Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Transformer-Based Deep Learning Architectures for Computer Vision: A Comprehensive Survey and Analysis

Bhavesh Prajapati

IT Department, L. D. College of Engineering,

Commissionerate of Technical Education, Government of Gujarat, India

Email: b.b.prajapati@gmail.com, ORCID: 0000-0002-8015-7934

ARTICLE INFO

Received: 01 Nov 2023

Revised: 19 Dec 2023

Accepted: 28 Dec 2023

ABSTRACT

The advent of transformer architectures has fun-damentally reshaped the landscape of deep learning, originally demonstrating remarkable success in natural language processing tasks before making substantial inroads into computer vision. This paper presents a comprehensive survey and analysis of transformer-based deep learning architectures and their applica-tions within computer vision domains. We systematically review key architectural innovations, including the Vision Transformer (ViT), Swin Transformer, and their numerous derivatives, exam-ining their design principles, strengths, and limitations compared to traditional convolutional neural network (CNN) approaches. Furthermore, we discuss the integration of self-attention mech-anisms into object detection, semantic segmentation, and image generation pipelines. We analyze performance benchmarks across widely adopted datasets such as ImageNet, COCO, and ADE20K, drawing comparative evaluations between transformer-based and CNN-based models. Additionally, we explore training strategies, data augmentation techniques, and transfer learning paradigms that underpin the success of these architectures. Finally, we highlight open research challenges and identify promising fu-ture directions, including efficient attention mechanisms, hybrid architectures, and self-supervised pre-training. This survey aims to serve as a structured reference for researchers and practi-tioners seeking a rigorous understanding of the current state of transformer-based deep learning for computer vision.

Keywords: deep learning, transformer, vision transformer, self-attention, convolutional neural networks, computer vision, image classification, object detection, semantic segmentation

Introduction

Deep learning has undergone a paradigm shift since the resurgence of neural networks with AlexNet [1], which demonstrated the power of deep convolutional neural networks (CNNs) on large-scale image recognition tasks. Subsequent advances such as VGGNet [2], GoogLeNet, and ResNet [3] further cemented CNNs as the dominant paradigm for com-puter vision for nearly a decade. However, the introduction of the Transformer architecture by Vaswani et al. [4] in the natural language processing (NLP) domain ushered in a new era of sequence modeling based entirely on self-attention mechanisms, dispensing with the recurrence and convolution operations that had previously defined the state of the art.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The success of transformers in NLP—exemplified by mod-els such as BERT [5] and GPT-3 [6]—naturally prompted re-searchers to investigate whether analogous architectures could achieve comparable results in vision tasks. The pivotal work of Dosovitskiy et al. [7], which introduced the Vision Trans-former (ViT), demonstrated that a pure transformer applied directly to sequences of image patches can attain competi-tive performance on image classification benchmarks when trained on sufficiently large datasets. This finding catalyzed an explosion of research into transformer-based vision models, culminating in architectures such as the Swin Transformer [9], DeiT [8], and ConvNeXt [10], among many others.

The significance of this transition cannot be overstated. Transformers introduce a fundamentally different inductive bias compared to CNNs: whereas CNNs exploit local spatial correlations through sliding convolutional filters, transformers model global dependencies from the outset via the attention mechanism. This capability to capture long-range interactions has proven advantageous in tasks requiring holistic under-standing of scenes, such as object detection and semantic segmentation, as evidenced by the Detection Transformer (DETR) [12] and its successor Deformable DETR [13].

Despite these advances, transformer-based models present distinct challenges. They typically require substantially larger training datasets, incur higher computational costs due to the quadratic complexity of self-attention with respect to sequence length, and are more difficult to optimize than their CNN coun-terparts. Addressing these limitations has motivated research into efficient attention mechanisms, hierarchical architectures, and self-supervised pre-training strategies [11].

This paper makes the following contributions:

- A structured taxonomy of transformer-based deep learn-ing architectures for computer vision, covering image classification, object detection, semantic segmentation, and image generation.
- A comprehensive comparative analysis of transformer and CNN models across standard benchmarks, highlighting trade-offs in accuracy, computational efficiency, and data requirements.
- An examination of training methodologies, including data augmentation, knowledge distillation, and self-supervised pre-training.
- A discussion of open research challenges and prospective directions for future work.

The remainder of this paper is organized as follows. Sec-tion II reviews foundational concepts. Section III surveys key transformer architectures. Section IV examines downstream applications. Section V discusses training strategies. Sec-tion VI presents comparative analysis. Section VII addresses research challenges, and Section VIII concludes the paper.

Background and Foundations

A. Convolutional Neural Networks

CNNs have dominated computer vision for over a decade. The foundational operations—convolution, pooling, and non-linear activation—provide translation equivariance and local feature extraction, which are highly effective for visual recog-nition tasks. The introduction of residual connections by He et al. [3] solved the vanishing gradient problem, enabling the training of very deep networks and significantly improving performance on the ImageNet benchmark. Batch normalization

[21] further stabilized training by reducing internal covariate shift, becoming a standard component in modern CNN archi-tectures.

B. The Attention Mechanism

The attention mechanism, originally proposed for neural machine translation [22], allows models to

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

dynamically weight the importance of different parts of the input when producing an output. The scaled dot-product attention formulated by Vaswani et al. [4] computes attention weights as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. Multi-head attention extends this by projecting Q, K, and V into h different subspaces and computing attention in parallel, enabling the model to capture diverse relationships simultaneously.

C. The Transformer Architecture

The original Transformer [4] consists of an encoder-decoder structure, where each encoder layer comprises a multi-head self-attention module followed by a position-wise feed-forward network (FFN), with residual connections and layer normalization applied around each sub-layer. Positional encodings are added to the input embeddings to inject se-quence order information, since self-attention is inherently permutation-invariant. This design proved remarkably effective for sequence-to-sequence tasks and established the blueprint for subsequent language models.

Transformer Architectures for Vision

A. Vision Transformer (ViT)

Dosovitskiy et al. [7] introduced the Vision Transformer (ViT) by reformulating image classification as a sequence modeling problem. An input image of resolution $H \times W$ is divided into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$, each linearly projected into a D-dimensional em-bedding space. A learnable [CLS] token is prepended to the patch sequence, and 1D positional embeddings are added prior to feeding the sequence into a standard transformer encoder. Classification is performed using the final state of the [CLS] token.

ViT demonstrated that, given sufficient pre-training data (e.g., JFT-300M), it could match or surpass state-of-the-art CNN models on ImageNet. However, when pre-trained only on ImageNet-21k or smaller datasets, ViT underperformed ResNets of comparable size, suggesting that transformers have weaker inductive biases and require more data to learn effective visual representations.

B. Data-Efficient Image Transformers (DeiT)

To reduce the data requirements of ViT, Touvron et al. [8] proposed Data-Efficient Image Transformers (DeiT). The key innovation is a knowledge distillation strategy using a distil-lation token, analogous to the [CLS] token, which is trained to mimic the output of a teacher CNN (e.g., RegNet). This approach enables ViT-scale models to be trained effectively on ImageNet-1k alone, achieving 85.2% top-1 accuracy, thereby democratizing the use of vision transformers without requiring access to massive proprietary datasets.

C. Swin Transformer

The Swin Transformer [9] addressed two key limitations of ViT: the fixed patch size and the quadratic attention complexity with respect to image resolution. It introduced a hierarchical feature map structure analogous to CNNs, where patch tokens are progressively merged to reduce resolution and increase channel dimension across stages. Crucially, attention is com-puted within non-overlapping local windows of fixed size, reducing the complexity to $O(n)$ with respect to image size. Shifted windows in alternating layers enable cross-window information flow while

maintaining computational efficiency. Swin Transformer achieved state-of-the-art performance on ImageNet classification (87.3% top-1 with Swin-L) and served as a versatile backbone for object detection and segmentation tasks, demonstrating clear improvements over prior art on the COCO and ADE20K benchmarks.

D. ConvNeXt

Liu et al. [10] revisited the design of pure CNN architectures in light of lessons from transformers, proposing ConvNeXt. By systematically modernizing a ResNet-50 baseline—including the adoption of depthwise convolutions, an inverted bottleneck design, larger kernel sizes (7 7), and layer normalization in place of batch normalization—the authors demonstrated that CNNs remain highly competitive with transformers on vision benchmarks. ConvNeXt-L achieved 87.5% top-1 accuracy on ImageNet, emphasizing that architectural design choices, rather than the attention mechanism per se, drive much of the performance gain.

E. Masked Autoencoders (MAE)

He et al. [11] introduced Masked Autoencoders (MAE) as a scalable self-supervised pre-training framework for vi-sion transformers. Inspired by masked language modeling in BERT, MAE randomly masks a large proportion (75%) of input patches and trains a transformer encoder-decoder to reconstruct the original pixel values. The asymmetric de-sign—encoding only visible patches and decoding the full sequence using a lightweight decoder, yields highly efficient pre-training. The Fine-tuned ViT-H / 14 achieved 87.8% accuracy on ImageNet, establishing MAE as a powerful self-supervised pre-training strategy.

Downstream Applications

A. Object Detection

The Detection Transformer (DETR) [12] represented a paradigm shift in object detection by eliminating hand-made components such as anchor generation and non-maximum suppression (NMS). DETR formulates detection as a set prediction problem, using a transformer encoder-decoder and bipartite matching loss. Although DETR matched the perfor-mance of Faster R-CNN on COCO, it suffered from slow convergence and poor performance on small objects.

Deformable DETR [13] addressed these issues by intro-ducing deformable attention modules that attend to a sparse set of key sampling points around a reference point, reducing computational cost and significantly accelerating convergence. These contributions have made transformer-based detectors competitive with anchor-based and anchor-free CNN detectors across the COCO benchmark.

B. Semantic Segmentation

Transformers have been extensively applied to semantic bridging the gap between vision and language modalities and enabling zero-shot transfer to downstream tasks.

Training Strategies

A. Data Augmentation

Modern training recipes for vision transformers rely heavily on aggressive data augmentation. Standard techniques include RandAugment, CutMix, MixUp, and Random Erasing, which collectively improve generalization and reduce overfitting, par-ticularly when training on smaller datasets such as ImageNet-1k. DeiT [8] demonstrated that these techniques are essential for data-efficient training of transformers.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

B. Knowledge Distillation

Knowledge distillation [23] transfers knowledge from a large teacher model to a smaller student model, producing compact models with minimal accuracy loss. In the context of vision transformers, token-based distillation [8] has proven highly effective, with CNN teachers providing complementary inductive biases that benefit transformer students.

C. Transfer Learning and Fine-Tuning

Pre-training on large-scale datasets (ImageNet-21k, JFT-300M) followed by fine-tuning on target tasks remains the dominant paradigm. ViT variants fine-tuned at higher resolu-tions (e.g., 384 384) consistently outperform those evaluated at the pre-training resolution. Self-supervised pre-training via MAE [11] and MoCo v3 [24] offers a compelling alternative to supervised pre-training, often achieving comparable or superior performance on dense prediction tasks.

Comparative Analysis

Table I summarizes the performance of representative mod-els on the ImageNet-1k benchmark. Table II presents object detection results on COCO val2017using standard metrics.

TABLE I

IMAGENET-1K TOP-1 ACCURACY COMPARISON

Model	Params (M)	FLOPs (G)	Top-1 (%)	Year
ResNet-50	25	4.1	76.1	2016
ResNet-152	60	11.6	78.3	2016
ViT-B/16	86	17.6	81.8	2021
DeiT-B	86	17.6	83.4	2021
Swin-T	29	4.5	81.3	2021
Swin-B	88	15.4	83.5	2021
Swin-L	197	34.5	87.3	2021
ConvNeXt-B	89	15.4	83.8	2022
ConvNeXt-L	198	34.4	87.5	2022
MAE (ViT-H)	632	–	87.8	2022

segmentation. The Segmentation Transformer (SETR) [18] replaced the conventional CNN encoder with a ViT encoder, using multiple decoder designs to produce dense predictions. More recent architectures such as Segmenter [19] employed mask embedding transformer decoders, while SegFormer [20] proposed a hierarchical mix transformer encoder combined with a lightweight MLP decoder, achieving excellent accuracy-efficiency trade-offs on ADE20K and Cityscapes.

C. Image Generation

Generative Adversarial Networks (GANs) [14] established the foundation for deep generative modeling. The integration of self-attention into GAN frameworks, as in SAGAN [15], enabled the generation of globally coherent high-resolution images by allowing the generator to attend to long-range dependencies. More recently, contrastive learning frameworks such as SimCLR [16] and CLIP [17] have demonstrated the power of self-supervised and vision-language pre-training, The results in Table I reveal several important trends. First, the Swin Transformer achieves a favorable accuracy-efficiency trade-off compared to plain ViT, largely due to its hierarchical design and local window attention. Second, ConvNeXt demonstrates that carefully designed CNNs can match transformers at comparable model sizes, challenging the assumption that the attention mechanism is necessary for

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE II

COCO VAL2017 OBJECT DETECTION RESULTS (AP^{BOX})

Model	Backbone	AP	Year
Faster R-CNN	ResNet-50	40.2	2015
DETR	ResNet-50	42.0	2020
Deformable DETR	ResNet-50	46.2	2021
Swin-T (Cascade)	Swin-T	50.5	2021
Swin-L (Cascade)	Swin-L	57.7	2021
ConvNeXt-L (Cascade)	ConvNeXt-L	57.9	2022

state-of-the-art visual recognition. Third, self-supervised pre-training via MAE yields the highest reported accuracy, under-scoring the potential of large-scale self-supervised learning as an alternative to supervised pre-training.

On the object detection benchmark (Table II), transformer-based detectors show clear advantages, with Swin-based Cas-cade detectors exceeding CNN-based counterparts by sub-stantial margins. The Deformable DETR provides a strong efficiency-accuracy trade-off relative to the original DETR, validating the importance of sparse attention for dense pre-diction tasks.

Open Research Challenges

A. Computational Efficiency

The quadratic complexity of standard self-attention remains a fundamental bottleneck for high-resolution inputs. While local window attention in the Swin Transformer mitigates this issue, achieving truly global attention efficiently remains an open problem. Approaches such as linear attention and sparse attention provide partial solutions but often sacrifice expressiveness.

B. Data Hunger

Despite advances in self-supervised pre-training, transform-ers generally require more training data than CNNs with equivalent inductive biases. Developing training recipes that generalize effectively from small or medium-scale datasets remains critical for real-world deployment, particularly in data-scarce domains such as medical imaging.

C. Interpretability and Robustness

The attention maps produced by vision transformers of-fer some interpretability, but rigorous characterization of the features learned by these models is lacking. Furthermore, the robustness of transformers to adversarial perturbations, distribution shifts, and common image corruptions requires further investigation [25].

D. Hybrid Architectures

Hybrid models that integrate convolutional and attention layers have shown promise in combining the strengths of both paradigms. Understanding how to optimally combine local and global processing across model scales remains an active area of inquiry. *Efficient Fine-Tuning*

Full fine-tuning of large pre-trained transformers is compu-tationally expensive. Parameter-

efficient fine-tuning methods, such as adapter layers and prompt tuning, offer compelling alternatives but require further study in the vision domain.

Conclusion

This paper has presented a comprehensive survey of transformer-based deep learning architectures for computer vision. Beginning from the foundational attention mechanism and the original Transformer model, we traced the evolution of vision-specific architectures—from ViT and DeiT to the Swin Transformer and ConvNeXt—and examined their applications in image classification, object detection, semantic segmenta-tion, and image generation. Our comparative analysis demon-strated that transformer-based models have achieved state-of-the-art performance across multiple vision benchmarks, while also identifying conditions under which CNN-based approaches remain competitive.

Key findings include: (1) hierarchical transformer designs with local window attention offer significant efficiency gains over plain vision transformers; (2) self-supervised pre-training via masked autoencoders is a powerful alternative to super-vised pre-training; (3) carefully modernized CNNs can achieve performance comparable to transformers, highlighting the im-portance of architectural design choices; and (4) transformer-based object detectors have largely superseded traditional anchor-based detectors on standard benchmarks.

Despite these advances, important challenges remain, in-cluding computational efficiency at high resolutions, data requirements, robustness, and interpretability. We anticipate that future work will focus on efficient attention mechanisms, hybrid convolutional-attention architectures, and scalable self-supervised learning to further close the gap between trans-former capabilities and real-world deployment requirements.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Infor-mation Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Rep-resentations (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computa-tional Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, pp. 4171–4186, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[6] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[7] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at

scale," in *Proc. Int. Conf. Learning Represen-tations (ICLR)*, Virtual, 2021. [Online]. Available: https://arxiv.org/abs/ 2010.11929

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Je´gou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Machine Learning (ICML)*, Virtual, vol. 139, pp. 10347–10357, 2021.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted win-dows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10012–10022, 2021.

[10] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 11976–11986, 2022.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dolla´r, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 16000–16009, 2022.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conf. Computer Vision (ECCV)*, Glasgow, UK, pp. 213–229, 2020.

[13] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, Virtual, 2021. [Online]. Available: https://arxiv.org/abs/2010.04159

[14] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 2672–2680, 2014.

[15] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, vol. 97, pp. 7354–7363, 2019.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Machine Learning (ICML)*, Virtual, vol. 119, pp. 1597–1607, 2020.

[17] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, Virtual, vol. 139, pp. 8748–8763, 2021.

[18] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Virtual, pp. 6881–6890, 2021.

[19] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Trans-former for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 7262–7272, 2021.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 12077–12090, 2021.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learning (ICML)*, Lille, France, vol. 37, pp. 448–456, 2015.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Rep-resentations (ICLR)*, San Diego, CA, USA, 2015.

[Online]. Available: https://arxiv.org/abs/1409.0473

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NeurIPS Workshop on Deep Learning and Representation Learning*, Montreal, QC, Canada, 2014. [Online]. Available: https://arxiv.org/abs/1503.02531

[24] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9640–9649, 2021.

[25] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10231–10241, 2021.