evolving alongside training techniques potentially narrowing performance gaps on smaller datasets. Fundamental architectural trade-offs will likely persist despite technological advances. Evidence-based guidance provided advances theoretical understanding while supporting practical deployment of computer vision models across diverse application domains.

**References**

[1] Muhammad Shafiq and Zhaoquan Gu, "Deep Residual Learning for Image Recognition: A Survey," MDPI, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/18/8972

[2] Alaaeldin El-Nouby et al., "Training Vision Transformers for Image Retrieval," arXiv, 2021. [Online]. Available: https://arxiv.org/pdf/2102.05644

[3] Wenjie Luo et al., "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," 30th Conference on Neural Information Processing Systems, 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf

[4] Penghao Wang et al., "DATA EFFICIENT ANY TRANSFORMER-TO-MAMBA DISTILLATION VIA ATTENTION BRIDGE," arXiv, 2025. [Online]. Available: https://arxiv.org/pdf/2510.19266

[5] Xi Cheng, et al., "Pest identification via deep residual learning in complex background," Computers and Electronics in Agriculture, 2017. [Online]. Available: https://drive.google.com/file/d/11weeEdu6kePeqvksWjDW4Fxo-c5mjCOG/view?pli=1

[6] Wen Wang et al., "Towards Data-Efficient Detection Transformers," arXiv, 2022. [Online]. Available: https://arxiv.org/pdf/2203.09507

[7] KAZIM ALI et al., "Adversarial Robustness of Vision Transformers Versus Convolutional Neural Networks," IEEE Access, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10614176

[8] Shashanka Venkataramanan et al., "Skip-Attention: Improving Vision Transformers by Paying Less Attention," arXiv, 2023. [Online]. Available: https://arxiv.org/pdf/2301.02240

[9] Yikang Zhang et al., "DYNET: DYNAMIC CONVOLUTION FOR ACCELERATING CONVOLUTIONAL NEURAL NETWORKS," arXiv, 2020. [Online]. Available: https://arxiv.org/pdf/2004.10694

[10] Ashish Vaswani et al., "Attention is all you need," 31st Conference on Neural Information Processing Systems, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf