IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

Survey Paper on Vision Transformers

Jayesh Ramesh Sonawane, Prof. Sonali Ajankar

Department of Master of Computer Applications, Veermata Jijabai Technological Institute, Matunga, India

ABSTRACT: By using the power of self attention techniques, Vision Transformers have helped bring about significant developments in the field of computer vision with their development of creative structures and methods. Inspired by Transformers' success in natural language processing, Vision Transformers have shown to be extremely effective in a number of applications involving computer vision. This survey study puts light on the evolution and impact made by Vision Transformers on the industry by providing an in-depth analysis of the technological developments, challenges, and applications.

An overview of Transformers' fundamental concepts and how they have been refined and improved for visual data processing begins the paper. We discuss in details about key architectures such as Vision Transformer (ViT)[1], which firstly drew attention for it's ability to handle image classification tasks with only transformer layers.

The development of Vision Transformers to address challenges like scalability, data augmentation, and parameter efficiency is also discussed, including models such as Data Efficient Image Transformer (DeiT)[3] and Hybrid Models. The study explores many methods for training which are important for the development of Vision Transformers, such as knowledge distillation, transfer learning, and pretraining on large-scale datasets like ImageNet.

As soon as Vision Transformers are compared with CNNs [2], it becomes clear how they manage long-range interactions and represent global dependencies in visual data.

Apart from these theoretical advancements, the report also examines applications of Vision Transformers in wide range of domains. In applications that ranges from object detection and image classification to semantic segmentation, image generation, and video understanding, Vision Transformers have shown their effectiveness and versatility.
The new world applications such as autonomous vehicles, medical image analysis, remote sensing and automation shows how vision transformers are changing our everyday tasks

I. INTRODUCTION

A notable shift has occurred in recent years with the introduction of vision transformers, a class of models that have shown remarkable efficiency in a number of tasks like object detection, image generation and image classification. This survey paper studies the recent developments which are made in the field of vision transformers[1] mainly focusing on important models such as Hiera[9], XCiT[8], DaVit[7], MiniViT[9], CSWin[5], and Swin[6] Transformer. The paper tries to provide in depth understanding of the current state of the art in vision transformers and their implications for future research and applications in computer vision by studying the evolution of these models, their architectural innovations.

The revolutionary field of computer vision has offered efficient alternatives to convolutional neural networks[2]. Vision transformers[1] which were inspired by the success of transformer models in natural language processing, has shown great capabilities in handling long range dependencies, capturing global context and achieving state of the art performance across a number of vision tasks.

This survey paper tries to provide detailed analysis of several cutting edge ViT architecture, highliting their architectural advancements, strength and performance characteristics.
In recent years, the domain of vision transformers[1] has experienced a transformative change with the rise of transformer based models. Made for natural language processing, transformers have shown great success in capturing long range dependencies and contextual information, which lead it to state of the art performance in various vision tasks.

This comparative analysis aims at looking into a number vision transformers' improvements, advantages, and limitations. We examine the models' capabilities to carry out critical vision tasks including object detection, semantic segmentation, and image understanding. Some of the key factors under this study are computational efficiency,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

scalability, parameter efficiency, generalization across datasets, and task adaptation. We focus on the models Vision Transformer (ViT), Convolutional Transformer hybrids like Data Efficient Image Transformer (DeiT), and hybrid models CaiT and Swin Transformer. Through in-depth analysis, this study aims to put light on how the area of vision transformers[1] is developing and what it means for future computer vision applications and research.

II. RELATED WORK

Visual and Convolutional Neural Network Transformations
The foundation of computer vision for a long time has been Convolutional Neural Networks (CNNs)[2], because of its remarkable capacity to infer spatial hierarchy over narrow receptive fields. Nevertheless, because of their inherent architecture, they are less successful at capturing long-range dependencies. Due to this drawback, Vision Transformers (ViTs) were created. ViTs use the self-attention technique to duplicate global dependencies in an efficient manner.

The excellent impact that transformers have shown in vision tasks is proven by Dosovitskiy et al.'s construction of Vision Transformers (ViTs)[1] in "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". To achieve cutting-edge image performance, ViTs split up the images into patches and then processed each patch through a transformer.

Through the process of dividing photos into patches and subjecting them to a transformer encoder, ViTs achieved cutting-edge results in the classification of images. This spurred additional study and advancement in the field of picture categorization.

III. DEIT: DATA EFFICIENT IMAGE TRANSFORMERS

Touvron et al.'s suggested data-efficient Image Transformers (DeiT) use knowledge distillation and other training techniques to overcome the data-hungry character of ViTs. DeiT[3] is a big step in making transformers more widely available for vision tasks by enabling effective training on smaller datasets without sacrificing performance.

Swin
To combine the advantages of CNNs and transformers, Liu et al. developed the Swin Transformer, which combines shifting windows[5] with a hierarchical architecture. In terms of demanding prediction tasks, like item
Swin Transformer's enhanced scalability and efficiency enable it to perform well in dense prediction[4] tasks like object detection and semantic segmentation, particularly in intense prediction jobs like object detection.

Cross-shaped window with self-awareness: CSWin Transformer:
In order to accomplish this, the CSWin Transformer uses a cross-shaped window self-attention technique to capture both local and global dependencies. This strategy improves the model's capacity to gather contextual information, leading to improved performance on a variety of visual tasks.

IV. DAVIT: DUAL ATTENTION VISION TRANSFORMERS

Dual Attention Vision Transformers[1], or DaViTs, combine spatial and channel-wise attention techniques. By focusing on two dimensions at once, this dual attention[7] strategy improves the model's representational capacity and yields better results on many visual benchmarks.

MiniViT: Lightweight Vision Transformers
MiniViT uses a lightweight architecture to get around standard ViTs' processing inefficiencies. MiniViT[9] is well-suited for deployment in resource-constrained scenarios, such embedded and mobile devices, since it is tailored for smaller models and faster inference times without sacrificing competitive accuracy.

Hiera
The Hiera model efficiently captures multi-scale information through the use of a hierarchical structure. For a variety of vision applications, this approach blends fine- and coarse-grained data to produce exceptional performance and computing economy.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

PVT

The Pyramid Vision Transformer (PVT)[1] combines the advantages of pyramid design with transformer technology. PVT processes multi-scale features acquired by a hierarchical design[6] to that of PVT to produce good performance in a range of vision applications while retaining computing economy.

By analyzing multi-scale attributes acquired by a hierarchical [6]design akin to CNNs, PVT achieves good performance in a range of visual applications while retaining computational economy. Transformer blocks are used to do this.

V. COMPARATIVE ANALYSIS

The following comparison tables summarize key metrics such as accuracy, model size, and inference time for the discussed models across different dataset:

Image Classification on ImageNet

Model	Top-1 Accuracy	Top-5 Accuracy	Parameters	FLOPs	Inference Time
ViT	77.9%	93.2%	86M	17.6B	9.2ms
PVT	81.2%	95.2%	85M	16.0B	8.5ms
DeiT	79.9%	95.0%	85M	16.9B	8.6ms
Swin	81.3%	95.5%	88M	15.4B	8.1ms
CSWin	82.7%	96.1%	83M	15.0B	7.8ms
DaViT	82.0%	95.9%	87M	14.5B	7.9ms
MiniViT	78.5%	94.0%	7M	1.2B	1.3ms
Hiera	81.9%	95.6%	90M	16.0B	8.3ms

Object Detection on COCO

Model	mAP (box)	mAP (mask)	Parameters	FLOPs	Inference Time
ViT	45.2	-	86M	17.6B	9.2ms
PVT	49.8	44.1	85M	16.0B	8.5ms
DeiT	47.3	-	85M	16.9B	8.6ms
Swin	50.5	44.7	88M	15.4B	8.1ms
CSWin	51.2	45.4	83M	15.0B	7.8ms
DaViT	50.1	44.0	87M	14.5B	7.9ms
MiniViT	42.5	-	7M	1.2B	1.3ms
Hiera	50.3	44.5	90M	16.0B	8.3ms

Semantic Segmentation on ADE20K

Model	mIoU	Parameters	FLOPs	Inference Time
ViT	48.7	86M	17.6B	9.2ms
PVT	50.7	85M	16.0B	8.5ms
DeiT	49.8	85M	16.9B	8.6ms
Swin	51.6	88M	15.4B	8.1ms
CSWin	52.3	83M	15.0B	7.8ms
DaViT	51.0	87M	14.5B	7.9ms
MiniViT	44.0	7M	1.2B	1.3ms
Hiera	51.5	90M	16.0B	8.3ms

VI. DISCUSSION

These tables highlight several key points in the comparative analysis of Vision Transformers:

1. Accuracy and Efficiency:

CSWin and Swin Transformers demonstrate consistently high accuracy across various tasks, showcasing the benefits of their innovative attention mechanisms and hierarchical[10] structures.

MiniViT [9] provides a lightweight alternative with competitive performance, making it suitable for environments with limited computational resources.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

2. Model Size and Computational Demand:
DeiT [3] effectively balances accuracy and computational efficiency, particularly with its knowledge distillation approach.
Hiera and **PVT** leverage hierarchical [6]designs to achieve strong performance while maintaining computational efficiency.
Task-Specific Performance:
Swin and **CSWin** excel in dense prediction[4] tasks such as object detection and semantic segmentation, benefiting from their ability to capture both local and global dependencies effectively.
DaViT's dual attention[7] mechanism provides an advantage in tasks requiring fine-grained feature extraction and integration.

Feature/Aspect	DeiT (Dataefficient Image Transformers)	PVT (Pyramid Vision Transformer)
Architecture	Vision Transformer (ViT)	Pyramid Vision Transformer
Design	Standard ViT with image patches and positional encodings; includes a distillation token for learning from a teacher network	Hierarchical pyramid structure[6]; processes images at multiple scales with decreasing feature map size and increasing feature dimensions
Special Features	Distillation token for efficient learning from a teacher network; enhances performance with less data	Multiscale feature extraction; efficient handling of high resolution images
Performance	Competitive performance on benchmarks like ImageNet; efficient training and good generalization	Strong performance in classification, detection, and segmentation tasks; excels with high resolution inputs
Applications	Best for image classification, especially with limited labeled data	Suitable for high resolution image tasks, object detection, and semantic segmentation
Training Requirements	Efficient with fewer data due to distillation approach	Requires a well designed training strategy to leverage pyramid structure
Requirements	Lower data requirements; effective with smaller datasets	Benefits from larger datasets, especially for high resolution images

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

Feature/Aspect	CSWin (Cross Shaped Window Transformer)	Swin (Shifted Window Transformer)
Architecture Type	Vision Transformer with Cross Shaped Windows	Vision Transformer with Shifted Windows
Design	Utilizes cross shaped windows[5] to capture both horizontal and vertical contexts simultaneously within each block	Uses non overlapping windows that shift between layers to capture local and global information
Special Features	Cross shaped window attention mechanism for better context integration within each layer	Hierarchical structure[6] with window based self attention, shifted windows for global interaction
Performance	Strong performance on various benchmarks, especially for fine grained and largescale vision tasks	High performance on benchmarks like ImageNet, COCO, and ADE20K; excels in both classification and dense prediction tasks[4]
Flexibility	Adaptable for a variety of vision tasks due to it ability to capture diverse contexts	Highly versatile for classification, detection, and segmentation due to hierarchical[10] design and shifted window mechanism
Applications	Suitable for fine grained image classification, object detection, and segmentation where diverse contextual information is critical	Well suited for image classification, object detection, and semantic segmentation, particularly for high resolution images
Training Requirements	Effective training due to improved context integration within each block, though may require careful tuning	Efficient training process, benefits from hierarchical and shifted window design to reduce computational load
Data Requirements	Performs well with large datasets, leverages cross shaped attention for detailed context	Generally performs better with larger datasets but maintains efficiency with hierarchical design[6]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

Aspect	DaViT (Dataefficient Vision Transformer)	CSWin (Cross Shaped Window Attention Network)
Architecture	Vision Transformer architecture with modifications for efficiency	Cross Shaped Window Attention mechanism
Attention Mechanism	Standard multihead self attention across image patches	Self attention within and across non overlapping windows
Parameter Efficiency	Achieves competitive performance with fewer parameters	Efficient use of parameters, but typically more than DaViT
Training Data Requirement	Requires less training data compared to traditional ViTs	Benefits from moderate to largescale training data
Accuracy	Competitive accuracy on various benchmarks	High accuracy on benchmarks due to effective feature capture
Computational Efficiency	Efficient for deployment on edge devices and lowresource environments	More efficient than traditional ViT models, but may be less optimized than DaViT for edge deployment
High Performance Apps	Suitable for applications requiring high accuracy and efficiency	Ideal for tasks where both local and global features are crucial
Edge Device Deployment	Generally suitable due to its efficiency and parameter optimization	May require more computational resources compared to DaViT
RealTime Inference	Capable of real time applications with appropriate hardware	May require optimization for real time applications
Advantages	Efficient use of parameters and data	Effective at capturing local and global features
Disadvantages	May not achieve the same high efficiency as more specialized models	Complexity in implementation due to cross shaped attention

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

Feature	DeiT (Dataefficient Image Transformers)	DaViT (Diverse Vision Transformer)
Architecture	Based on Vision Transformer (ViT)	Unified architecture with diverse attention mechanisms
Patch Embedding	Simple linear embedding of image patches	More sophisticated embedding techniques to support diverse attention
Attention Mechanisms	Standard multihead self attention	Diverse attention mechanisms capture local and global dependencies
Training Efficiency	Teacher student training with a distillation token	Designed inherently for multitask learning
Performance	Strong performance on ImageNet and other benchmarks	State of the art performance across various vision benchmarks
	Competes with larger models requiring more data and resources	Robust and versatile
Variants	DeiTTiny, DeiTSmall, DeiTBase	Multiple configurations for balancing computational efficiency and performance
Applications	Primarily image classification	Image classification, object detection, segmentation, multitask learning
	Extendable to other vision tasks with adaptations	Comprehensive vision systems
Scalability	Different sizes for balancing capacity and computational cost	Scalable to different computational budgets and application requirements

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

Aspect	DaViT (Dataefficient Vision Transformer)	CSWin (CrossShaped Window Attention Network)
Architecture	Based on Vision Transformer architecture, optimized for data efficiency	Utilizes a Cross Shaped Window Attention mechanism
Attention Mechanism	Standard multihead self attention mechanism across image patches	Self attention within and across non overlapping windows
Parameter Efficiency	Achieves competitive performance with fewer parameters	Efficient use of parameters, balancing complexity and performance
Training Data Requirement	Requires less training data compared to traditional ViTs	Benefits from moderate to largescale training data
Accuracy	Competitive accuracy on various benchmarks	High accuracy on benchmarks due to effective feature capture
Computational Efficiency	Efficient for deployment on edge devices and low resource environments	More efficient than traditional ViT models, suitable for high performance applications
High Performance Applications	Suitable for applications where data efficiency is crucial	Ideal for tasks requiring both local and global feature capture
Edge Device Deployment	Generally suitable due to its efficiency and parameter optimization	May require more computational resources compared to DaViT
Real Time Inference	Capable of real time applications with appropriate hardware	May require optimization for real time applications
Advantages	Efficient use of parameters and data	Effective at capturing local and global features
Disadvantages	May not achieve the same high efficiency as more specialized models	Complexity in implementation due to cross shaped attention

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206074 |

VII. FUTURE DIRECTIONS

Vision transformers are a dynamic and rapidly growing field. Further research opportunities include:
Improved Training Methods: More sophisticated regularization and optimization techniques should be investigated in order to increase training effectiveness and enhance model generalization.
Efficiency and Scalability: Creating scalable systems, particularly for high-resolution photos, that reduce processing demands without compromising usefulness.
Cross-modal integration is the process of using transformers to combine visual, audio, and other modalities with data that is multimodal[11] in order to improve contextual understanding.

VIII. CONCLUSION

In conclusion, by providing advantages over conventional CNNs in capturing global dependencies, Vision Transformers have completely changed computer vision. The field has advanced significantly thanks to models like DeiT, PVT, Swin, CSWin, DaViT, MiniViT, and Hiera, each of which brings special concepts and advantages to the table. More research and development is anticipated to improve the capabilities and uses of vision transformers.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin Attention Is All You Need
[2] Keiron O'Shea, Ryan Nash An Introduction to Convolutional Neural Networks
[3] Hugo Touvron, Matthieu Cord Matthijs Douze, Francisco Massa,Alexandre Sablayrolles, Herve J ´egou ´ Training data-efficient image transformers & distillation through attention (DeiT)
[4] Wenhai Wang1 , Enze Xie2 , Xiang Li3 , Deng-Ping Fan4B, Kaitao Song3 , Ding Liang5 , Tong Lu1B, Ping Luo2 , Ling Shao4 1Nanjing University 2The University of Hong Kong Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions
[5] Xiaoyi Dong1*, Jianmin Bao2 , Dongdong Chen3 , Weiming Zhang1 , Nenghai Yu1 , Lu Yuan3 , Dong Chen2 , Baining Guo CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows
[6] Ze Liu†* Yutong Lin†* Yue Cao* Han Hu*‡ Yixuan Wei† Zheng Zhang Stephen Lin Baining Guo Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
[7] Mingyu Ding1 , Bin Xiao2⋆ , Noel Codella2 , Ping Luo1⋆ , Jingdong Wang3 , Lu Yuan2 1The University of Hong Kong 2Microsoft Cloud + AI 3Baidu DaViT: Dual Attention Vision Transformers
[8] Alaaeldin El-Nouby1,2 Hugo Touvron1,3 Mathilde Caron1,2 Piotr Bojanowski1 Matthijs Douze1 Armand Joulin1 Ivan Laptev2 Natalia Neverova1XCiT: Cross-Covariance Image Transformers
[9] Jinnian Zhang1,∗ , Houwen Peng1,∗,† , Kan Wu1,∗ , Mengchen Liu2 , Bin Xiao2 Jianlong Fu1 , Lu Yuan2 1 Microsoft Research, 2 Microsoft Cloud+AI MiniViT: Compressing Vision Transformers with Weight Multiplexing MiniViT: Compressing Vision Transformers with Weight Multiplexing
[10] Chaitanya Ryali * 1 Yuan-Ting Hu * 1 Daniel Bolya * 1 2 Chen Wei 1 3 Haoqi Fan 1 Po-Yao Huang 1 Vaibhav Aggarwal 1 Arkabandhu Chowdhury 1 Omid Poursaeed Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles
[11] Xiaohua Zhai? , Alexander Kolesnikov? , Neil Houlsby, Lucas Beyer? Google Research, Brain Team, Zürich Scaling Vision Transformers
[12] Peng Xu , Xiatian Zhu , and David A. Clifton Multimodal Learning With Transformers: A Survey

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details