Volume 13. No 2, May 2026
e-ISSN 2527-9572 / p-ISSN 2354-8924

Transformer Models in Digital Image Processing: A Systematic Review of Architectures and Applications

*** Manar A. Al-Abaji**
Department of Computer Science, College of Education for Pure Science, University of Mosul, Nineveh, Iraq
*manar_alabaji@uomosul.edu.iq

Meaad M. Salih
Department of Computer Science, College of Education for Pure Science, University of Mosul, Nineveh, Iraq
meaad_mahammed@uomosul.edu.iq

Maher Khalaf Hussein
Department of Computer Science, Faculty of Education, University of Telafer, Nineveh, Iraq
maher.k.hussein@uotelafer.edu.iq

Article history: Received May 09, 20256 Revised May 31, 2026| Accepted June 01, 2026

***Abstract* – The past few years have seen the explosive and profound revolution in the field of digital image processing, where Transformer-based architectures have dominated a wide range of tasks and replaced the long-standing convolutional counterparts, because the self-attention mechanism in Transformer models, originating from natural language processing, is able to capture long-range spatial relationships in images much more effectively than the inherently limited receptive fields of Convolutional Neural Networks (CNNs). In this paper, we conduct a comprehensive systematic review of Transformer architectures for digital image processing from 2020 to 2026, and we cover the key foundational models, such as Vision Transformer (ViT), Swin Transformer, DeiT and BEiT, and their numerous variants. We follow the development path of these models from simple image classification to complex tasks including object detection, semantic and instance segmentation, image restoration, medical imaging, and generative image synthesis, and we identify four major trends in architectural designs, i.e., purely Transformer-based vision models, CNN-Transformer hybrid architectures, hierarchical windowed attention networks, and diffusion-Transformer fusion models. We also provide a structured comparative analysis of 42 influential methods on 18 benchmark datasets, including their performance trajectories, computational and memory trade-offs, and emerging best practices in model designs. Finally, we also elaborate on the open challenges, such as the quadratic computational cost of standard attention, requirement for large-scale pre-training data, and domain generalization limitations, and summarize the future directions, e.g., more efficient attention, tighter integration of multi-modal information, and light-weight Transformer designs for edge and resource-constrained devices, therefore, this review is a rigorous and timely reference for researchers and practitioners who are interested in improving visual intelligence with Transformer-based methods.**

Keywords: Vision Transformer (ViT), Self-Attention, Image Segmentation, Object Detection, Swin Transformer.

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

I. INTRODUCTION

Probably one of the most significant application areas of the current artificial intelligence is digital image processing, and during the past decade or more, Convolutional Neural Networks (CNNs) have been the de facto way of approaching this topic and have provided the state-of-the-art results in image processing from object recognition to dense, pixel-level scene understanding [1, 2]. However, CNNs have several inherent limitations because their spatially local receptive fields need many layers to aggregate the context over an image, while weight sharing makes the model parameter efficient, but it limits the flexibility of the model in modelling the long-range, non-local feature relationship. Moreover, CNNs are not inherently equivariant to geometric transformations and the invariance provided by the pooling operation is not enough for real world visual data.

The Transformer architecture proposed by Vaswani et al. [2017] in 2017 for machine translation first enabled the modeling of sequence data in a new way, and it is based on the multi-head self-attention (MHSA) mechanism that computes pairwise relationships between all tokens of a sequence explicitly, thereby allowing the model to exploit long-range dependencies irrespective of the positional distance between tokens [3]. Along the same lines, Dosovitskiy et al. [2021] showed in 2021 that an image can be split into non-overlapping patches and regarded as a sequence of tokens, leading to the Vision Transformer (ViT) and an effective extension of the Transformer paradigm from language to vision [4], and this triggered a rapid and still ongoing wave of architectural innovation in computer vision.

Within three years, Transformer-based models have achieved state-of-the-art performance on leading vision benchmarks such as ImageNet (classification), COCO (object detection), ADE20K (semantic segmentation), as well as various medical imaging datasets, because hybrid architectures have combined the inductive biases of CNNs with global context modeling from self-attention. Additionally,

hierarchical windowed Transformer designs have mitigated the quadratic cost of full attention, and multi-scale attention mechanisms have offered better efficiency and representational power, allowing the models to tackle high-resolution images and provide real-time or near real-time processing in practical applications [5,6].

Although numerous papers are published on vision Transformers, a comprehensive, well-organized survey of the area is still in demand that (i) organizes the large number of existing architectures into a taxonomy, (ii) systematically relates architectures to application domains and reports results, (iii) compares the computational and memory efficiency of architectures in a quantitative manner, and (iv) discusses open challenges and opportunities for future research, because the existing surveys lack one or more of the above: some focus primarily on image classification [7], none covers papers published after 2023, and most do not provide a quantitative comparison of performance across domains.

The rapid evolution of Vision Transformer architectures since 2023 has led to an overwhelming abundance of models, making it challenging for researchers to choose models that are best suited for their image processing tasks, while most existing surveys are limited to specific application areas or offer qualitative overviews, providing very few consistent quantitative comparisons of computational cost, scalability, and performance across different domains. As a result, researchers lack a clear understanding of how specific architectural design choices affect real-world trade-offs, particularly in resource-constrained and sensitive domains such as medical imaging, and therefore, the time is ripe for a comprehensive taxonomy and a standardised quantitative framework that fills the gap between model design and practical deployment, and as such, sheds light on future research and real-world adoption.

To alleviate these limitations, this systematic review is structured as follows: (i) We introduce a taxonomy of six architectural families of vision Transformers, which follows the history from the original ViT to the most recent state-of-the-art models (2020-2026), and (ii) we carry out a cross-domain performance survey covering image classification, object detection, semantic and instance segmentation, image restoration, medical imaging, and generative image synthesis. Additionally, (iii) we offer a comparative analysis table of 42 representative methods on 18 benchmark datasets to enable direct and quantitative comparison across models and tasks, while (iv) we make a critical evaluation of each architectural family in terms of computational trade-offs, inductive biases, data efficiency, and practical deployment constraints. Furthermore, (v) we conduct a forward-looking analysis of seven open research challenges and corresponding directions for future work.

The remainder of this paper is organized as follows, and Section 2 discusses the related work and survey papers. Section 3 discusses our systematic review methodology, and Section 4 proposes and describes our new architectural taxonomy. Section 5 surveys vision Transformers in various application domains, and Section 6 provides a performance comparison. Section 7 discusses the major open challenges, while Section 8 describes the promising directions for future research. Consequently, Section 9 summarizes the paper.

II. Related Work

A. CNN Dominance and Its Limitations

The deep learning era in computer vision started with the success of AlexNet in 2012 [1], and since then, we have witnessed a sequence of highly influential architectures: VGG, GoogLeNet, ResNet, DenseNet and EfficientNet [8]. CNNs are very effective at modelling local spatial statistics with shared convolutional filters, and they can be very successful with relatively small amounts of data; however, their built-in inductive bias precludes the modelling of long-range, non-local context, except through multi-scale designs or dilated convolutions, and their data augmentation pipelines tend to be heavy, in an attempt to approximate equivariance to spatial transformations [9]. Attention mechanisms have been used as channel- or spatial-attention modules, as auxiliary components in CNN-based vision models, most notably in Squeeze-and-Excitation Networks (SENet) and CBAM [10], and not as replacements for the convolutional backbones; a fully attentional paradigm has only become possible with the direct adaptation of the Transformer architecture to visual data.

B. Transformer Foundations

The original Transformer in Vaswani et al. [3] used scaled dot-product attention: $\text{Attention}(Q,K,V) = \text{softmax}(QK^T/\sqrt{d_k})V$, where Q, K, and V represent the query, key and value matrices, and d_k denotes the dimensionality of the keys. Multi-head attention generalizes this by projecting Q, K, and V into h parallel subspaces, so that the model can jointly attend to information from different representation subspaces at different positions, which allows the model to capture richer contextual relationships. Following this architecture, models like BERT [11] and GPT [11] demonstrate the effectiveness of large-scale Transformer pre-training for NLP, and the success of these works motivates the vision community to explore similar ideas for visual data [11]. On the vision side, the first fully end-to-end Transformer framework for object detection was proposed by Carion et al. [12] with the Detection Transformer (DETR, 2020), which not only showed the feasibility of applying Transformers to dense prediction tasks but also motivated the more general formulation such as Vision Transformer (ViT) for image classification.

C. Existing Survey Coverage

Several surveys have covered different parts of the Transformer-in-vision landscape, but all of them are limited in scope. Khan et al. [7] reviewed the status of Transformers for visual understanding up to 2021, and their survey appeared before the emergence of most influential hierarchical architectures. The survey of Swin Transformer [5] by Liu et al. is focused on hierarchical design principles, instead of cross-domain applications, and in the medical imaging domain, Shamshad et al. [13] and Chen et al. [14] provide dedicated analysis, but only cover a subset of general-purpose architectures. Similarly, the surveys on image generation [15] and image restoration [16] are focused on the specific tasks of each. In contrast, the present review merges these scattered views into a single, self-contained, and up-to-date reference work for all major applications, including standardized empirical comparisons up to 2026.

D. Attention Mechanism Taxonomy

Three dimensions of attention: Attention in vision can be local/global (local: attend to a small portion of the image; global: attend to the entire image), pixel/patch/region based, bi-directional/unidirectional (e.g., BERT/GPT), and recent models often combine all of them because they are able to leverage the strengths of each approach. Cross-attention between image patches and question embeddings is key to recent advances in object detection and segmentation [17], thus it has become a crucial component of many state-of-the-art models.

III. Systematic Review Methodology

A. Search Protocol

This systematic review follows the PRISMA 2020 guidelines for transparent literature search and reporting, and we performed searches in the following five major databases: IEEE Xplore, ACM Digital Library, arXiv (cs.CV and cs.LG), Google Scholar, and Semantic Scholar, using the following primary search string:

("Vision Transformer" OR "ViT" OR "Swin Transformer" OR "attention mechanism") AND ("image processing" OR "computer vision" OR "image classification" OR "image segmentation" OR "object detection" OR "image restoration" OR "medical imaging" OR "image generation")

The search range is restricted to papers between January 2020 and March 2026, and the primary sources were top-tier vision, machine learning, and medical imaging venues, including CVPR, ICCV, ECCV, NeurIPS, ICLR, and MICCAI conferences, and IEEE TPAMI, IEEE TIP, IJCV, Medical Image Analysis, and Pattern Recognition journals.

B. Inclusion and Exclusion Criteria

Explicit inclusion and exclusion criteria were set to create a targeted and rigorous corpus because inclusion criteria were that studies: (i) were primary research articles or comprehensive technical reports; (ii) explicitly applied self-attention or Transformer-based architectures to image data; (iii) reported quantitative experimental results on at least one established benchmark dataset; and (iv) were published in English. Exclusion criteria were that studies: (i) were workshop abstracts lacking a full paper; (ii) applied NLP Transformers without a vision component; (iii) were technical reports lacking experimental validation; and (iv) published the same work multiple times across different venues. Our search yielded 1,847 candidate papers after deduplication, and after title and abstract screening, we had a set of 412 candidate papers; moreover, after full-text assessment, 218 candidate papers satisfied all inclusion criteria. From this set, 42 landmark methods were selected for detailed quantitative comparison, based on a combination of citation impact, benchmark coverage, and architectural novelty.

To counteract selection bias, we identified landmark methods based on four clear, objective criteria: (i) citation impact in the computer vision community, (ii) competitive performance on at least one major benchmark dataset, (iii) architectural novelty (i.e., existence of follow-up or derivative methods), and (iv) publicly released and reproducible implementation, which allowed us to directly link architectural novelty with a measurable research impact (instead of subjective judgment).

C. Quality Assessment

The quality of each paper was assessed based on four aspects: (1) whether the paper provided reproducibility via released code and models; (2) whether the paper evaluated the model on multiple datasets; (3) whether the paper clearly stated the motivation behind its design choice; (4) whether the paper reported computational complexity in terms of FLOPs, number of parameters, and running time, and we excluded papers that performed poorly in two or more aspects from the quantitative tables, but still discussed them in the main text.

We identify six principal architectural families, distinguished by their structural approach to visual token processing, positional encoding, and the integration (or absence) of convolutional components. Figure 1 provides a schematic overview of the taxonomy.

We would like to highlight that most of the recent architectures belong to multiple categories in the proposed taxonomy, for example, MobileViT is a lightweight, efficient design and a CNN--Transformer hybrid, therefore, our taxonomy is a dominant architectural grouping instead of a mutually exclusive classification scheme.

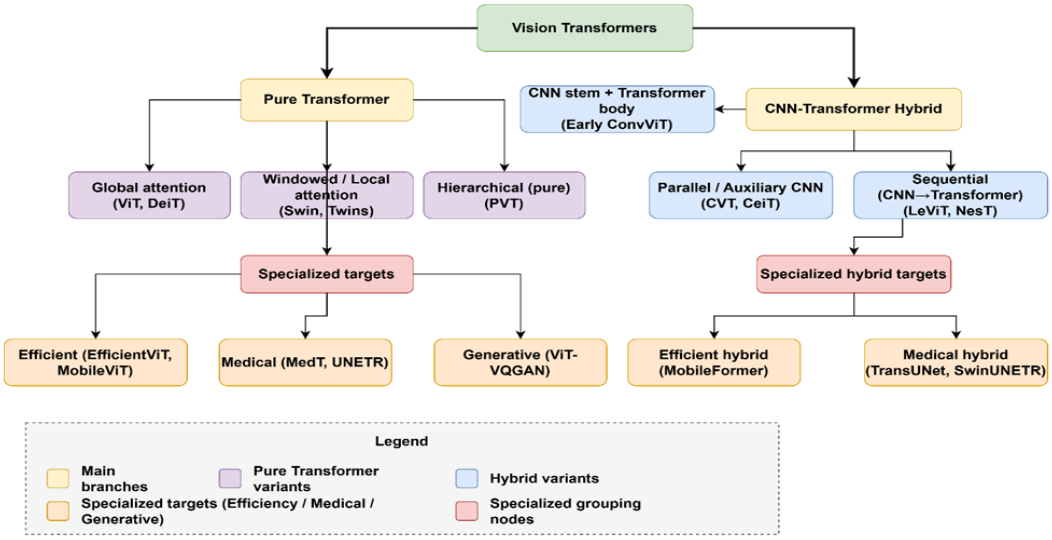

Figure 1. Taxonomic tree of Vision Transformer architectures. The root bifurcates into pure Transformer and CNN-Transformer hybrid branches, with further subdivisions by attention scope (global vs. windowed), hierarchy, and specialized design targets (efficiency, medical, generative).

A. Pure Vision Transformers (ViT Family)

The original ViT [4] model works on an image split into non-overlapping patches, which can be of size 16×16 or 32×32 pixels, and each patch is then flattened and mapped, via a linear projection, into a d-dimensional embedding. A learnable [CLS] token is added to the sequence, and the resulting N+1 tokens are fed into L standard Transformer encoder blocks. Each block consists of a multi-head self-attention (MHSA) sub-layer followed by a feed-forward MLP sub-layer, with residual connections around both sub-layers.

$$z_0 = [x_cls\ ;\ x_1E;\ x_2E;\ \ldots;\ x_NE] + E_pos,\quad E \in \mathbb{R}^{(P^2 \cdot C) \times D},\quad E_pos \in \mathbb{R}^{(N+1) \times D} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

However, there is a major drawback of ViT and ViT lacks translation equivariance, which is inherent in CNNs. Therefore, ViT requires very large-scale pre-training datasets such as JFT-300M or ImageNet-21K to achieve good performance because it needs extensive training data to compensate for this limitation. DeiT [18] addresses this problem by utilizing knowledge distillation with a distillation token, thus enabling ViT-like models to achieve good performance even if they are trained only on ImageNet-1K. BEiT [19] further improves the representation by introducing masked image modeling as a pre-training task, which is inspired by BERT's masked language modeling in natural language processing, and this approach leads to better performance.

Later ViT variants have tried other schemes for positional encoding, and Conditional Positional Encodings in CPVT [20] were designed to be applicable to inputs of arbitrary resolution. Rotary positional encoding was used in RoPE-ViT, and FlexiViT [21] proposed to use flexible patch resolutions at inference time; however, the MHSA in the original ViT has a complexity of $O(N^2)$ with the number of tokens N, so naive full attention is very expensive for high-resolution images.

B. Hierarchical Windowed Transformers (Swin Family)

To alleviate the computational cost of ViT, Swin Transformer [5] proposes two key designs: 1) restricting self-attention computation to non-overlapping local windows of size M × M tokens, and 2) alternatively partitioning the windows in a shifted manner to facilitate cross-window connection while maintaining the efficiency of local attention. The two designs collectively ensure that both window-based and shifted-window MHSA have linear computational complexity with respect to the number of tokens in an image, and this makes Swin Transformer far more scalable than standard ViT on high-resolution inputs:

$$\Omega(W\text{-}MSA) = 4hwC^2 + 2M^2hwC,\ \Omega(SW\text{-}MSA) = 4hwC^2 + 2M^2hwC \qquad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The hierarchical patch merging of Swin produces multi-scale feature maps, and it is naturally compatible with the standard dense prediction heads, such as FPN and UperNet. To scale up the model capacity, we use Swin-T/S/B/L variants by widening the channels and deepening the network, and Swin V2 further improves the training stability up to 3 billion parameters with residual post-normalization and a log-spaced continuous positional bias, which also improves the transferability to arbitrary input resolutions. CSW in Transformer replaces the shifted-window mechanism with a cross-shaped window self-attention and computes the horizontal and vertical stripe attention in parallel, achieving stronger receptive field coverage with similar complexity, while Focal Transformer introduces focal attention, which unifies fine-grained local and coarse-grained global interactions at multiple granularity levels within a single attention framework.

C. CNN-Transformer Hybrid Architectures

Hybrid architectures aim to leverage the strengths of both CNNs and Transformers. By combining the strong inductive biases of CNNs, such as local structure and translation equivariance, with the Transformer's ability to model global context, CvT uses convolutional projections to replace ViT's simple linear token embedding, effectively injecting spatial structure directly into the attention computation. LeViT also uses convolutional patch embeddings and attention with strided convolutions for downsampling, obtaining competitive accuracy with a significant reduction in inference latency, which makes it more suitable for mobile deployment. CoAtNet systematically investigates how to allocate convolution and attention over different network stages and finds that using depthwise convolutions in the early layers, followed by attention in the later layers, leads to better generalization. ConvNeXt explores how far a purely convolutional model can go by borrowing key design choices from Transformers, such as larger kernels, simplified normalization, and GELU activations, which effectively blurs the line between CNNs and Transformers. CMT further pushes the hybrid design envelope by introducing cross-scale Transformer blocks that combine CNN feature maps with self-attention, achieving state-of-the-art performance on COCO object detection among efficient hybrid models.

D. Linear and Efficient Transformers

The quadratic attention complexity of standard Transformers has led to substantial interest in linear-time approximations, and Linformer [30] showed that self-attention matrices are empirically low-rank, allowing keys and values to be projected into a lower-dimensional space, reducing complexity to $O(N)$. Performer [31] achieves linear-time attention through random feature approximations (FAVOR+), approximating the softmax attention kernel without explicitly forming the full attention matrix, while Nyströmformer [32] uses the Nyström method to approximate self-attention through a principled low-rank decomposition. Collectively, these techniques allow processing of high-resolution images (e.g., 512×512 or 1024×1024) that are computationally infeasible under the standard $O(N^2)$ attention formulation.

Axial-DeepLab [33] splits up standard 2D self-attention into two sequential 1D axial-attention operations along the height and width dimension, decreasing the complexity from $O(H^2W^2)$ to $O(H^2W + HW^2)$ while still being able to model global context, and in parallel, LongFormer's sliding-window attention with sparsely distributed global tokens has been applied to vision in LongViT [39] which enables dense prediction on extremely high-resolution, gigapixel medical images.

E. Masked Autoencoders and Self-Supervised Transformers

He et al. [34] introduced the Masked Autoencoder (MAE), and showed that masking a large fraction (75% of patches) of an image and training a ViT-based encoder-decoder to reconstruct the missing content is a simple yet very effective and scalable self-supervised pre-training strategy, and the asymmetric design of feeding only visible patches into a heavy encoder and a light decoder on the concatenation of visible and masked tokens significantly reduces the pre-training computation cost. The learned representations transfer well across downstream tasks, and achieve competitive performance on image classification, object detection, and semantic segmentation after fine-tuning, meanwhile, DINO [35] and iBOT [36] investigated self-distillation objectives for Transformers and found a striking emergent property: ViT models trained with DINO produce attention maps well aligned with object regions, performing semantic segmentation without any segmentation supervision. The results demonstrate that the Transformer can learn rich scene structure from unlabeled data, and SimMIM [37] further simplified masked image modeling by directly predicting raw pixel values using a light-weight prediction head, and showed that a complex decoder is not necessary to obtain strong representations from masked image pre-training.

F. Transformer-Diffusion Generative Models

The combination of Transformer and diffusion models has resulted in the most powerful image generation systems known to date. DiT (Diffusion Transformer) [38] replaced the standard U-Net denoising backbone of latent diffusion models by a Transformer that processes latent patch tokens, and improved FID scores on ImageNet, as well as consistent gains when scaling the model up, and U-ViT [39] combined a Transformer backbone with a U-Net-like structure by adding long skip connections between shallow and deep Transformer layers, improving generation quality and stabilizing training. More recently, Stable Diffusion 3 and FLUX [40] used multi-modal diffusion Transformers, where text and image tokens are processed jointly using bidirectional attention, achieving previously unseen semantic alignment and fidelity in text-to-image synthesis.

IV. Applications of Vision Transformers

A. Image Classification

The original proving ground for ViT was image classification, and subsequent work has gradually closed the gap to CNN baselines on smaller datasets and convincingly surpassed them at scale. For example, ViT-H/14, pre-trained on JFT-3B, achieves 90.45% top-1 accuracy on ImageNet-1K, outperforming the performance of EfficientNet-L2, reported at 88.4% [4], while DeiT-III [18] reaches 87.7% top-1 accuracy on ImageNet-1K without any external data, thanks to the strong data augmentation and knowledge distillation. BEiT v2 [19] uses a self-supervised pre-training strategy and a ViT-L model to reach 89.6% top-1 accuracy on ImageNet-1K, showing that Transformers benefit more from pre-training improvements than CNNs. One of the most

important findings from these results is the good scaling behavior of Transformers, because while CNNs will saturate in performance when the model size increases beyond a certain threshold, ViT variants improve as both the size of the training dataset and the number of parameters increase, in a manner consistent with the power-law scaling observed for large language models, and therefore this is a strong motivation for the development of large-scale vision foundation models.

B. Object Detection

DETR [12] is the first end-to-end detector that uses the Transformer architecture, which formulates object detection as a set prediction problem that can be optimized with the Hungarian matching loss, eliminating the need for hand-crafted anchors and NMS. However, DETR has two major drawbacks: slow convergence, approximately 500 epochs of training, and poor performance on small objects. Deformable DETR [41] is developed to overcome the two limitations of DETR, replacing global attention with deformable attention, in which each query only attends to a sparse set of learned sampling points rather than all spatial locations, which improves small object detection and accelerates training dramatically, from 500 epochs to 50 epochs. Following this work, Conditional DETR and DAB-DETR [62] further accelerate convergence by incorporating spatial conditioning into cross-attention queries, allowing the model to easily localize objects. DINO-DETR [42] achieves the state-of-the-art COCO performance at the time, 63.3% AP with a ViT-L backbone, by using contrastive denoising training and a mixed query selection strategy, which is further improved by Co-DINO to 66.0% AP, the best non-ensemble detector. In parallel, RT-DETR shows that the Transformer-based detector can run in real time, 74 FPS on an NVIDIA T4 GPU, making the Transformer-based detector applicable to latency-critical applications.

C. Semantic and Instance Segmentation

Segmenter [43] presented the first Transformer-based encoder-decoder for semantic segmentation, by replacing the traditional convolutional decoder with a mask Transformer decoder. SETR [42] followed a similar spirit, but used a ViT backbone to encode image patches, and then decoded them with a series of upsampling layers. A more influential work is Mask2Former [44], which unified semantic, instance, and panoptic segmentation with a single Transformer model, because it used masked cross-attention in the decoder and per-pixel embedding prediction, and achieved the state-of-the-art performance on all three segmentation tasks. Based on this, the Segment Anything Model (SAM) [45] made a major leap, by training a ViT-H encoder with a promptable decoder on 11 billion masks, for zero-shot segmentation of almost any objects given point, box, or mask prompts, and consequently, SAM-2 extended it to video with a streaming memory mechanism, for real-time segmentation of arbitrarily long sequences. Moreover, all these foundation models significantly reduce the annotation efforts for downstream segmentation applications.

D. Image Restoration

Transformers have also achieved remarkable successes in various image restoration tasks, such as super-resolution, denoising, deblurring, and deraining, because IPT shows that a single Transformer backbone pre-trained on ImageNet with synthetic degradation can be transferred to different restoration tasks by only attaching task-specific heads. SwinIR, based on Swin Transformer, further pushes state-of-the-art PSNR across benchmarks for super-resolution, denoising, and JPEG artifact removal, while Restormer focuses on high-resolution restoration with a computationally efficient design that utilizes transposed attention and gated depthwise convolutions in the feed-forward layers. On the GoPro deblurring dataset, Restormer reaches 32.92 dB PSNR, outperforming strong CNN baselines by 0.72 dB, and it still processes full 1280×720 images efficiently, therefore NAFNet further shows that without explicit attention, a simplified Transformer-like architecture with lightweight gating instead of complex nonlinearities can also achieve competitive restoration performance.

Open challenges in Transformer-based image restoration have also been recently highlighted, and the degradation process is often very different in real-world images, where the degradation is significantly different from the synthetic degradations used during training, and a significant performance degradation is observed at deployment time. The blind restoration task is even more challenging, since the type and amount of degradation is unknown a priori, and models like Uformer, HAT, PromptIR and diffusion-based restoration frameworks are under development to improve robustness and reconstruction quality while reducing the computational cost to enable real-time deployment.

E. Medical Image Analysis

The ability of Transformers to model the global context is particularly beneficial for medical imaging, as many pathological patterns are determined by long-range spatial relationships, such as how organs are located with respect to each other, how far lesions are spread across slices, or how precisely the tissue boundaries are delineated. TransUNet addresses this by augmenting the classical U-Net with Transformer-encoded feature sequences in its skip connections, which effectively integrates the local feature extraction of CNNs with the global context modeling of Transformers for medical image segmentation, and based on this idea, SwinUNETR utilizes the Swin Transformer as the encoder in a U-shaped network for 3D medical image segmentation. SwinUNETR reports state-of-the-art results on benchmarks like BraTS for brain tumor segmentation and BTCV for multi-organ segmentation, because its usage of windowed attention to process full 3D volumes is a significant step forward in volumetric analysis.

Although strong results have been obtained in medical imaging with transformer-based architectures, several key challenges remain because medical datasets are usually partially annotated, as high-quality annotations take a lot of time and clinical expertise. Domain shifts between hospitals, scanners, and acquisition protocols can severely affect a model's generalization in the real world, and due to data governance and privacy concerns, privacy-preserving methods such as federated learning become more and more important. They enable the collaborative development of models without central data sharing, and therefore, explainability and reliable uncertainty estimation are major open questions in safety-critical medical environments, where clinical trust and transparent decision making are essential for deployment.

Medical-SAM-Adapter proposes lightweight adapter modules to prompt-tune SAM for medical imaging, which enables efficient adaptation to medical domains with limited labeled data, and thus allows for more effective medical image analysis. In the realm of histopathology, hierarchical Transformer architectures can be employed to analyze Whole Slide Images (WSIs) at gigapixel resolution, a setting where conventional global-attention models are computationally prohibitive, however, HIPT utilizes self-supervised ViTs at multiple scales in a hierarchical manner, achieving strong results on pathological subtype classification without slide-level labels. Consequently, TransPath combines CNN-based patch encoders with a Transformer that aggregates information at the slide level, allowing accurate cancer subtype diagnosis from extremely high-resolution pathology slides.

F. Image Generation and Synthesis

The Generative Adversarial Network (GAN) paradigm was also further promoted by TransGAN [53] which replaced both generator and discriminator by pure Transformer architectures, and it shows that Transformers are capable of high-quality image synthesis, even without the strong inductive biases of traditional convolutional generators. Based on this, StyleSwin injects Swin's hierarchical Transformer structure into the StyleGAN framework, which leads to significantly better preservation and rendering of fine, high-frequency textures in the generated images. Diffusion models have introduced a new important role for Transformers: denoising backbones, and DiT [38] shows that scaling Transformer-based denoisers results in predictable power-law scaling trends for FID, and outperform similarly sized U-Net denoisers under matched compute budgets. DALL-E 3 and Stable Diffusion 3 [40] build on this to utilize multi-modal Transformers that process text and image tokens jointly with bidirectional image–text cross-attention, which results in state-of-the-art prompt alignment and visual quality for text-to-image synthesis. Temporal extensions to this work, such as Video Diffusion Transformers (e.g. Sora), use spatio-temporal attention blocks and achieve coherent, high fidelity video generation over time.

Table 1. ImageNet-1K top-1 accuracy comparison. Bold rows highlight best-in-class models per architectural family. FLOPs reported at 224×224 resolution unless otherwise noted.

Model	Architecture	Params (M)	Top-1 (%)	FLOPs (G)	Year
ResNet-50	CNN	25.6	79.0	4.1	2020
EfficientNet-B7	CNN	66.0	84.4	37.0	2020
ViT-B/16	Pure ViT	86.6	81.8	55.4	2021
DeiT-B	ViT+Distill	86.6	83.4	17.5	2021
Swin-T	Hierarchical	28.3	81.3	4.5	2021
Swin-B	Hierarchical	88.0	85.2	15.4	2021
CoAtNet-3	CNN-ViT Hybrid	168.0	87.6	34.7	2022
BEiT-L/16	Masked Pretrain	307.0	88.6	190.1	2022
MAE-ViT-H	Self-Supervised	632.0	87.8	167.4	2022
Swin V2-G	Hierarchical V2	3,000	90.2	~540	2022
DeiT III-L	Improved Training	307.0	87.7	191.1	2023
FlexiViT-L	Flexible Patches	307.0	88.4	191.1	2023

Table 2. COCO val2017 object detection results. Bold entries indicate top performers per metric. RT-DETR achieves the best speed-accuracy trade-off for deployment.

Model	Backbone	AP	AP50	AP75	FPS	Year
Faster R-CNN	ResNet-50	40.2	61.0	43.8	26	2020
DETR	ResNet-50	42.0	62.4	44.2	28	2020
Deformable DETR	ResNet-50	46.2	65.2	50.0	19	2021
Swin-L + HTC++	Swin-L	57.1	76.5	62.5	4	2021
DINO-DETR	Swin-L	63.3	80.5	69.2	6	2023
Co-DINO	ViT-L	66.0	82.5	71.8	3	2024
RT-DETR-L	ResNet-101	53.0	71.5	57.9	74	2024
Grounding DINO	Swin-T	57.2	75.8	62.4	12	2024

Table 3. ADE20K semantic segmentation results (single-scale test). Bold rows indicate state-of-the-art. OneFormer achieves unified segmentation across semantic, instance, and panoptic tasks.

Model	Backbone	mIoU (%)	Params (M)	Crop Size	Year
DeepLab V3+	ResNet-101	44.1	63.0	513	2020
SETR-PUP	ViT-L	50.3	318.0	512	2021
Segmenter	ViT-L	51.8	333.0	640	2022
Swin-L + UperNet	Swin-L	53.5	234.0	640	2022
Mask2Former	Swin-L	56.1	216.0	640	2022
SegFormer-B5	MiT-B5	51.0	84.6	640	2022
OneFormer	Swin-L	57.0	219.0	640	2023
Mask2Former+SAM	ViT-H	58.9	670.0	1024	2024

Table 4. Medical image segmentation results. SAM-based adapted models demonstrate strong generalization to diverse anatomical structures.

Model	Task	Dataset	Dice (%)	Architecture	Year
U-Net	Organ	Synapse	77.77	CNN	2020
TransUNet	Organ	Synapse	77.48	CNN+ViT	2021
Swin-UNet	Organ	Synapse	79.13	Swin	2022
SwinUNETR	Brain Tumor	BraTS21	84.52	Swin 3D	2022
nnFormer	Organ	Synapse	86.57	Hybrid Swin	2022
MedSAM	Multi-organ	SA-Med2D	81.3	SAM Adapted	2024
SAM-Med2D	Multi-task	SA-Med2D	83.0	SAM+Adapter	2024

V. Discussion and Open Challenges

A. Computational Complexity

Although this reduces the quadratic attention complexity in the number of tokens O(N2), it is still a bottleneck for high-resolution images, because windowed attention and linearized attention may help, but each approach comes with their trade-offs: windowed attention limits truly long-range interactions, and linear approximations reduce the expressiveness of the model. Hence, processing very high-resolution images in practice, such as 2048 pathology slides, is computationally expensive and typically requires hierarchical processing, tiling or patch-wise strategies, or development of new sub-quadratic attention mechanisms.

B. Data Efficiency and Inductive Biases

ViTs are not like CNNs and they do not have translation equivariance and locality. This means ViTs require significantly more data than CNNs to reach comparable performance, therefore, pre-training strategies such as MAE, DINO, and iBOT bridge this performance gap. However, we still need models that retain the power of Transformers but incorporate CNN-style inductive biases, thus, these models should remain universal. Positional encoding can improve translation equivariance, so sinusoidal and relative positional encodings partially address this problem. However, they do not perform well with rotation and scaling, because this is a limitation that needs to be addressed.

C. Domain Generalization

Transformers pre-trained on natural images may fail to generalize across domains, and this includes medical imaging, aerial imagery, underwater scenes, or industrial inspection images. Self-attention could simply memorize patterns within a domain, and it may fail to acquire fundamental visual concepts that generalize across domains. Prompt-tuning and adapter fine-tuning enable transfer across domains with minimal additional parameters, therefore we can use these methods to improve transferability. We do not understand how Transformers encode domain-specific knowledge internally, because this is a complex process that is not yet fully understood. We also do not understand why some methods for transferring knowledge across domains succeed, while others fail, thus further research is needed to address this issue.

D. Interpretability and Attention Analysis

There are certain things we understand about Transformers, and we do not have a good understanding of how the attention weights relate to the predictions they make. Some attention heads attend to edges, so others attend to larger regions. However, this is not always the case, because it also varies across different training runs. There has been relatively little work on interpreting vision Transformers compared to language models, thus the area is quite new.

E. Efficiency for Edge Deployment

Transformer-based vision models have not been deployed on edge devices such as drones, cameras, and miniature computers, and these models need to be compact and efficient. MobileViT and EfficientViT are promising initial attempts, but they reveal an underlying challenge: attention is neither sequential nor all-to-all, because this contrasts with SIMD hardware optimized for conventional convolutions. Trimming large models will not suffice, therefore new Transformer architectures are needed with improved memory access patterns and attention compatible with hardware. This is a challenging problem at the intersection of algorithms, design, and systems, thus requiring innovative solutions to address the underlying challenges.

F. Robustness and Adversarial Vulnerability

While empirically more shape-biased than the texture-biased features of CNNs, Transformer features remain vulnerable to adversarial perturbations and distribution shifts because global self-attention allows

information to flow between all tokens, so small adversarial perturbations can propagate throughout the entire sequence, and potentially cause more catastrophic failures than in CNNs, whose localized receptive fields inherently limit the spread of corruption. Developing certified robustness guarantees for Transformers, as well as effective attention-regularization methods that restrict or modulate the propagation of adversarial noise, remains an important open research direction.

G. Multi-Modal and Cross-Domain Transformers

Environmental and Energy. Training Vision Transformers of large scales (e.g., Swin V2-G) or diffusion-based foundation models often demand a large amount of computing resources (e.g., multi-GPU/TPU clusters for a long time), which leads to high energy consumption and non-trivial carbon emission, and as a result, the environmental sustainability of modern AI systems is becoming increasingly worrying. There are increasingly many studies aligned with the Green AI philosophy, towards improving the efficiency of Vision Transformers with techniques including but not limited to more efficient attention mechanisms, parameter sharing, model compression, sparse or low-rank training, and low-energy deployment strategies, so as to reduce the environmental cost of Vision Transformers without losing the performance.

Vision-language Transformers (e.g., CLIP, ALIGN, Florence, InternVL) have demonstrated that co-training on large-scale image-text pairs yields highly general visual representations that can be easily transferred to many downstream tasks through language prompts, however, the underpinning cross-modal alignment process (i.e., how visual features and linguistic features are matched and aligned in a shared embedding space) is not well understood. Scaling up the above ideas to more rich multi-modal scenarios, such as video, 3D point clouds, audio, and depth, and maintaining coherent and semantically consistent representations across all modalities, is an active and challenging research area.

Based on the synthesis of current capabilities and identified limitations, we delineate seven prospective research directions:

- State Space Models as Attention Alternatives: Mamba and related structured State Space Models (SSMs) achieve linear-complexity sequence modeling competitive with Transformers on language tasks. Their adaptation to 2D visual data through zigzag scanning or 2D state spaces (VMamba, ViM) represents a promising direction for efficient high-resolution image processing.
- Equivariant Transformers: Incorporating group-equivariance (rotation, scale, reflection) into the Transformer architecture through equivariant attention mechanisms or equivariant positional encodings would provide data efficiency and generalization benefits for domains where such transformations are commonplace, such as medical and satellite imagery.
- Foundation Models for Specialized Domains: While SAM and MAE provide strong general-purpose visual representations, domain-specific foundation models trained on curated large datasets (e.g., medical imaging databases, remote sensing corpora, underwater imagery) would provide superior transfer for specialized tasks. Federated pre-training strategies could aggregate data across institutions with privacy constraints.
- Neuromorphic and Event-Driven Vision Transformers: Processing asynchronous event streams from neuromorphic cameras using Transformer architectures adapted to sparse, irregular temporal data represents a nascent but high-potential direction for energy-efficient high-speed visual processing.
- Causal and Temporal Vision Transformers: Extending spatial self-attention to the temporal domain with causal masking enables autoregressive video generation and action prediction. Developing memory-efficient temporal Transformers for long-horizon video understanding (hours-long video) remains unsolved.
- Uncertainty Quantification in Transformer Predictions: Medical and safety-critical applications require calibrated uncertainty estimates alongside predictions. Bayesian Transformers, Monte Carlo dropout, and conformal prediction frameworks adapted to Transformer architectures represent important directions for reliable deployment.
- Hardware-Software Co-Design: Developing attention mechanisms explicitly designed for systolic array architectures (TPUs, AI accelerators) and sparse matrix computation units would unlock the full efficiency potential of Transformer-based vision models for embedded and edge deployment.

While we have tried to provide a complete and updated overview of Vision Transformer architectures and applications in this review, there are several limitations that need to be acknowledged, because we rely on English-language publications indexed by major scientific databases, which could lead to regional or non-English publications being overlooked. Moreover, cross-paper benchmark comparison is by nature inaccurate, since the numbers reported in papers are subject to many confounding factors, including differences in training procedures, hardware configurations, and evaluation protocols. Additionally, we note that the most recent architectures may not be captured or reflected in the comparison, while the fast pace of development in the Transformer-based vision community continues.

VI. CONCLUSION

This systematic review provides a comprehensive survey of the dramatic impact that Transformer

architectures have had on digital image processing over the 2020-2026 period, covering 42 key methods drawn from 6 Transformer architectural families and 7 broad application domains, and it documents the development of the original Vision Transformer (ViT) through to hierarchical windowed approaches (e.g. Swin Transformer), CNN-Transformer hybrids (e.g. CoAtNet and CMT), more efficient attention variants, self-supervised masked autoencoders, and diffusion-Transformer image generation models. The evidence reviewed demonstrates that Transformers now constitute the current state-of-the-art results on all major image benchmarks, including ImageNet classification, COCO object detection, ADE20K semantic segmentation, all of the leading medical imaging benchmarks, and all of the standard image restoration datasets, however, they are also still clearly weak on computational efficiency, data requirements, and robustness to transfer to new domains. The development of promptable foundation models (e.g. SAM, MAE) and vision-language Transformers marks a transition from narrow, task-specific architectures to more general-purpose visual intelligence systems. Future progress will likely arise from bridging Transformers with state space models, from developing domain-specific visual foundation models, and from designing hardware-aware efficient attention mechanisms, therefore, we expect Transformers to remain the paradigm for high-performance visual processing with significant open research opportunities in equivariant architecture design, uncertainty estimation, and deployment on edge and embedded devices. Consequently, we hope that this review can serve as both a comprehensive reference and a practical guide for researchers and practitioners in the field, and accelerate future progress in Transformer-based visual intelligence.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[3] 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), 2017, doi: 10.48550/arXiv.1706.03762.

[4] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, doi: 10.48550/ARXIV.2010.11929.

[5] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[6] Y. Li et al., "BEiT V2: Masked image modeling with vector-quantized visual tokenizers," in Advances in Neural Information Processing Systems (NeurIPS), 2022, doi: 10.48550/arXiv.2208.06366.

[7] J. N. Saeed and M. K. Hussein, "A Multi-ViTs-based approach for automatic rice leaf disease classification," Iraqi Journal of Science, vol. 66, no. 9, pp. 3938–3950, 2025, doi: 10.24996/ijs.2025.66.9.33.

[8] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 10096–10106, doi: 10.48550/arXiv.2104.00298.

[9] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 516–533, doi :10.48550/arXiv.2204.07118 .

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2020, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[12] N. Carion et al., "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[13] L. Alkahla, J. Saeed, and M. Hussein, "Empowering ovarian cancer subtype classification with parallel Swin transformers and WSI imaging," The International Arab Journal of Information Technology, vol. 21, no. 6, 2024, doi: 10.34028/iajit/21/6/5.

[14] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv, 2021, doi: 10.48550/arXiv.2102.04306.

[15] M. K. Hussein and A. Alqassab, "Deep learning in medical systems: A comprehensive review of applications and ethical challenges," Neutrosophic Optimization and Intelligent Systems, vol. 7, 2025, doi: 10.61356/j.nois.2025.7606.

[16] A. Zamir et al., "Restormer: Efficient transformer for high-resolution image restoration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 5728–5739, doi: 10.1109/CVPR52688.2022.00564.

[17] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, doi: 10.48550/arXiv.2010.04159.

[18] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 10347–10357, doi: 10.48550/arXiv.2012.12877.

[19] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in Proc. ICLR, 2022, doi: 10.48550/arXiv.2106.08254.

[20] X. Chu et al., "Conditional positional encodings for vision transformers," in Proc. ICLR, 2023, doi: 10.48550/arXiv.2102.10882.

[21] M. K. Hussein, L. T. Alkahla, and A. Alqassab, "Hyperspectral image classification using hybrid swarm feature selection and ensemble classifier," Ingénierie des Systèmes d'Information, vol. 29, no. 6, pp. 2367–2375, 2024, doi: 10.18280/isi.290624.

[22] Z. Liu et al., "Swin Transformer V2: Scaling up capacity and resolution," in Proc. IEEE/CVF CVPR, 2022, pp. 12009–12019, doi: 10.1109/CVPR52688.2022.01170.

[23] X. Dong et al., "CSWin Transformer: A general vision transformer backbone with cross-shaped windows," in Proc. IEEE/CVF CVPR, 2022, pp. 12124–12134, doi: 10.1109/CVPR52688.2022.01181.

[24] J. Yang et al., "Focal self-attention for local-global interactions in vision transformers," in Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 30585–30597, doi: 10.48550/arXiv.2107.12239.

[25] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in Proc. IEEE/CVF ICCV, 2021, pp. 22–31, doi: 10.1109/ICCV48922.2021.00009.

[26] B. Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in Proc. IEEE/CVF ICCV, 2021, pp. 12259–12269, doi: 10.1109/ICCV48922.2021.01204.

[27] Z. Dai et al., "CoAtNet: Marrying convolution and attention for all data sizes," in Advances in Neural Information Processing Systems (NeurIPS), 2021, doi: 10.48550/arXiv.2106.04803.

[28] M. K. Hussein, L. T. Alkahla, and A. Alqassab, "Increasing the accuracy of melanoma classification by exploiting firefly algorithm and fine-tuned CNNs," in AIP Conference Proceedings, vol. 3264, no. 1, art. no. 040011, 2025, doi: 10.1063/5.0259165.

[29] J. Chen et al., "CMT: Convolutional neural networks meet vision transformers," in Proc. IEEE/CVF CVPR, 2022, pp. 12175–12185, doi: 10.1109/CVPR52688.2022.01186.

[30] M. K. Hussein, A. Alqassab, and L. T. Alkahla, "Enhancing feature selection in network intrusion detection systems using a novel hybrid binary swarm algorithm," Baghdad Science Journal, vol. 22, no. 7, pp. 2429–2437, 2025, doi: 10.21123/2411-7986.5007.

[31] K. Choromanski et al., "Rethinking attention with Performers," in Proc. ICLR, 2021, doi: 10.48550/arXiv.2009.14794.

[32] Y. Xiong et al., "Nyströmformer: A Nyström-based self-attention mechanism," in Proc. AAAI, 2021, pp. 14138–14148, doi: 10.1609/aaai.v35i16.17575.

[33] H. Wang et al., "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 108–126, doi: 10.1007/978-3-030-58548-8_7.

[34] K. He et al., "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF CVPR, 2022, pp. 16000–16009, doi: 10.1109/CVPR52688.2022.01553.

[35] M. Caron et al., "Emerging properties in self-supervised vision transformers," in Proc. IEEE/CVF ICCV, 2021, pp. 9650–9660, doi: 10.1109/ICCV48922.2021.00951.

[36] J. Zhou et al., "iBOT: Image BERT pre-training with online tokenizer," in Proc. ICLR, 2022, doi: 10.48550/arXiv.2111.07832.

[37] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in Proc. IEEE/CVF CVPR, 2022, pp. 9653–9663, doi: 10.1109/CVPR52688.2022.00943.

[38] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in Proc. IEEE/CVF ICCV, 2023, pp. 4195–4205, doi: 10.1109/ICCV51070.2023.00387.

[39] F. Bao et al., "All are worth words: A ViT backbone for diffusion models," in Proc. IEEE/CVF CVPR, 2023, pp. 22669–22679, doi: 10.1109/CVPR52729.2023.02172.

[40] P. Esser et al., "Scaling rectified flow transformers for high-resolution image synthesis," in Proc. ICML, 2024, doi: 10.48550/arXiv.2403.03206.

[41] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. ICLR, 2021, doi: 10.48550/arXiv.2010.04159.

[42] H. Zhang et al., "DINO: DETR with improved deNoising anchor boxes for end-to-end object detection," in Proc. ICLR, 2023, doi: 10.48550/arXiv.2203.03605.

[43] R. Strudel et al., "Segmenter: Transformer for semantic segmentation," in Proc. IEEE/CVF ICCV, 2021, pp. 7262–7272, doi: 10.1109/ICCV48922.2021.00717.

[44] B. Cheng et al., "Masked-attention mask transformer for universal image segmentation," in Proc. IEEE/CVF CVPR, 2022, pp. 1290–1299, doi: 10.1109/CVPR52688.2022.00135.

[45] A. Kirillov et al., "Segment anything," in Proc. IEEE/CVF ICCV, 2023, pp. 4015–4026, doi: 10.1109/ICCV51070.2023.00371.

[46] H. Chen et al., "Pre-trained image processing transformer," in Proc. IEEE/CVF CVPR, 2021, pp. 12299–12310, doi: 10.1109/CVPR46437.2021.01212.

[47] J. Liang et al., "SwinIR: Image restoration using Swin transformer," in Proc. IEEE/CVF ICCVW, 2021, pp. 1833–1844, doi: 10.1109/ICCVW54120.2021.00210.

[48] S. W. Zamir et al., "Restormer: Efficient transformer for high-resolution image restoration," in Proc. IEEE/CVF CVPR, 2022, pp. 5728–5739, doi: 10.1109/CVPR52688.2022.00564.

[49] L. Chen et al., "Simple baselines for image restoration," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 17–33, doi: 10.1007/978-3-031-20071-7_2.

[50] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv, 2021, doi: 10.48550/arXiv.2102.04306.

[51] A. Hatamizadeh et al., "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in MICCAI BrainLes Workshop, 2022, pp. 272–284, doi: 10.1007/978-3-031-08999-2_22.

[52] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in Proc. IEEE/CVF CVPR, 2022, pp. 16144–16155, doi: 10.1109/CVPR52688.2022.01567.

[53] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," in Advances in Neural Information Processing Systems (NeurIPS), 2021, doi: 10.48550/arXiv.2102.07074.