Vision Transformers in Image Understanding: A New Paradigm

Erfan Yekehzare, Seyyed Parsa Sharifi, Sayedmehdi Ayati Najafabadi, Pouria Matinifard
Department of IEF, University of Rostock, Rostock, Germany
{erfan.yekehzare, seyyed.sharifi, sa.najafabadi, pouria.matinifard}@uni-rostock.de

***Abstract*—Vision Transformers (ViTs) have emerged as a powerful alternative to convolutional neural networks (CNNs) for a variety of image understanding tasks. Derived from the Transformer architecture employed in natural language processing, ViTs eliminate convolutional inductive biases, using self-attention instead to capture global dependencies across image regions. This review surveys key developments since the original ViT model by Dosovitskiy et al. [1], including hierarchical architectures such as the Swin Transformer [2] and the Pyramid Vision Transformer [3]. These architectures improve scalability and localization for dense prediction. Additionally, we examine data-efficient training techniques and highlight the impact of large-scale pretraining and self-supervised methods, such as masked autoencoders (MAEs) [4] and DINO [5]. Furthermore, we explore the performance of ViTs in domains such as object detection, medical imaging, and video understanding. Finally, we discuss core limitations—computational cost, data dependence, and weak locality—and outline promising directions for future research. This work provides a concise yet comprehensive overview of ViTs as a transformative paradigm in computer vision.**

***Index Terms*—Vision Transformers, Self-Attention, Image Understanding, Transfer Learning, Self-Supervised Learning, Object Detection, Semantic Segmentation, Video Recognition, Deep Learning.**

I. Introduction

The Vision Transformer (ViT) presents notable benefits over conventional convolutional neural networks (CNNs), primarily due to its reliance on fewer inductive biases and its strong performance on large-scale datasets [1]. In comparison, CNNs embed strong assumptions such as locality and translation invariance, which enhance their data efficiency, particularly on small to medium-sized datasets [6]. Nonetheless, in low-data settings, ViTs often lag behind CNNs unless complemented with extensive data augmentation or hybrid architectures [6], [7].

Despite the longstanding dominance of CNNs in visual recognition, ViTs have demonstrated superior scalability as data volume and model size increase [1], [7]. For example, when trained on large datasets like JFT-300M, ViTs can significantly outperform CNN models such as ResNet on benchmarks like ImageNet [1]. Additionally, advancements like the Swin Transformer incorporate hierarchical structures and localized attention mechanisms to refine ViT's capabilities, narrowing the performance gap in dense prediction tasks such as object detection and segmentation [2], [3].

Consequently, while CNNs are generally more suitable for data-limited tasks with strong spatial priors, ViTs are well-suited for large-scale applications, offering adaptable architectures and substantial representational power when adequately trained.

II. Architectural Variants and Improvements

Several enhancements have been proposed to improve the original Vision Transformer (ViT) architecture, targeting its limitations in spatial resolution, scalability, and efficiency. A key focus has been the trade-off in patch size, where smaller patches improve spatial detail but increase computational cost. Hierarchical structures, as in Swin Transformer [2] and Pyramid Vision Transformer (PVT) [3], introduce multi-scale representations for better feature learning in dense prediction tasks. Swin employs window-based attention and shifted windows for efficient computation, while PVT reduces resource usage with spatial-reduction attention. Hybrid models [1], [6] combine CNN and Transformer strengths for better data efficiency. Additionally, advances in token reduction and positional embeddings improve ViT performance on high-resolution and variable-sized inputs.

A. Patch Size vs. Performance Trade-offs

The patch size in Vision Transformers (ViTs) significantly influences the balance between spatial detail, computational cost, and overall performance. While the original ViT used 16×16 patches to achieve strong results on large datasets like ImageNet-21k and JFT-300M [1], later models such as Swin Transformer [2] demonstrated that smaller patches (e.g., 4×4) preserve spatial information better—beneficial for dense prediction tasks. However, smaller patches increase the number of tokens, raising computational complexity, whereas larger patches reduce cost but risk losing fine details. DeiT [7] retained the 16×16 setting, focusing instead on data efficiency through training strategies. Empirical studies, such as those in [8], suggest that performance changes are modest when patch size and image resolution are scaled together, emphasizing that the total number of patches plays a more critical role. These trade-offs are quantitatively illustrated in Table I.

B. Hierarchical Vision Transformers

Hierarchical Vision Transformers address limitations of the original ViT architecture, particularly the fixed spatial resolution and lack of locality. Unlike ViT, which processes a flat sequence of patches, hierarchical models such as Swin Transformer [2] and Pyramid Vision Transformer (PVT) [3]

TABLE I

Top-1 accuracy on INet10 for ViT models with varying patch sizes and resolutions. B/S denote base/small models; numbers indicate patch size [8].

Model	B/32	B/48	B/64	S/16	S/24	S/32
Res.	224	336	448	224	336	448
INet10	64.43	64.65	64.67	63.42	63.79	63.50

adopt a multi-scale architecture, gradually reducing spatial resolution across layers. This structure allows better modeling of both local and global features, improving performance in dense prediction tasks. Swin Transformer, for instance, uses shifted window-based attention to enable cross-window interactions efficiently. These architectural innovations yield models that scale better and achieve state-of-the-art results in semantic segmentation and object detection.

Swin Transformer. The Swin Transformer addresses key challenges in adapting Transformer models from language to computer vision, specifically the large variations in visual entity scale, unlike fixed-size word tokens [9]–[11], and high image resolution, which makes global self-attention inefficient due to quadratic complexity [1], [12]. Tasks like semantic segmentation also require pixel-level predictions, which are intractable with standard Transformers.

Swin solves this by using non-overlapping local windows for self-attention, keeping complexity linear to image size. Shifted windows connect across windows between layers, improving modeling and hardware efficiency [13], [14]. Furthermore, hierarchical feature maps are built by starting from small patches and merging them deeper in the network, similar to CNN pyramids [9], [15].

Architecturally, Swin (e.g., Swin-T) begins by splitting an RGB image into 4x4 non-overlapping patches, projected into a dimension C. The model uses four stages: Stage 1 has standard Transformer blocks, while Stages 2–4 reduce token count via patch merging, forming a hierarchical representation. Each Swin Transformer Block replaces global attention with Window-based Multi-head Self-Attention (W-MSA) and Shifted Window MSA (SW-MSA). Global attention scales quadratically $(hw)^2$, while window-based attention scales linearly when window size M is fixed A cyclic shift strategy is used for efficient shifted attention, and relative positional bias is added in attention [16].

TABLE II

Swin models compared with key baselines on ImageNet-1K .

Model	**Params**	**FLOPs**	**Top-1 Acc.**
ResNet-50	25M	4.1G	76.5%
DeiT-S	22M	4.6G	79.8%
ViT-B/16	86M	55.4G	81.8%
Swin-T	29M	4.5G	81.3%
Swin-B (224^2)	88M	15.4G	83.5%
Swin-B (384^2)	88M	47.0G	84.5%
Swin-L	197M	103.9G	87.3%

Thanks to these innovations, Swin Transformer performs exceptionally well across tasks: 87.3% top-1 accuracy on ImageNet-1K [17] (classification), 58.7 box AP / 51.1 mask AP on COCO [18] (detection), and 53.5 mIoU on ADE20K (segmentation), surpassing previous methods by large margins. Swin-T outperforms ResNet-50 [19] and DeiT-S [7]. Swin-B outperforms ViT-B [1] and Swin-L achieves 87.3%, +0.9% over Swin-B.

Pyramid Vision Transformer (PVT). A novel Transformer backbone, the Pyramid Vision Transformer (PVT) [3], addresses key limitations of conventional Transformer architectures in computer vision. PVT overcomes these challenges by using fine-grained image patches (e.g., 4x4 pixels) for high-resolution representation, crucial for dense prediction tasks. It employs a progressive shrinking pyramid to reduce sequence length and computational cost as the network deepens, and incorporates a spatial-reduction attention (SRA) layer to further optimize resource consumption for high-resolution features.

TABLE III

Object detection performance on COCO val2017 using RetinaNet. PVT models outperform CNN-based backbones under comparable parameter counts.

Backbone	**Params**	**AP (val2017)**
ResNet-50 [20]	37.7M	36.3
PVT-Small [3]	34.2M	**40.4**
ResNeXt101-64x4d [21]	95.5M	41.0
PVT-Large [3]	71.1M	**42.6**

The PVT offers significant advantages over traditional CNNs and the original ViT. Unlike CNNs, which have local receptive fields, PVT consistently provides a global receptive field, making it more suitable for detection and segmentation. Compared to ViT, PVT's pyramid structure allows for easier integration into dense prediction pipelines like RetinaNet [22] and Mask R-CNN [23]. It can also form convolution-free pipelines when paired with other Transformer decoders, such as PVT+DETR [24] for object detection.

In performance, PVT demonstrates strong results. With RetinaNet [22], PVT-Small achieves 40.4 AP on COCO val2017, outperforming ResNet50 by 4.1 points. PVT-Large reaches 42.6 AP, 1.6 points better than ResNeXt101-64x4d, with 30% fewer parameters.

Similar to CNNs [25], PVT [3] features a four-stage hierarchical architecture, each producing feature maps at different scales. A key innovation is its progressive shrinking strategy via patch embedding layers, which controls feature map scale by dividing input into patches and projecting them into lower-dimensional embeddings. This offers greater flexibility in constructing the feature pyramid than traditional convolutional strides.

The spatial-reduction attention (SRA) layer is another critical component, replacing the standard multi-head attention (MHA) [12] in the Transformer encoder. SRA reduces the spatial scale of the key (K) and value (V) inputs before attention calculation. This dramatically cuts computational and memory overhead for high-resolution feature maps, allowing PVT [3] to operate efficiently with limited resources.

C. Hybrid CNN-Transformer Architectures

Hybrid architectures combine the locality and inductive biases of CNNs with the global modeling capacity of Transformers. The original ViT explored a hybrid model that uses a ResNet-50 backbone to extract features before applying the Transformer encoder [1]. CvT (Convolutional Vision Transformer) further integrates convolutions into both the embedding and attention mechanisms [6], enhancing spatial encoding and locality. These models exhibit better data efficiency, faster convergence, and improved performance on datasets like ImageNet. By merging strengths from both paradigms, hybrid models provide a balanced solution for medium-sized datasets.

D. Token Reduction and Efficient Attention Mechanisms

Token reduction and efficient attention mechanisms are essential for scaling ViTs to high-resolution images. The quadratic complexity of standard self-attention, as in the original ViT [1], limits scalability. PiT introduces pooling-based token reduction that mimics CNN-like downsampling [26], while NesT adopts a nested hierarchical structure to process fewer tokens progressively [27]. Techniques like Performer [28] and Linformer [29] approximate self-attention to reduce complexity to linear time. Swin Transformer [2], with its windowed attention, also achieves linear computational cost while maintaining strong performance. These innovations significantly improve the applicability of ViTs to dense tasks and large inputs.

E. Learned vs. Fixed Positional Embeddings

Since Transformers lack an inherent understanding of spatial structure, positional embeddings are necessary in ViTs. The original ViT employs both learned and fixed 2D sine-cosine embeddings [1]. Fixed embeddings offer better generalization in low-data regimes, while learned embeddings provide more flexibility and often better performance on large datasets. Later models like Swin Transformer [2] and Focal Transformer [30] use relative positional encodings within local windows to support variable input sizes. DETR [24] uses learned sinusoidal embeddings to align spatial locations effectively for object detection. The choice of positional embedding method depends on the task and data scale—fixed embeddings for robustness, learned ones for adaptability.

III. Training Strategies and Data Efficiency

A. Pre-training on Large-Scale Datasets

Pre-training on massive datasets, such as JFT-300M, significantly improves the performance of ViTs. Dosovitskiy et al. [1] showed that a ViT-L/16 model trained on JFT-300M outperforms CNN baselines on ImageNet. Zhai et al. [8] confirmed that both the size of the dataset and the size of the model are critical factors. Raghu et al. [31] added a representational view, finding that ViTs exhibit smoother CKA similarity across layers when pretrained on larger datasets. This indicates that they have more transferable representations.

TABLE IV
Top-1 accuracy of ViT models with different pre-training datasets.

Pre-training Dataset	Model	Top-1 Accuracy (%)
ImageNet-1k	ViT-B/16	76.5
ImageNet-21k	ViT-L/16	84.2
JFT-300M	ViT-L/16	**88.55**

B. Fine-tuning in Low-Data Regimes (VTAB Suite)

The VTAB benchmark includes 19 tasks, each with 1,000 training samples. Pretraining on JFT-300M improves performance; for example, ViT-H/14 achieves 77.6% on VTAB [8]. Raghu et al. [31] found that lower-layer representations remain consistent with as little as 3% of the JFT data, aiding transfer. Unlike CNNs, which build locality hierarchically, ViTs' self-attention allows early access to global context.

TABLE V
VTAB benchmark results for models with various pretraining sources.

Model	Dataset	VTAB Score (%)
BiT-M	JFT-300M	75.7
S4L	ImageNet-1k	67.2
ViT-H/14	JFT-300M	**77.63**
DeiT-B	ImageNet-1k	72.7

C. Self-Supervised Learning: MPP vs. Contrastive Learning

Self-supervised learning (SSL) enables ViTs to learn without human-labeled data. MAE [4] uses masked patch prediction (MPP) to reconstruct missing image regions and achieves 79.1%, outperforming fully supervised ViT-B/16 (77.9%) on ImageNet-1k, despite using no labels during pretraining. Raghu et al. [31] showed that ViTs preserve spatial relationships better than CNNs, which supports the effectiveness of MPP. Contrastive methods such as DINO and MoCo [5], [32] maximize agreement between augmented views, with DINO leveraging a self-distillation framework and MoCo using a memory bank-based approach.

TABLE VI
Comparison of MAE and supervised training methods on ImageNet-1k.

Method	Dataset	Top-1 Accuracy (%)
MAE	ImageNet-1k	**79.1**
Supervised	ImageNet-1k	77.9
SimMIM	ImageNet-1k	78.8

D. Transfer Learning and Few-Shot Performance

ViTs' uniform layer-wise representation and global token mixing benefit few-shot learning. Dosovitskiy et al. [1] showed that ViTs excel in five-shot transfer learning. Raghu et al. [31] linked this to strong residual connections and stable intermediate features. Early access to global features and layer-to-layer similarity facilitate feature reuse.

TABLE VII
5-SHOT ACCURACY OF VITS AND CNNS ON FEW-SHOT LEARNING TASKS.

Dataset	Model	Accuracy (%)
CIFAR-100	ViT-L/16 (JFT)	**87.1**
CIFAR-100	BiT-L (JFT)	83.9
VTAB Avg.	ViT-H/14 (JFT)	**77.6**
VTAB Avg.	DeiT-B (IN1k)	72.7

IV. Interpretability and Attention Analysis

Understanding how Vision Transformers (ViTs) make decisions is crucial for model transparency, particularly in safety-critical applications. Below, we summarize three key perspectives on ViT interpretability and attention behavior.

A. Visualizing Self-Attention Maps

In ViTs, a learnable *class token* aggregates information from patch tokens via self-attention across all layers. Each layer's attention map quantifies how much each patch contributes to every other patch, including the class token. The *attention rollout* method recursively multiplies attention matrices from all layers to trace the flow of information and reveal the cumulative contribution of each patch to the class token. Incorporating gradients of the loss with respect to attention weights using methods such as Grad-Rollout yields more accurate importance maps that highlight image regions that are heavily attended to and critical to the final decision [33]. Visualizing these maps as heatmaps over the input image provides explanations of ViT predictions that are semantically meaningful.

B. Patch Level Semantic Attention vs. Pixel Level Representations

Unlike CNNs, which build representations from local pixels upward, ViTs operate on non-overlapping image patches from the beginning. This patch-based design enables semantic attention across the entire image, even in the initial layers. In "Patch-Level Representation Learning for Self-Supervised Vision Transformers," Yun et al. [34] propose *SelfPatch*, which leverages the assumption that neighboring patches share semantic context. During contrastive learning, SelfPatch treats spatially adjacent patches as positive pairs. The model learns to bring these positive pairs (semantically similar patches) closer together in representation space. This produces richer, localized features than random patch selection. Ablation studies demonstrate that using neighboring patches as positives yields superior patch-level representations and show that ViTs inherently capture global semantics earlier than pixel-focused CNNs.

C. Attention Distance and Receptive Fields in Transformer Layers

Although ViTs theoretically enable each token to attend globally from the first layer, empirical analyses reveal a progression from local to global attention. Raghu et al. [31] measure *attention distance*—the average spatial separation between source and attended tokens—and show that early layers focus on nearby patches while deeper layers attend more globally. Similarly, the receptive fields of ViTs expand rapidly, becoming nearly global by the middle of the network, whereas the receptive fields of ResNets grow more gradually. Thus, despite their full self-attention, ViTs emulate CNN-like locality in the early layers before leveraging the global context in the later layers.

V. Applications and Specialized Domains

A. Object Detection and Segmentation

Vision Transformers (ViTs) introduced a new approach to object detection, enabling end-to-end set prediction without anchor boxes or non-maximum suppression. Key milestones in this shift include DETR [24], Deformable DETR [35], and Swin Transformer [2].

The DETR model formulates object detection as a bipartite matching problem. In this model, each object query predicts a bounding box and a class label. Although this approach is architecturally elegant, it suffers from slow convergence and suboptimal detection of small objects. Deformable DETR addresses these issues by applying sparse attention to a limited number of sampling points, thereby reducing complexity and improving performance. Deformable DETR achieves faster convergence and improves the AP for small objects from 20.5 to 25.1 on COCO [35].

The Swin Transformer acts as a hierarchical ViT backbone that uses shifted-window attention to scale linearly with image resolution. When integrated into Mask R-CNN, Swin Transformer surpasses ResNet-based detectors. In COCO evaluations, Swin-B achieves 50.5 box AP and 44.5 mask AP [2] (see Table VIII).

TABLE VIII
COCO VAL2017 PERFORMANCE OF DETR AND VARIANTS [24]

Model	#Params	AP	$\mathrm{AP_S}$	$\mathrm{AP_L}$
DETR (R50)	41M	42.0	20.5	61.1
DETR-DC5 (R50)	41M	43.3	22.5	61.1
DETR (R101)	60M	43.5	21.9	61.8
DETR-DC5 (R101)	60M	44.9	23.7	62.3

B. Medical Image Classification

Tasks in medical imaging benefit from the ability of ViTs to model long-range spatial dependencies across anatomical structures or modalities. In TransMed [36], CNN-encoded T1 and T2 MRI sequences are fused using a Transformer to improve cross-modality representation and boost tumor classification accuracy by 10.1% over CNN baselines.

ViTs have also proven effective in diagnosing COVID-19. For example, Chen et al. [37] fine-tuned ViT-Base [1] on chest X-rays and achieved 95.79% accuracy and 98.58% recall on a multi-class classification task. AUC scores approach 0.999 for cases of the disease. Their analysis shows that 10 encoder blocks optimize performance while avoiding overfitting. These results are summarized in Table IX.

TABLE IX
COVID-19 CHEST X-RAY CLASSIFICATION RESULTS [37]. EFFICIENTNET MODELS ARE FROM [38], MVIT2 FROM [39], AND VIT-BASE VARIATIONS FROM [1].

Model	Accuracy	Recall	F1 Score
EfficientNet-B0	94.65%	98.58%	98.44%
EfficientNet-B3	94.51%	98.59%	98.31%
EfficientNet-B5	93.94%	99.43%	98.32%
MViT2	94.41%	96.88%	97.29%
ViT-Base-patch8	95.41%	99.15%	98.87%
ViT-Base-patch16	95.22%	98.31%	98.58%
ViT-Base-patch32	93.71%	98.29%	97.88%
Proposed Model	**95.79%**	**98.58%**	**98.73%**

C. *Video Understanding and Temporal Attention*

Video ViTs must efficiently process spatio-temporal sequences. The Video Swin Transformer [40] extends the 2D Swin Transformer to the 3D domain by incorporating spatiotemporal shifted windows, which enables scalable video modeling. On the SSv2 dataset [41], Swin-B achieves a top-1 accuracy of 69.6%, outperforming MViT [42] and ViViT.

The TimeSformer [43] uses a factorized attention mechanism that decouples spatial and temporal attention. Experiments show that applying temporal attention before spatial attention improves accuracy by 0.5%, while reducing patch granularity lowers performance by up to 3%. These results highlight the importance of temporal ordering and spatial resolution in ViT-based video models. (see Table X).

TABLE X
TOP-1 AND TOP-5 ACCURACY ON SSV2 BENCHMARK. **K400**: KINETICS-400 DATASET [44]. [40], [43]

Model	Pretrain	Top-1	Top-5
TimeSformer-HR [43]	ImageNet-21K	62.5	–
TSM-RGB [45]	K400	63.3	88.2
SlowFast R101, 8×8 [46]	K400	63.1	87.6
ViViT-L/16x2 [47]	–	65.4	89.8
MViT-B, 64×3 [42]	K400	67.7	90.9
Swin-B (proposed)	K400	**69.6**	**92.7**

Overall, these applications demonstrate the versatility of ViTs in various domains, ranging from structured clinical data to dense spatio-temporal tasks. This supports the idea that ViTs are emerging as a general-purpose vision backbone.

VI. CHALLENGES, LIMITATIONS, AND FUTURE DIRECTIONS

Although Vision Transformers (ViTs) have demonstrated impressive performance across classification, segmentation, and video understanding tasks, several fundamental challenges remain.

A. *Data Requirements and Transferability*

ViTs heavily rely on large-scale pretraining to achieve strong performance, particularly in low-data regimes. Models such as ViT-L/16 show substantial gains when trained on datasets like JFT-300M rather than ImageNet-1k [1]. However, this reliance limits transferability to specialized domains, such as medical or remote sensing imagery, without extensive fine-tuning [48]. Future work may focus on enhancing data efficiency through self-supervised methods (e.g., MAE, DINO) and lightweight hybrid CNN-ViT models to better adapt to limited or domain-specific datasets.

B. *Computational Complexity at Scale*

Standard self-attention scales quadratically with the number of tokens, posing challenges for high-resolution images and spatiotemporal sequences. As shown in temporal models like TimeSformer and Video Swin, increasing spatial or temporal resolution incurs significant training costs [40]. While strategies such as window-based attention mitigate some of this cost, scalable and efficient attention mechanisms—such as sparse or linear attention—remain an open research direction.

C. *Locality and Fine-Grained Detail*

Unlike CNNs, which encode strong locality priors, ViTs operate on global self-attention from the start, which can limit their ability to capture fine-grained spatial details. Interpretability studies reveal that ViTs progressively increase their receptive field across layers rather than starting globally in practice [31]. Incorporating architectural components such as convolutions or localized attention may improve fine-detail recognition, particularly in domains like medical imaging and semantic segmentation.

Addressing these limitations is essential for improving the robustness, efficiency, and applicability of ViTs across both general-purpose and specialized image understanding tasks.

VII. CONCLUSION

Vision Transformers (ViTs) have redefined the landscape of image understanding, offering scalable, flexible, and effective alternatives to convolutional architectures. Initially introduced by ViT [1] and advanced through hierarchical designs such as the Swin Transformer [2], ViTs now achieve competitive performance in image classification, object detection, and semantic segmentation.

In parallel, training strategies such as masked autoencoding [4] and contrastive self-supervised learning [5] have enhanced ViTs' robustness and data efficiency, even in low-resource settings.

Despite these advances, challenges persist in computational complexity, weak locality priors, and reliance on large-scale datasets. Future research will likely focus on more efficient attention mechanisms, hybrid CNN-Transformer models, and domain-specific adaptations. The continued evolution of ViT architectures underscores their emergence as a transformative framework in modern computer vision.

ACKNOWLEDGMENT

All authors contributed substantially to the development of this review paper. Erfan Yekehzare was responsible for Section I: Introduction and Section II: Architectural Variants and Improvements. Seyyed Parsa Sharifi contributed Section V: Applications and Specialized Domains, Section VI: Challenges, Limitations, and Future Directions, the Abstract, and Section VII: Conclusion. Sayedmehdi Ayati Najafabadi worked on Section IV: Interpretability and Attention Analysis. Pouria Matinifard developed Section III: Training Strategies and Data Efficiency.

Seyyed Parsa Sharifi and Erfan Yekehzare collaboratively wrote, edited, and organized the entire manuscript.

The authors are affiliated with the Institute of Electrical Engineering and Information Technology (IEF), University of Rostock, and gratefully acknowledge the department's academic support.

Language and editorial improvements were supported using AI-based tools. ChatGPT (OpenAI) was used to assist in rephrasing and improving clarity, and DeepL Write was used to enrich sentence structure and paragraph flow. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[3] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021.

[4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *CVPR*, pp. 16000–16009, 2022.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

[6] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, 2021.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning (ICML)*, 2021.

[8] X. Zhai, A. Hassani, S. Mustikovela, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," *CVPR*, 2022.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection - snip," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] 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*, vol. 30, pp. 5998–6008, 2017.

[13] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," *arXiv preprint arXiv:1904.11491*, 2019.

[14] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351 of *LNCS*, pp. 234–241, Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[16] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," *CoRR*, vol. abs/1904.09925, 2020.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, (Zürich), September 2014.

[19] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," *arXiv preprint arXiv:2006.02334*, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[21] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2017.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations* (Y. Bengio and Y. LeCun, eds.), 2015.

[26] B. Heo, S. J. Yun, D. Han, S. Yoo, and S. Chun, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11936–11945, 2021.

[27] Z. Zhang, Y. Luo, H. Yan, Q. Zhao, Y. Lin, Z. Ma, and H. Li, "Aggregating nested transformers," *arXiv preprint arXiv:2105.12723*, 2021.

[28] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, *et al.*, "Rethinking attention with performers," in *International Conference on Learning Representations (ICLR)*, 2021.

[29] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

[30] J. Yang, Q. Hou, L. Zhou, M.-M. Cheng, and P. Luo, "Focal self-attention for local-global interactions in vision transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[31] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "Do vision transformers see like convolutional neural networks?," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12116–12127, 2021.

[32] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *ICCV*, 2021.

[33] X. Zhang, Y. Li, and Z. Wang, "Visual explanations of vision transformer guided by self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10785–10794, 2021.

[34] M. Yun, J. Lee, and H. Kim, "Patch-level representation learning for self-supervised vision transformers," in *International Conference on Learning Representations (ICLR)*, 2022.

[35] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[36] Z. Dai, Y. Chen, X. Xu, and Y. Yang, "Transmed: Transformers advance multi-modal medical image classification," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 246–256, 2021.

[37] Q. Chen, L. Zhao, M. Wang, and H. Li, "A vision transformer model for covid-19 diagnosis using chest x-ray images," *Diagnostics*, vol. 14, no. 3, p. 384, 2024.

[38] H. Cai, J. Lin, M. Hu, C. Gan, and S. Han, "Efficientvit: Multi-scale linear attention for high-resolution dense prediction." https://api.semanticscholar.org/CorpusID:262824134, 2022. Online; accessed 2025-06-04.

[39] Y. Li, C. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4804, 2022.

[40] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3202–3211, 2022.

[41] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5842–5850, 2017.

[42] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *arXiv preprint arXiv:2104.11227*, 2021.

[43] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *International Conference on Machine Learning (ICML)*, pp. 813–824, 2021.

[44] W. Kay, J. Carreira, K. Simonyan, *et al.*, "The kinetics human action video dataset," in *arXiv preprint arXiv:1705.06950*, 2017.

[45] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7083–7093, 2019.

[46] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6202–6211, 2019.

[47] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.

[48] Y. Ruan, Y. Li, P. Zhou, J. Wang, X. Zhao, X. Wang, E. Xie, Z. Li, and P. Luo, "Vision transformers: State of the art and research challenges," *arXiv preprint arXiv:2207.03041*, 2022.