

Transformer architectures for computer vision: A comprehensive review and future research directions

Tukaram Ugile^{1}, Dr. Nilesh Uke²*

¹PhD Research Scholar, Department of Computer Engineering, VIIT, Affiliation to Savitribai Phule Pune University (SPPU) Pune -411048, India

²Research Supervisor, Department of Computer Engineering, VIIT, Affiliation to Savitribai Phule Pune University (SPPU) Pune- 411048, India

Abstract. Long-range dependencies and contextual relationships in videos were captured by using Convolutional Neural Networks (CNNs) in past. Recently the use of Transformers is started for capturing the long-range dependencies and contextual relationships in videos. Transformers have made revolutionary impacts in Natural Language Processing (NLP) area and started making significant contributions in Computer Vision problems. So, it was required to perform the review of different Transformer Architectures in Computer Vision which will help to use them for different applications in Computer Vision. This paper provides a comprehensive review of the Transformer Architectures in Computer Vision, providing a detailed view about their evolution from Vision Transformers (ViTs) to more advanced variants of transformers like Swin Transformer, Transformer-XL, and Hybrid CNN-Transformer models. We have tried to make the study of the advantages of the Transformers over the traditional Convolutional Neural Networks (CNNs), their applications for Object Detection, Image Classification, Video Analysis, and their computational challenges. Finally, we discuss the future research directions, including the self-attention mechanisms, multi-modal learning, and lightweight architectures for Edge Computing.

Keywords

Abnormal Event Detection, Transformers, Transformer-XL, Vision Transformers, Video Vision Transformers, Long-range dependencies, Contextual Relationships

1 Introduction

Convolutional Neural Networks (CNNs) [1–3] depend upon the local receptive fields for capturing long range dependencies and contextual relationships. They face several challenges. Transformers have gained importance in Computer Vision field. Their self-

* Corresponding author: ugiletukaram@gmail.com

attention mechanism helps models to capture long range dependencies and contextual relationships. Transformers help to provide powerful framework for understanding complex spatial and temporal patterns in videos.

In Computer Vision, Transformers are widely used to different tasks like image classification, object detection, segmentation, action recognition, and detection of abnormalities. Vision Transformers (ViT) [4] showed that Transformers outperform when pure transformer architectures are trained on large-scale datasets as compared to CNNs. They consider the image as sequences of patches analogous to the words in the sentence. Subsequent models such as Swin Transformer [5], DeiT [6], and TimeSformer [7] have extended this concept with hierarchical attention mechanism, efficient token representations, and spatio-temporal modelling capabilities.

The development of multimodal and hybrid approaches which combine CNNs or optical flow networks with self-attention modules for enhancement of feature representation are encouraged by adapting the Transformer-based architectures. So, this paradigm shifts from convolutional to attention-based vision models has marked a important step in the evolution of Computer Vision domain by providing the different possibilities for fine-grained video understanding, cross-modal learning, and temporal reasoning in video analysis.

There is a huge paradigm shift in Computer Vision with the adoption of Transformers, which were originally designed for Natural Language Processing (NLP). They overpowered the traditional deep learning models like Convolutional Neural Networks (CNNs) because of their abilities to capture the spatial hierarchies using local receptive fields. CNNs faced problems with long-range dependencies and global contexts which are very important for computer vision tasks like image classification, object detection, and segmentation.

Convolutions for image recognition were replaced by self-attention mechanisms provided by Vision Transformers (ViTs) by Dosovitskiy et al. (2020). Self-attention and multi-head attention mechanisms were leveraged by Transformers for modelling the complex relationships between image regions which enabled to provide superior performance on large scale datasets. After that many advanced adaptations of Transformers have been proposed like Data-efficient ViT (DeiT), Swin Transformer, Pyramidal Vision Transformer (PVT) [8], and Convolutional Vision Transformer (CVT) [9], each of which improved the efficiency, scalability, and performance for various computer vision tasks.

There were no reviews done about the transformers comparing their evolution, architectures, advantages, and applications. There is need to review for utilising them for solving the different problems in Computer Vision domain.

This paper explores the evolution of Transformer Architectures in computer vision, and their key design principles, applications, and future research directions. The study of advantages of transformers over the traditional Convolutional Neural Networks (CNNs) is performed, their applications for Object detection, Image Classification, Video analysis, and their computational challenges are discussed. Future research directions including self-attention mechanisms, multi-modal learning, and light weight architectures for edge computing are also discussed.

Section 2 provides the evolution of Transformers in Computer Vision domain. Section 3 lists the applications of Transformers in Computer Vision. Section 4 provides the comparative analysis of transformer architectures in Computer Vision. Limitations and challenges of transformers are provided by Section 5. Section 6 provides the future research directions for Transformers in Computer Vision domain. Section 7 provides the conclusion.

2 Evolution of transformers in computer vision

From the initial Natural Language Processing (NLP) applications to the state-of-art vision architectures, Transformers in Computer Vision have been significantly evolved.

The initial architecture was proposed by Vaswani et al. (2017)[10]. Researchers soon started exploring whether the self-attention mechanisms will replace the Convolutional Neural Networks (CNNs) or not for the computer vision tasks. Dosovitskiy et. al (2020) introduced the Vision Transformer (ViT). In ViTs, images are split into patches, and each patch is considered just like a token in NLP. The standard Transformer encoder is being fed by these patches, and self-attention mechanisms are used instead of convolutional layers. Hybrid architectures like Data -efficient ViT (DeiT) was proposed to train ViTs with fewer data so that they can be made more accessible for smaller datasets. Convolutional layers were added before the Transformers to extract local features better in Convolutional Vision Transformer (CVT). Dai et al. proposed hybrid transformers CoAtNet [11] where they combined convolutional and attention mechanisms for improving efficiency.

Liu et al. introduced Swin Transformer using a feature pyramid with shifted window attention to process images more efficiently which outperformed ViT and CNNs on classification, object detection, and segmentation tasks. TimeSFormer, ViViT , Video Swin Transformer [12] were proposed for video understanding. Light weighted and efficient Vision Transformers like EfficientFormer [13] and EdgeViT [14] were proposed for real-time applications on mobile and edge devices.

Multimodal and generative Vision Transformers like DINO [15], Blip-2 [16] & Flamingo [17], Segment Anything Model (SAM) [18] were proposed.

3 Applications of Transformers in Computer Vision

3.1 Image Classification

Before Transformers, CNNs were heavily used for Image Classification tasks. Global context can be captured easily using Transformers and self-attention mechanisms are used for modelling the long-range sequences. Out of these transformers, very well-known model is Vision Transformer (ViT). Vision Transformer was first proposed by Dosovitskiy et. al (2020) where they adapted the Transformer architecture from NLP domain to Computer Vision domain. ViTs outperform over CNNs on large datasets but they require significant amount of pretraining. Later many variations of Vision Transformers were proposed for them.

ViT divides the entire image into fixed-sized patches (i.e. 16*16 pixels) rather than processing it as whole image. Input tokens are created from these patches after flattening them into a vector. In CNNs the convolutions are used to have a built-in understanding of the spatial relationships. Instead, positional embeddings are used to model the spatial relationships in ViT. Transformer has encoder layers each containing Multi-Head Self Attention (MHSA), feedforward layers, Layer Normalization and skip connections. The sequence of tokens is passed through these layers. A special classification token is prepended each patch sequence. After passing through Transformer encoding process, this classification token contains the final representation of the feature which is then passed to the fully connected layer (MLP) for the purpose of classification.

ViT provides several benefits like it captures the global relationship using self-attention mechanism. There is no need to design the hand-crafted filters as it was done in CNNs. Attention mechanisms are used for learning the spatial relationships. ViTs provide better performance over the large datasets like ImageNet-21k [19,20]. The major limitations of ViTs are that they computationally expensive as they are using quadratic self- attention mechanisms and large datasets are required for training them as compared to CNNs.

3.1.1 Variants and improvements in ViTs Table 1 lists the different variants in ViTs and describes the improvements done in ViTs.

Table 1 Variants and Improvements in ViTs

Sr. No	Variant	Improvements
1	Data Efficient Transformer (DeiT) [6]	More data efficient, trainable on smaller datasets (e.g. ImageNet 1k [21])
2	Swin Transformer [5]	Shifted window for reducing computations, scalable, efficient, and suitable for high resolution images
3	Pyramid Vision Transformer (PVT) [8]	Multi-scale feature extraction due to pyramidal structure, used for object detection and segmentation
4	Hybrid CNN Transformer Models	Combination of CNNs and Transformers for better performance. e.g. Convolutional Multi-Head Self-Attention Transformer (CMT) [22]

3.1.2 Applications of transformers

Table 2 provides the detailed applications of the Transformers.

Table 2 Applications of transformers

Sr. No	Area	Application
1	Medical Imaging	Disease classification in X-rays, MRIs, CT scans
2	Autonomous Vehicle	Recognition of road signs and objects
3	Remote Sensing	Land cover classification
4	Face recognition	Identity Recognition
5	Retail & E-commerce	Product categorization and visual search

3.1.3 Future directions for transformer based image classification

Image Classification is improved significantly by using self-attention and global feature modelling in Transformers. However, data and computational efficiency are one of the major challenges. More efficient Transformers like Swin Transformer [5], PVT [8], CoAtNet [11] can be used for Image Classification. Dependency on the labeled data can be reduced by using Self supervised learning methods for ViTs. Lightweight Transformers can be used for mobile and Edge devices. Hybrid models integrating CNNs, Transformers and Graph Neural Networks (GNNs) [23,24,25] can be used. Future research should focus on making the Transformers faster, scalable, and more data efficient for real-time applications.

3.2 Object detection

Object detection is significantly improved by utilizing the self-attention mechanisms, feature extractions and spatial relations using Transformers compared to CNNs. Transformers are used in following ways for detection of objects.

3.2.1 Detection transformers (DETR)

One of the first Transformer-based Object Detection Models, DEtection TRansformer (DETR) was proposed by Carion et, al (2020) [26]. It uses CNN backbone (e.g. ResNet [27]) for the feature extraction purposes. These features are processed by encoder and decoder of Transformers. Detection of objects is done via bipartite matching by using the object queries which are the learned embeddings.

Major advantages are that there is no need of hand-crafted components like anchors and it can directly model the relationships between the objects. End-to-end training is done with strong performance and using large-scale datasets. Limitation is that they are computationally expensive for high resolution images and show slower convergence as compared to CNN-based methods like Faster R-CNN [28].

3.2.2 Deformable DETR

DETR is enhanced by Zhu et al. (2021) [29] into Deformable DETR using deformable attention which focuses only on a sparse set of key points instead of the entire feature map.

It improves the DETR by providing faster convergence, low computational costs, and better performance on the small objects.

3.2.3 Swin transformer for object detection

Liu et al. (2021) [5] introduced a hierarchical vision Transformer with shifted windows called Swin Transformer which improved the efficiency for object detection. It achieves state-of-the-art accuracy on COCO [30] dataset by reducing computational complexity. It works well with the standard object detection frameworks like Faster R-CNN [28] and Mask R-CNN [31]

3.2.4 Hybrid CNN-transformer object detection models

Hybrid CNN-Transformer Object Detection models combine the CNN and Transformers balancing the efficiency and accuracy. YOLO [31-36] applies pure Transformers just like ViTs for the detection of the objects. ViT with additional localization constraints, ViDT (Vision Transformer for Detection) [37] is used for Object detection. For optimizing the efficiency a light weight hybrid model, EfficientFormer-Det [38] is proposed.

Transformers provide end to end training, awareness about global context, removal of hand-crafted components like anchor boxes for detection of objects. On the other hand, the accuracy and interpretability are improved but there are still challenges like high computational costs and slow training as compared to traditional CNNs.

3.3 Video analysis and action recognition

Powerful self-attention mechanisms are used for modelling the spatial and temporal dependencies used by Transformers for Video analysis and action recognition. Video analysis includes challenges like Spatial Temporal relationships, long range dependencies, high computational costs and occlusions and motion blurs. Video representation is learned effectively by Transformers by modelling both spatial and temporal aspects.

3.3.1 Video vision transformer (ViViT)

ViT is extended for videos by Arnab et al. (2021) [39] by processing spatial and temporal tokens separately. It divides the video into spatiotemporal tokens like patches in ViT. Spatial features are captured using Self-attention and motion is tracked using temporal attention.

ViViT [39] can handle the long sequences of videos without RNNs [40],[41] or 3D CNNs [42],[43]. It required very large datasets for training of the models.

3.3.2 TimeSformer

Video frames are processed efficiently using divided space time attention by TimeSFormer [44]. Video frames are divided into patches and separate attention mechanisms are used for spatial and temporal dimensions.

As compared to 3D full attention, computational costs are reduced. It is efficient and scales well for the long sequences of videos but computational costs are still higher than CNNs.

3.3.3 Video swin transformer

Swin Transformer is extended to videos by hierarchical learning of features [45]. Local and global features are captured by shifted windows in Swin Transformers. Computational costs are reduced by processing the video frames hierarchically. Video Swin Transformer is faster and more scalable than the standard Transformers but it requires the pretraining on the large datasets.

3.3.4 Motion-transformer (MFormer) for action recognition

Motion feature layers are used explicitly for efficient capturing of the motions [46]. This Transformer uses motion cues instead of the just static frames. Movement trajectories are modelled using motion self-attention. Recognition of actions is done more accurately by focusing on how objects move, although it costs more computational requirements.

Human Action Recognition, Video captioning, Anomaly Detection, Autonomous Vehicles and Gesture Recognition etc. are the applications of Transformers for Video analysis. Future research directions should be towards reducing overhead for computations for real-time analysis, use of self-supervised learning for reducing the dependency on labelled data, Multi- Modal Fusion by combining the video with audio/text for better understanding of the scene and Long-Video understanding for handling the multi-minute video clips efficiently.

Models like ViViT, TimeSFormer and Video Swin Transformer are making significant progress and are replacing 3D CNNs and RNNs with powerful attention mechanisms. Although future research work should be done to make them faster, more efficient, and better at handling the long-term dependencies.

3.4 Medical imaging and remote sensing

Various tasks such as detection of diseases, segmentation and classification are done using Transformers which provide the powerful feature extraction, spatial understanding, and self-attention mechanisms that improve the accuracy and robustness.

3.4.1 Medical imaging

3.4.1.1 Medical image segmentation

Segmentation of medical images such as MRI, CT scans and histopathological images is done by using Vision Transformers (ViTs) and Swin Transformers. TransUNet [47] combines the Transformer and U-net architectures for the purpose of the medical image segmentation.

3.4.1.2 Disease diagnosis and classification

Images are analyzed by Transformers for detection of diseases such as cancer, pneumonia, and COVID-19. Detection of COVID-19 in Chest X-rays and CT scans was done by ViTs [4].

3.4.1.3 Multi-modal fusion (combination of medical text and images)

Integration of text reports with relevant images for comprehensive diagnosis is performed using Transformers. Radiology reports are combined with medical reports using Vision Language models like BioViL-T [48].

3.4.1.4 3D medical image processing

Volumetric segmentation and detection of lesions is performed for 3D medical images using Transformers. Medical image segmentation is performed using Swin UNETR [49] which is combination of Swin Transformer and U-Net Encoder-Decoder [50].

3.4.2 Remote sensing

Analysis of satellite and aerial images is done using remote sensing for environmental monitoring, land use classification, and management of disasters. Transformers are highly used for high-resolution image analysis.

3.4.2.1 Land cover classification

Classification of land cover types such as vegetation, water bodies, and urban areas are done by using ViTs and Swin Transformers. Satellite Vision Transformer (SATViT) [51] is used for observation of Earth.

3.4.2.2 Change detection

Detection of environmental changes like deforestation, urban expansion is done by Transformers. ChangeFormer [52] uses Transformers for bi-temporal remote sensing change detection.

3.4.2.3 Hyperspectral and multispectral image analysis

Processing of hyperspectral images for agriculture, mineral mapping, and pollution monitoring is performed by Transformers.

3.4.2.4 Disaster monitoring

Detection and monitoring of natural disasters like floods, wildfires and earthquakes etc. is done using Vision Transformers and Spatio-Temporal Transformers.

4 Comparative Analysis of Transformer Architectures in Computer Vision

Comparative analysis of Transformer architectures in computer vision is performed in Table 3.

Table 3 Comparative analysis of transformer architectures in computer vision

Architecture/Model	Core concept	Strength	Limitations
Vision Transformer (ViT)	Image patches are treated as tokens and global self-attention is applied	Long range dependencies are captured and outperforms CNNs	Large scale pretraining required, quadratic complexity, lacks spatial inductive bias
DeiT (Data-efficient ViT)	Optimized training for smaller datasets and knowledge distillation	Data efficiency improved, effectively trained on ImageNet-scale data	Sensitive to hyperparameters, lower performance on small datasets
Swin Transformer	Shifted window attention and hierarchical architecture	Efficiently scaled to high-resolution images, dense prediction is done strongly	Limited global context, complex design
DETR (Detection Transformer)	Object detection using end to end Encode-decoder Transformer	Object relations modelled globally, anchors and NMS removed	High computational cost and slow convergence
TimeSformer	Video understanding using factorized spatial-temporal attention	Spatial and temporal dependencies are captured, Stronger video recognition	High computation costs and memory required for long clips and no explicit motion priors

5 Challenges and Limitations

Computational complexity is one of the challenges as the self-attention scales quadratically with input size. Transformers require large datasets for improving the effectiveness of training. Unlike CNNs, the spatial hierarchies are not modelled by Transformers inherently. Limited interpretability is another challenge as it is challenging to understand the attention mechanisms in Transformers.

6 Future Research Directions

Sparse attention mechanisms and linear transformers can be used for reducing the complexity. Multimodal learning can be performed by combining computer vision and language models for tasks like image captioning and visual question answering (VQA). Optimized transformers can be used for mobile and embedded devices for Edge AI. Self-supervised Learning can be used for Vision Transformers by reducing the dependency on the labeled datasets using contrastive and masked image modelling techniques.

Identified Research gaps and potential research directions are provided in Table 4.

Table 4 Research gaps and future research directions

Research area	Limitation	Potential research direction
Long-range temporal modelling	High computational costs	Sparse/efficient or recurrent Transformers can be developed for extended temporal context
Data and Label efficiency	Dependency on large pretraining datasets	Domain-adaptive self-supervised or contrastive pretraining can be used
Semantic alignment in SSL	Reconstruction without semantic understanding	Multi-task or contrastive pretext objectives can be used
Interpretability and Robustness	Explainability of attention patterns is limited	Robust training methods and attention priors can be developed
Edge Deployment	Heavy models are not suitable for real-time	Lightweight or quantized Transformer variants can be designed
Multimodal Fusion	Cross modal attention is Inefficient	Temporal and modality aligned fusion modules are efficient

7 Conclusion

Transformers have shown considerable good performance in Computer Vision problems as compared to CNNs in several aspects. However, for their better utilization in future, challenges such as computational costs and data requirements should be considered. Future research works should focus on optimizing the architectures of transformers for computational efficiencies, interpretability, and multimodal capabilities so that they can be adopted in real-time and Edge AI applications.

References

1. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.
2. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
3. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012–10022).
5. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International conference on machine learning (pp. 10347–10357). PMLR.
6. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Icml* (Vol. 2, p. 4).
7. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., ... Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 568–578).
8. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 22–31).
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
10. Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34, 3965–3977.
11. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3202-3211.
12. Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). EfficientFormer: Vision Transformers at MobileNet Speed. *arXiv preprint arXiv:2206.01191*.
13. Chen, Z., Zhong, Z., Liu, Z., & Li, S. (2022). EdgeViT: Efficient Visual Modeling for Edge Computing. *arXiv preprint arXiv:2205.03436*.
14. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9650–9660). IEEE.
15. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
16. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*.
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

18. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
19. Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). ImageNet-21K pretraining for the masses. arXiv preprint arXiv:2104.10972.
20. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
21. Thapak, P., & Hore, P. (2020). Transformer++. arXiv preprint arXiv:2003.04974. <https://arxiv.org/abs/2003.04974>
22. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. IEEE Transactions on Neural Networks, 20(1), 61–80. <https://doi.org/10.1109/TNN.2008.2005605>
23. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. 5th International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1609.02907>
24. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1710.10903>
25. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision (ECCV). <https://arxiv.org/abs/2005.12872>
26. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. <https://doi.org/10.1109/CVPR.2016.90>
27. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS), 28. <https://arxiv.org/abs/1506.01497>
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/2010.04159>
29. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
30. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988. <https://doi.org/10.1109/ICCV.2017.322>
31. Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7263–7271. <https://doi.org/10.1109/CVPR.2017.690>
32. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767. <https://arxiv.org/abs/1804.02767>
33. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. <https://arxiv.org/abs/2004.10934>

34. Jocher, G., Chaurasia, A., & Qiu, J. (2022). YOLOv5 by Ultralytics. GitHub Repository. <https://github.com/ultralytics/yolov5>
35. Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696. <https://arxiv.org/abs/2207.02696>
36. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., & Yang, M.-H. (2021). ViDT: An efficient and effective fully transformer-based object detector. arXiv preprint arXiv:2110.03921.
37. Li, Y., Yuan, G., Wen, Y., Hu, E., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). EfficientFormer: Vision Transformers at MobileNet Speed. arXiv preprint arXiv:2206.01191.
38. Arnab, A., Deghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 6836-6846.
39. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
40. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. <https://arxiv.org/abs/1406.1078>
41. Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 221-231. <https://doi.org/10.1109/TPAMI.2012.59>
42. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, 4489-4497. <https://doi.org/10.1109/ICCV.2015.510>
43. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In M. Meila & T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 813–824). PMLR.
44. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3202-3211.
45. Patrick, M., Campbell, D., Asano, Y. M., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., & Henriques, J. F. (2021). Keeping your eye on the ball: Trajectory attention in video transformers. arXiv preprint arXiv:2106.05392.
46. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Xing, L. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
47. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M. P., Nori, A., Alvarez-Valle, J., & Oktay, O. (2023). Learning to exploit temporal structure for biomedical vision–language processing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
48. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv preprint arXiv:2201.01266.

49. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arXiv:1505.04597.
50. Fuller, A., Millard, K., & Green, J. R. (2022). SatViT: Pretraining Transformers for Earth Observation. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.
51. Bandara, W. G. C., & Patel, V. M. (2022). A Transformer-Based Siamese Network for Change Detection. In *IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium* (pp. 207-210). IEEE.