Doupe Journal of Top Trending Technologies, Vol 1(2), July 2025, PP, 36-45

Transformer-Based Models in Image Segmentation and Classification: A New Era in Vision AI

Atako, Nelson Rachael. *1

[1]Dept of Public Administration. University of Nigeria Nsukka. (Policy Analysis PhD in view)

Abstract: Over the past decade, deep learning has revolutionized computer vision, with convolutional neural networks (CNNs) dominating tasks like image classification and segmentation. However, a new paradigm emerged as transformer-based models – originally developed for natural language processing – have begun to surpass previous CNN-based approaches across vision tasks. This marks a new era in Vision AI, where transformers' ability to capture long-range dependencies and global context is reshaping how we design vision systems. Transformer models have achieved state-of-the-art performance in image classification (assigning labels to entire images) and segmentation (partitioning images into labeled regions), often with simpler pipelines and stronger results than their CNN predecessors.

***Keywords*:** Transformer, Vision Transformer (ViT), Image Classification, Image Segmentation, Convolutional Neural Networks (CNN), Attention Mechanism, Hybrid Models.

1 Introduction

This review provides an in-depth literature survey (2015–2025) on transformer-based models in image segmentation and classification. We highlight key milestones and architectural innovations that enabled transformers to thrive in computer vision. We also compare these transformer models to traditional CNNs and to hybrid architectures that combine convolution and attention. Performance benchmarks on standard datasets are discussed, along with evaluation metrics and remaining challenges. The article is organized as a structured IEEE-style review with sections covering Background, Transformer Architectures, Applications in Segmentation, Applications in Classification, Hybrid Models, Benchmarks, Challenges, and Future Directions. Throughout, we cite representative works from peer-reviewed conferences (e.g., CVPR, ICCV, NeurIPS) and journals (IEEE/ACM), as well as influential arXiv preprints, to trace the evolution of this rapidly developing field.

2 Background

2.1 CNN dominance (2015–2020)

In the mid-2010s, CNNs became the de facto standard for vision tasks. For image classification, architectures such as ResNet (2015) [1] and EfficientNet (2019) [2] achieved remarkable accuracy on ImageNet and other benchmarks, leveraging convolutional layers' inductive biases (local receptive fields and weight sharing) to learn hierarchical feature representations. Similarly, for image segmentation, CNN-based fully convolutional networks (e.g. FCN and subsequent models like U-Net and DeepLab) brought breakthroughs by performing pixel-wise labeling with convolutional encoders and decoders [3]. These CNN approaches excelled at learning from relatively small datasets due to built-in assumptions such as as translation equivariance and locality, which act as useful inductive biases. However, CNNs struggle to capture long-range dependencies because convolution kernels and pooling operate locally; global context only emerges after many layers. As a result, complex scenes requiring understanding of distant relationships could be challenging for pure CNNs.

*1Corresponding Author Email: onyemauche.inweregbuh@unn.edu.ng

https://www.doupe.in | DOI: https://doi.org/10.71063/DJTTT.2025.1205
Creative Commons CC BY-ND 4.0 License

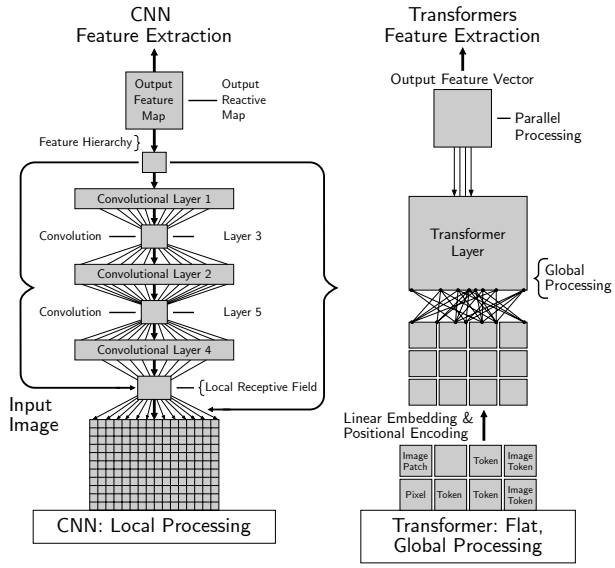

Figure 1: CNN vs. Transformer Feature Extraction

2.2 Attention in vision – early attempts

Inspired by the success of self-attention in NLP (e.g. the 2017 Transformer model), researchers began integrating attention mechanisms into CNN architectures. For example, Non-local Neural Networks [4] introduced a self-attention module into CNNs to capture global correlations, improving video classification and image segmentation results. In segmentation, PSANet applied point-wise spatial attention to model long-range pixel relationships [5]. Such hybrid approaches augmented CNN feature maps with attention-driven context modeling. Subsequent works went further: some replaced certain convolutional layers entirely with self-attention layers. Stand-alone self-attention networks attained competitive image recognition accuracy without convolutions [6]. Similarly, Local Relation Networks used learned local attention in place of conv filters to better model interactions among neighboring pixels [7]. These experiments hinted that with proper design, self-attention could match or exceed CNN performance, albeit they still used limited attention scope or required more computation.

2.3 Vision Transformer breakthrough

The watershed moment came in 2020 with the introduction of the Vision Transformer (ViT) [8]. ViT was the first pure transformer applied to image classification: it divides an image into a sequence of patch tokens and processes them with a transformer encoder, much like words in a sentence. Remarkably, ViT achieved state-of-the-art accuracy on image classification benchmarks, provided it was pre-trained on sufficient data. For instance, a ViT model pre-trained on a large private dataset (JFT-300M) and fine-tuned on ImageNet matched or exceeded the best CNNs of similar size. This demonstrated the scalability of transformers – with enough training data, their lack of inductive bias can be overcome, and their representational capacity may surpass CNNs. In parallel, DETR (Detection Transformer) applied a transformer in object detection by using a set of learned object queries in a transformer decoder [9]. DETR eliminated the need for hand-crafted components like anchor boxes, simplifying the detection pipeline and even enabling a direct segmentation output via predicted masks. These two works were key milestones signaling transformers' potential in vision. By 2021, vision researchers rapidly embraced and extended these ideas to numerous tasks.

2.4 Early transformer-based segmentation

Building on ViT and DETR, researchers applied transformers to semantic and instance segmentation. Notably, SETR was the first to replace a CNN backbone with a pure transformer encoder (ViT) for

semantic segmentation [9]. SETR fed image patches into a transformer and then used a simple decoder to predict pixel-wise labels, achieving state-of-the-art results on the ADE20K dataset. Around the same time, MaX-DeepLab introduced an end-to-end panoptic segmentation model with a "mask transformer" decoder, using learned query vectors to directly predict segmentation masks for entire objects [10]. MaX-DeepLab was one of the first fully transformer-driven segmentation frameworks, and it achieved top results on COCO panoptic segmentation. These advances confirmed that transformer architectures could handle dense prediction tasks, not just image-level classification.

In summary, by 2021 the field had shifted from CNN-only solutions to a new paradigm where transformers play a central role. In the following sections, we detail the architectures of vision transformers, their application to segmentation and classification tasks, how they compare to CNNs, and how hybrid CNN-transformer models have combined the best of both worlds.

3 Transformer Architectures in Vision

Transformers were first developed for sequence modeling in NLP, and adapting them to images required rethinking how to represent spatial data as sequences. The vision transformer architecture generally follows the encoder of the original Transformer [11] but with modifications for 2D images. Key components include patch embedding, positional encoding, multi-head self-attention layers, and feed-forward networks, as described below.

3.1 Patch Embedding

Instead of pixels, ViT and similar models operate on image patches as input tokens. An input image (e.g. 224×224) is divided into fixed-size patches (e.g. 16×16 pixels), yielding a grid of patches (14×14 = 196 patches in this example). Each patch is flattened into a vector and passed through a linear projection (learned fully-connected layer) to produce a patch embedding of a desired dimension (e.g. 768). This process is analogous to word embeddings in NLP. All patch embeddings are of equal dimension D, which remains constant throughout the transformer layers. The set of embedded patches constitutes the input sequence for the transformer [12].

3.2 Positional Encoding

Because transformers are permutation-invariant to input order, we must inject information about patch positions. ViT adds a learned position embedding vector to each patch embedding, indicating its location in the 2D grid. Dosovitskiy et al. experimented with various 2D positional encodings (learned vs. sinusoidal) and found simple learnable 1D encodings (flattening 2D positions) to work well [13]. Some later models use relative positional embeddings or convolutional tokenization to encode spatial structure implicitly, as discussed shortly. Additionally, ViT introduces a special learnable classification token ([CLS]), which is prepended to the sequence of patch embeddings. This extra token (analogous to BERT's [CLS]) serves to aggregate global information from all patches via self-attention. At the output, the transformer's representation of this class token is used for image-level predictions [14].

3.3 Transformer Encoder Layers

The core of the model is a stack of transformer encoder blocks (e.g. 12 or 24 layers) [15]. Each layer consists of multi-head self-attention (MHSA) followed by a feed-forward network (two linear layers with an activation, often called an MLP block) [16]. Layer normalization and residual connections are employed around both the attention and MLP sub-layers, as in the standard transformer. Multi-head self-attention allows the model to attend to relationships between patches: each attention head computes weighted interactions among all patch embeddings (plus the class token) using learned query, key, and value projections. Multiple heads (each a different learned projection) capture diverse aspects of patch relationships, and their outputs are concatenated [17]. This self-attention mechanism enables global receptive field from the first layer – every patch can potentially interact with any other patch in the image in one attention step. This is a stark contrast to CNNs, where the receptive field grows layer by layer and

https://www.doupe.in | DOI: https://doi.org/10.71063/DJTTT.2025.1205
Creative Commons CC BY-ND 4.0 License

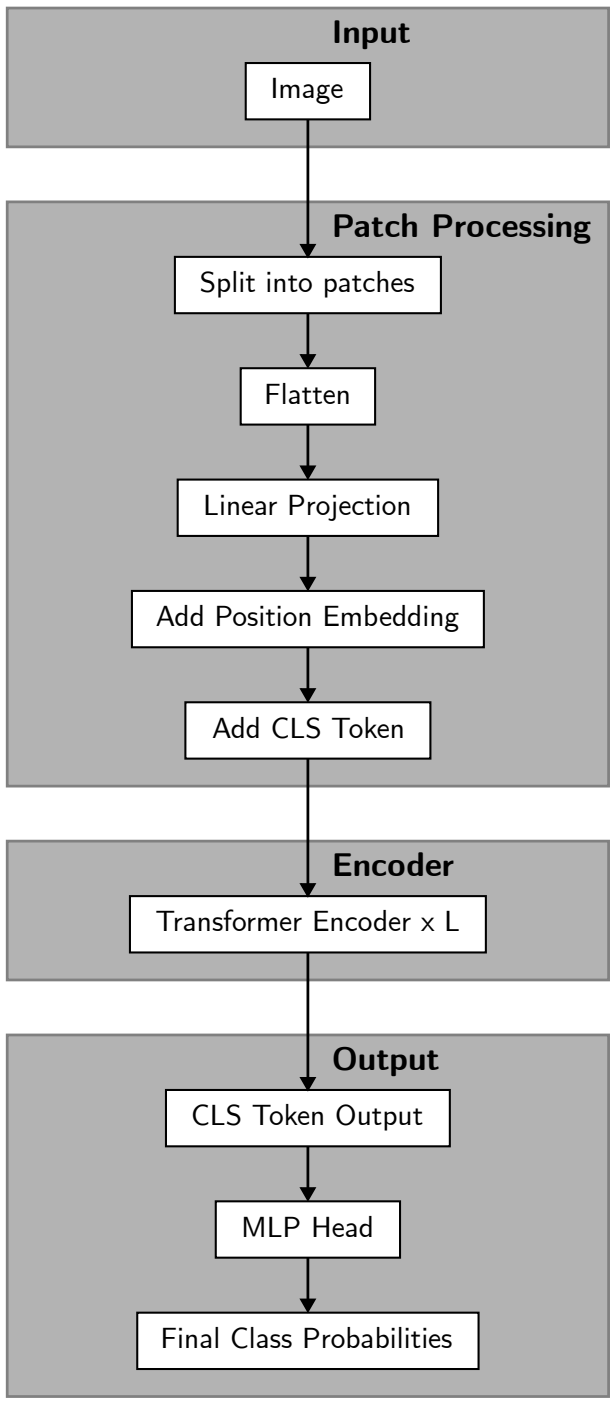

Figure 2: Vision Transformer (ViT) Architecture Flow

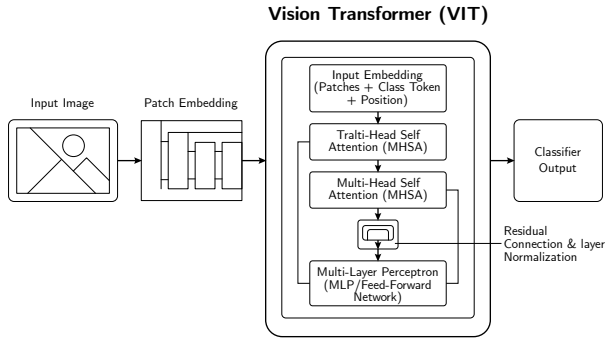

Figure 3: Vision Transformer (ViT) Architecture Flow

is inherently local at first. By stacking multiple MHSA layers, the transformer can build complex global feature representations. Notably, the computational cost of self-attention is quadratic in the number of patches ($O(N^2)$ for N patches) [18]. For typical image sizes, N is on the order of a few hundred (e.g. 196 for 224×224 with 16×16 patches), which is manageable, but for higher resolutions or dense prediction (where one might not want to downsample too much), this cost becomes a concern (addressed by later variants).

3.4 Classification/Output Layer

After the final encoder layer, the transformer's output corresponding to the [CLS] token is fed to a classification head (a simple feed-forward layer or small MLP) to predict the image class [11]. The patch outputs can be ignored for classification, as all necessary information is meant to be condensed in the class token representation. In tasks like segmentation, however, the outputs for all patch tokens (or an upsampled version of them) are used to produce a spatial prediction.

3.5 Hierarchical and Efficient Vision Transformers

The original ViT used a single-scale sequence of fixed-size patches. Subsequent architectures introduced hierarchical structures similar to CNN feature pyramids. For example, the Swin Transformer builds a hierarchy of tokens by merging patches progressively (analogous to pooling in CNNs) and limits attention to local windows that shift across layers [11]. Swin's hierarchical design yields multi-scale feature maps (high-resolution shallow layers, low-resolution deep layers) useful for downstream tasks like detection and segmentation, while its shifted window attention scheme reduces complexity from quadratic in the image to roughly linear, since self-attention is computed within small windows. Many other variants followed a similar motivation: integrating CNN-like locality and pyramid structure for efficiency and inductive bias. Pyramid Vision Transformer (PVT), CVT (Convolutional ViT), and CeiT in 2021 all employed multi-stage transformers or convolutional tokenizers to better capture local structures and reduce the sequence length as depth increases. These designs echo CNNs (which progressively reduce spatial resolution and increase channel depth) but still use self-attention instead of convolution for feature mixing[19].

4 Applications in Image Segmentation

Image segmentation is a dense prediction task where the model assigns a label (e.g. object class) to each pixel in an image. Transformers have disrupted this area by providing powerful global context modeling and a new paradigm for mask prediction.

4.1 Transformers as segmentation backbones

A straightforward way to apply transformers to segmentation is to use a transformer encoder in place of the CNN encoder. The aforementioned SETR is a prime example [20]. SETR took a ViT (pre-

trained on ImageNet) as the feature extractor, then added a lightweight decoder that upsamples the transformer's patch-based outputs to full resolution for pixel classification. Similarly, Segmenter and SegFormer proposed efficient transformer backbones tailored for segmentation [21, 22]. SegFormer uses a hierarchical transformer and a simple multilayer perceptron (MLP) decoder, yielding both accuracy and efficiency improvements. Using vision transformers as backbones provides rich global features that improve segmentation quality, especially in scenes where context is important for labeling a pixel correctly.

4.2 Query-based mask prediction

A different approach to segmentation leverages the set prediction capabilities of transformers. MaX-DeepLab was a seminal work in this direction: it employed a dual-path architecture (CNN + transformer) where the transformer produces a set of mask embeddings (queries) that directly output full-resolution masks. Building on this, MaskFormer showed that a single method can tackle semantic, instance, and panoptic segmentation by predicting a set of masks with associated class labels [23]. An improved version, Mask2Former, introduced masked attention in the transformer decoder: queries attend only to relevant image regions, refining the masks in iterative decoder layers [23]. Mask2Former, combined with a strong backbone (e.g. a Swin-L transformer), achieved cutting-edge results on COCO and ADE20K.

4.3 Vision transformers for specific segmentation domains

Transformers have also been embraced in specialized segmentation problems. In medical image segmentation, hybrid architectures (CNN + ViT) are popular. For example, TransUNet kept a CNN encoder for low-level feature extraction but added a ViT encoder on top to capture global context [23]. This improved segmentation of organs in 3D scans by modeling long-range dependencies that CNNs might miss. We see a similar pattern in remote sensing and other domains: Swin-Unet and other variants have been proposed to segment satellite imagery, leveraging transformers for global context and CNN-like layers for detail.

4.4 Performance and benchmark progress

The introduction of transformers led to rapid improvements on segmentation benchmarks. On the popular Cityscapes dataset, transformer-based models quickly matched the best CNNs. On ADE20K, a more challenging benchmark, the best CNNs reached 45-47% mIoU. Transformers broke through this ceiling: SETR reached 50%, and by 2022, advanced models like SeMask and DAT crossed 58% mIoU [24]. These results underscore that transformers not only simplify segmentation architectures but also deliver superior accuracy by virtue of their global modeling capabilities.

5 Challenges and Limitations

Despite their impressive performance and rapid adoption, transformer-based models in vision face several enduring challenges and open issues that influence their research and deployment trajectories.

1. **Data Requirements and Inductive Bias:** Vision transformers (ViTs) generally require very large training datasets to achieve competitive results from scratch because they lack the strong locality and translation equivariance biases inherent in convolutional neural networks (CNNs). While pretraining on massive datasets like ImageNet-21k or JFT mitigates this, such resources are inaccessible to many practitioners. Techniques like hybrid Conv+Attention designs, data augmentation, and distillation help but do not completely close the gap in low-data regimes.

2. **Computational Complexity:** The vanilla multi-head self-attention mechanism has a quadratic complexity $\mathcal{O}(N^2 d)$ with respect to the number of tokens N, which scales poorly for high-resolution images common in segmentation and medical imaging. This leads to prohibitive memory and compute requirements, especially during training. Approaches like windowed attention (Swin), token pruning, and low-rank approximations help but can trade off global context modeling.

3. **Lack of Spatial Hierarchy:** Plain ViT models treat an image as a flat sequence of tokens, losing multi-scale spatial context early in the pipeline. This can hinder performance on dense prediction tasks such as segmentation or detection, where fine-grained spatial relationships are crucial. Hierarchical transformers (PVT, Swin) address this, but they add complexity to the design space.

4. **Optimization and Training Dynamics:** Transformers are sensitive to hyperparameters such as learning rate, weight decay, and warm-up schedules. They often require more careful tuning than CNNs, particularly in low-data settings. Large-batch training, gradient clipping, and adaptive optimizers (e.g., AdamW) are common remedies, but these add engineering complexity.

5. **Interpretability:** While attention maps offer a direct way to visualize learned relationships, they are not always faithful explanations of the model's decision process [25]. Attention can highlight regions unrelated to the actual prediction, and multiple attention heads can focus on disparate, non-intuitive features. Interpretability in vision transformers remains an active and unresolved research area.

6. **Overfitting and Generalization:** Large transformers pretrained on extensive datasets tend to generalize well, but smaller models or those trained from scratch on limited datasets can overfit quickly. Regularization techniques such as dropout, stochastic depth, and label smoothing are necessary but may still fall short without large-scale pretraining [26].

7. **Domain Adaptation:** Adapting pretrained transformers to specialized domains (e.g., remote sensing, medical imaging) with scarce labeled data can be challenging. Domain shifts can cause performance degradation, and naive fine-tuning often fails to recover performance without sophisticated domain adaptation techniques or targeted data augmentation.

8. **Compatibility with Existing Pipelines:** The broader vision ecosystem—frameworks, pretrained backbones, and deployment optimizations—has been heavily optimized for CNN architectures over the past decade. Adapting these pipelines for transformer backbones often requires re-engineering of model serving stacks, feature extractors, and even hardware kernels.

6 Future Directions

The period 2015–2025 firmly established transformers as a cornerstone of modern computer vision. Looking ahead, several research trajectories are likely to define the "new era" of vision AI:

- **Unifying Architectures:** Convergence towards backbones that can handle diverse tasks—classification, detection, segmentation, and even video understanding—through task-conditioned heads or prompt-based interfaces. Architectures like OneFormer already demonstrate the feasibility of unifying semantic, instance, and panoptic segmentation.

- **Larger and Multimodal Foundation Models:** The emergence of billion-parameter vision-language models pretrained on web-scale multimodal corpora promises stronger generalization and zero-shot capabilities. Integrating audio, depth, or temporal modalities could yield holistic perception systems.

- **Improved Efficiency and New Attention Mechanisms:** Research into sub-quadratic attention (linear, sparse, low-rank), dynamic tokenization, and content-aware computation will be central to deploying transformers at the edge and in resource-constrained environments without sacrificing accuracy.

- **Better Theoretical Understanding:** Despite empirical success, a rigorous theoretical account of why transformers work so effectively for vision is still lacking. Insights into their inductive biases, optimization landscapes, and representation properties could inform more principled architecture design.

https://www.doupe.in | DOI: https://doi.org/10.71063/DJTTT.2025.1205
Creative Commons CC BY-ND 4.0 License

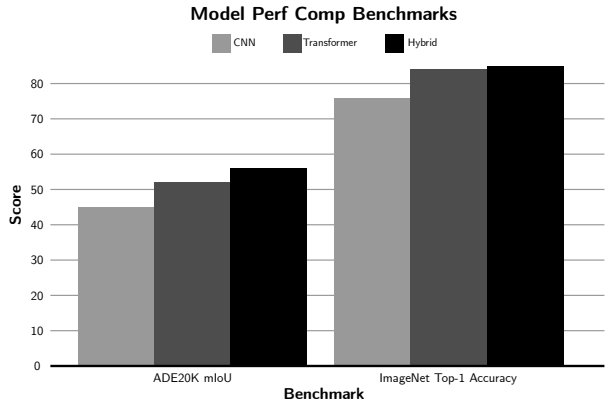

Figure 4: ADE20K mIoU and ImageNet Top-1 for CNN vs. Transformer vs. Hybrid models.

- **Robustness and Trustworthiness:** Future models must be robust to adversarial perturbations, distribution shifts, and corruptions, while being explainable and fair. This will require integrating robustness objectives into training and developing reliable interpretability tools beyond raw attention visualization.
- **Domain-Specific Models and Applications:** Transformers tailored for specialized applications—3D point cloud analysis, satellite imagery, medical imaging, and low-light vision—will likely proliferate, combining task-specific inductive biases with global modeling power.
- **Training Paradigm Shifts:** Self-supervised learning, particularly masked image modeling (e.g., MAE) and contrastive methods (e.g., CLIP, DINO), will continue to dominate as the preferred pretraining strategy, reducing dependence on large labeled datasets and enabling rapid adaptation to new tasks.

7 Conclusion

Over the last decade, transformer-based models have transitioned from an experimental adaptation of NLP architectures to a dominant paradigm in computer vision, particularly for image segmentation and classification. This survey has traced their evolution from the early incorporation of attention modules into convolutional pipelines to the development of pure and hybrid vision transformers capable of state-of-the-art performance across diverse benchmarks.

We have examined key architectural innovations—such as hierarchical attention, multi-scale tokenization, and hybrid convolutional embeddings—that have enabled transformers to address the limitations of plain ViT designs. Applications in semantic, instance, and panoptic segmentation, as well as in single-label, multi-label, fine-grained, and zero-shot classification, demonstrate the remarkable versatility of these models. Benchmark analyses show that, with sufficient pretraining, transformers consistently match or surpass CNN-based systems in accuracy, generalization, and cross-task transfer.

Nevertheless, significant challenges remain. Data and compute requirements, interpretability gaps, domain adaptation hurdles, and deployment inefficiencies still limit widespread adoption, especially in resource-constrained environments. The next wave of research will likely focus on efficient attention mechanisms, unified multitask architectures, and large multimodal foundation models trained with self-supervised objectives.

In closing, transformers have not merely augmented the vision research landscape—they have redefined it. As the field moves toward greater efficiency, robustness, and cross-domain integration, transformer-based architectures are poised to serve as the backbone of a new generation of vision AI systems that are more powerful, adaptable, and accessible than ever before.

References

[1] Xiangtai Li et al. "Transformer-based visual segmentation: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* (2024).

[2] Zihang Dai et al. "Coatnet: Marrying convolution and attention for all data sizes". In: *Advances in neural information processing systems* 34 (2021), pp. 3965–3977.

[3] Yaoli Wang et al. "Vision transformers for image classification: A comparative survey". In: *Technologies* 13.1 (2025), p. 32.

[4] Krishna Teja Chitty-Venkata et al. "Neural architecture search for transformers: A survey". In: *IEEE Access* 10 (2022), pp. 108374–108412.

[5] Hengshuang Zhao et al. "Psanet: Point-wise spatial attention network for scene parsing". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 267–283.

[6] Prajit Ramachandran et al. "Stand-alone self-attention in vision models". In: *Advances in neural information processing systems* 32 (2019).

[7] Han Hu et al. "Local relation networks for image recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3464–3473.

[8] Seyed Mohammad Hassan Erfani. "Developing a Vision-Based Framework for Measuring and Monitoring Water Resource Systems Using Computer Vision and Deep Learning Techniques". PhD thesis. University of South Carolina, 2023.

[9] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[10] Huiyu Wang et al. "Max-deeplab: End-to-end panoptic segmentation with mask transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5463–5474.

[11] Salman Khan et al. "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.

[12] Guy Dar et al. "Analyzing transformers in embedding space". In: *arXiv preprint arXiv:2209.02535* (2022).

[13] Zobeir Raisi et al. "2lspe: 2d learnable sinusoidal positional encoding using transformer for scene text recognition". In: *2021 18th Conference on Robots and Vision (CRV)*. IEEE. 2021, pp. 119–126.

[14] Hangbo Bao et al. "Beit: Bert pre-training of image transformers". In: *arXiv preprint arXiv:2106.08254* (2021).

[15] Xinxin Zhu et al. "Captioning transformer with stacked attention modules". In: *Applied Sciences* 8.5 (2018), p. 739.

[16] Feng Zhao et al. "Multi-head self-attention mechanism-based global feature learning model for ASD diagnosis". In: *Biomedical Signal Processing and Control* 91 (2024), p. 106090.

[17] Jie Liu et al. "Attention as relation: learning supervised multi-head self-attention for relation extraction". In: *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 2021, pp. 3787–3793.

[18] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. "On the computational complexity of self-attention". In: *International conference on algorithmic learning theory*. PMLR. 2023, pp. 597–619.

[19] Xuran Pan et al. "On the integration of self-attention and convolution". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 815–825.

[20] Sixiao Zheng et al. "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.

[21] Shehan Perera, Pouyan Navard, and Alper Yilmaz. "Segformer3d: an efficient transformer for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 4981–4988.

[22] Enze Xie et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[23] Bowen Cheng et al. "Masked-attention mask transformer for universal image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299.

[24] Shuyi Yin. "Automated Road Environment Perception for Safety Assessment With Tuning-Free Vision Foundation Models". PhD thesis. University of Washington, 2025.

[25] Junkang An and Inwhee Joe. "Attention map-guided visual explanations for deep neural networks". In: *Applied Sciences* 12.8 (2022), p. 3846.

[26] Ali Hassani et al. "Escaping the big data paradigm with compact transformers". In: *arXiv preprint arXiv:2104.05704* (2021).