2023 7th IEEE Congress on Information Science and Technology (CiSt) | 978-1-6654-6133-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CiSt56084.2023.10410015

Vision Transformers: A Review of Architecture, Applications, and Future Directions

Abdelhafid Berroukham
Department of Computer Science
Faculty of Science, Ibne Tofail University
Kenitra, Morocco
a.berroukham@gmail.com

Khalid Housni
Department of Computer Science
Faculty of Science, Ibn Tofail University
Kenitra, Morocco
housni.khalid@uit.ac.ma

Mohammed Lahraichi
Department of Computer Science
Faculty of Science, Ibn Tofail University
Kenitra, Morocco
lahraichi.mohamed@gmail.com

Abstract— **In recent years, the development of deep learning has revolutionized the field of computer vision, especially the convolutional neural networks (CNNs), which become the preferred approach for numerous tasks handling images. However, CNNs have difficulty interpreting massive and complicated datasets, which has led to the creation of alternative architectures such as vision transformers. The transformer architecture, which was initially developed for natural language processing (NLP), is modified for image-related applications via vision transformers. In this paper, we present an outline of the main concepts and components of vision transformers. We review various variations and modifications to the architecture, and compare different approaches based on their effectiveness, complexity, and other attributes. Additionally, we examine the applications and uses of vision transformers, such as image classification, object detection, and semantic segmentation, and provide illustrations of relevant real-world situations. Finally, we discuss the potential impact of vision transformers on computer vision, while exploring the challenges and restrictions associated with their usage. We conclude by outlining potential new directions and advancements in the field of computer vision, as well as areas that require further study and investigation.**

Keywords—vision transformers, deep learning, computer vision

I. Introduction

In recent years, computer vision advanced significantly in various applications, such as image recognition[1], [2], object detection[3], [4], and semantic segmentation[5], [6]. Convolutional neural networks (CNNs) have been the dominant approach to tackling computer vision tasks[7], but they have several limitations. One of them is their fixed receptive field, which refers to the size of the input that the network takes into account when computing a particular feature. Small kernels are frequently used by CNNs to capture local features, and the information from various regions of the input image is integrated hierarchically through a number of layers. However, this approach might not properly capture global properties. Another limit of CNNs is the lack of explicit modeling of relationships between features, which can result in a subpar performance in tasks requiring comprehension of intricate relationships between objects or components of an image. Vision transformers aim to address these limitations by introducing a novel architecture based on the self-attention mechanism, this architecture can directly describe the relationships between features without the requirement of hierarchical feature extraction and this enhances the performance of vision transformers in a variety of computer vision tasks by enabling them to more accurately capture global information and represent complicated relationships between various components of an image[8]. Vision transformers perform remarkably well on image classification tasks, outperforming CNNs current state-of-the-art accuracy, and other computer vision tasks, where they have been successfully used. New opportunities for enhancing computer vision research and applications have emerged with the development of vision transformers.

The key idea behind vision transformers is self-attention, which allows the model to directly capture interactions between distinct regions of an input image. Unlike standard CNNs, which employ convolutional layers to extract local information hierarchically, vision transformers describe the relationships between all features in the input using a sequence of self-attention layers. The self-attention mechanism computes a weighted sum of the feature vectors at every position in the input based on their correlations with all other positions. The self-attention process is then applied to a series of patches of the input image, which are vectorized and supplied into the self-attention layers. Another important component of vision transformers is positional encoding, which adds information about the position of each patch in the input to the feature vectors. This enables the model to recognize spatial relationships among patches in the input image. Furthermore, vision transformers often employ a feedforward neural network to interpret the output of the self-attention layers and generate final predictions. Overall, fundamental concepts and components of vision transformers allow the model to capture global information and complicated interactions between different sections of an image, resulting in increased performance on a wide range of computer vision applications. In this paper, we will discuss the key concepts, architectures, applications, and future directions of vision

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.

transformers with the goal of providing insights into this quickly expanding topic.

II. Background

A. History of deep learning in computer vision

Deep learning has revolutionized computer vision, enabling considerable advances in picture classification [9], object detection [10], semantic segmentation[6], and other tasks. Deep learning in computer vision has a long history, dating back to the early 2010s, with the appearance of AlexNet model [4], which won the ImageNet Large Scale Visual Recognition Challenge in 2012 and achieved a considerable gain in accuracy over previous approaches. This achievement accelerated the development of convolutional neural networks (CNNs), which have now emerged as the dominant approach in computer vision. CNNs extract local features hierarchically using convolutional layers and pooling techniques to minimize the spatial resolution of the feature maps. This method has proven to be particularly effective at capturing meaningful patterns in images, allowing significant advancement in a variety of computer vision tasks. Since the rise of CNNs, researchers have continued to investigate novel architectures and approaches to increase the accuracy, efficiency, and interpretability of deep learning models in computer vision. Vision transformers are a recent example of this technique, which promises to extend deep learning's capabilities in computer vision beyond standard CNNs.

B. Development of vision transformers

Despite CNNs have been quite successful in many computer vision tasks, they have presented significant limitations that have prompted the development of additional techniques such as vision transformers[8]. One of CNNs' essential disadvantages, is its limited receptive field, because the network can only consider a limited spatial context when computing each feature. Which presents a difficulty for CNNs to recognize long-term dependencies or global features in the input image. Furthermore, CNNs do not explicitly model the relationships between distinct characteristics in the input, which can be useful for applications requiring complicated interactions between objects or portions of an image. Vision transformers overcome these restrictions by employing a self-attention mechanism that enables the model to capture global information and represent complicated interactions between different regions of the input image. This technique has shown considerable promise in delivering advanced results on a variety of computer vision tasks, and it has opened up opportunities for developing computer vision research and applications.

C. The key features of transformers

Transformers[11] is a neural network architecture that was originally designed for natural language processing (NLP) problems. They employ a self-attention mechanism, which enables the network to recognize complicated links between different components of the input sequence. This technique is especially excellent at capturing long-term dependencies and has outstanding success in NLP applications. This architecture is adapted for image-based tasks via vision transformers[8], which consider the input image as a sequence of patches that are fed into the transformer network. The core features of vision transformers are the utilization of self-attention layers to represent the interactions between the patches, as well as positional encoding to provide spatial information about the placement of each patch in the image. Furthermore, the transformer network often contains feedforward layers that process the output of the self-attention layers and generate the final predictions. By utilizing these fundamental features, vision transformers can successfully capture global information and complicated interactions between different sections of the input image, resulting in increased performance on a wide range of computer vision tasks.

III. Architectures of ViT and Variants

A. The original vision transformer architecture

In [8], the authors provided the original vision transformer architecture. It was created to address typical convolutional neural network difficulties in capturing global information and long-range relationships in images. The architecture is made up of a succession of transformer blocks, each with a multi-head self-attention mechanism and a feedforward neural network. The input image is first separated into non-overlapping patches, which are then linearly embedded and fed into the transformer blocks. Furthermore, the model encodes the spatial information of each patch using a learnable positional embedding. The output of the transformer blocks is pooled across all patches and fed through a feedforward neural network to obtain the final classification logits. Figure 1 shows an overview of the vision transformers architecture. The authors trained the model using the large-scale ImageNet dataset[12] and demonstrated that it beat state-of-the-art convolutional neural networks on image classification tasks, showing the usefulness of the vision transformer model for computer vision applications, Figure 2 shows some examples of the attention maps generated by ViT using the attention mechanism.

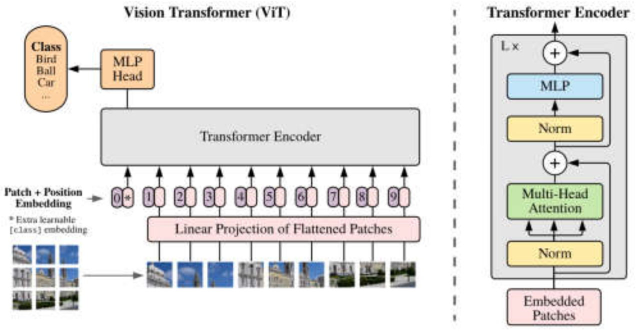

Figure 1: The original architecture of Vision transformer [8]

B. Variants of the original architecture of ViT

Since the original vision transformer paper was published[8], researchers have proposed several tweaks and variants of the architecture to increase its performance and efficiency on various computer vision applications. A good one is the DeiT (Data-efficient image Transformers) architecture introduced by [13], which is a type of Vision Transformer for image classification tasks that incorporates a teacher-student training technique that allows training the model with less labeled samples.

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.

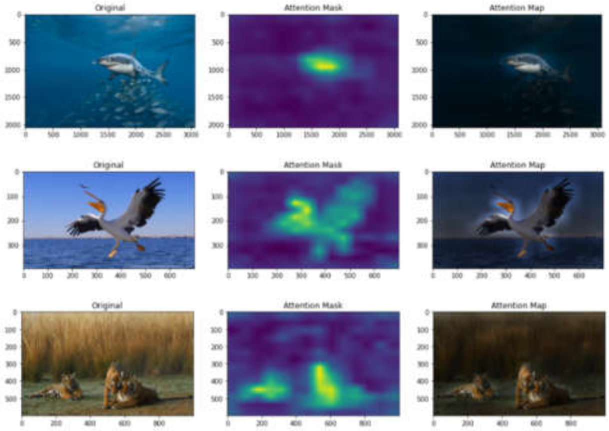

Figure 2: Samples of Attention mask and Attention map Generated by ViT [8]

DeiT also combines several regularization approaches, such as stochastic depth and knowledge distillation, to improve generalization performance, Figure 3 shows the distillation process in DeiT. The ViT-Hybrid model introduced by [14] is another variant of the vision transformer that combines the vision transformer with convolutional neural networks to leverage their complementing capabilities. The model utilizes a CNN backbone to extract low-level features from the input image, which are then coupled with the vision transformer's high-level features. This approach has shown improved performance on a variety of image classification benchmarks compared to employing only CNNs or vision transformers, figure 4 shows the architecture of this approach. Other researchers have proposed improvements to the original architecture to increase its efficiency, such as the Lite Transformer architecture [15], which employs a lightweight transformer block design and a dynamic patch size to decrease the model's computational and memory requirements, Figure 5 Shows the architecture of Lite Transformer. In general, these variants highlight the potential of vision transformers for furthering computer vision research and applications.

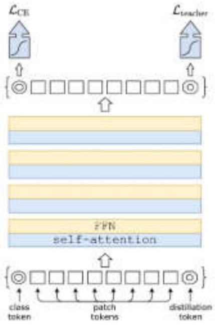

Figure 3: An overview of DeiT [13]

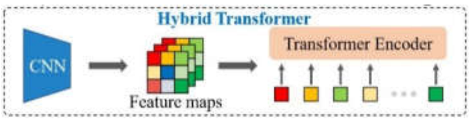

Figure 4: The architecture of ViT-Hybrid [14]

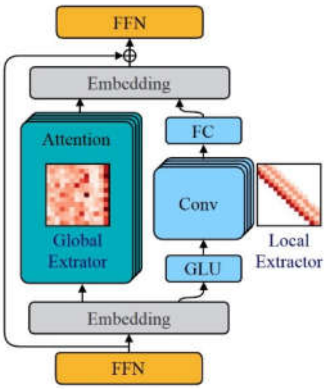

Figure 5: Lite Transformer architecture [15]

C. *Comparing the different variants of vision transformers*

The comparison and contrast of the various architectures and versions of vision transformers can provide insights into their strengths and limitations in different situations. In terms of performance, recent research has demonstrated that several versions, such as DeiT[13] and ViT-Hybrid[14], obtain state-of-the-art results on various image classification benchmarks while requiring significantly less computational and memory than the original architecture. However, performance may vary according to the task and dataset. Some variations, such as Lite Transformer[15] and Pyramid Vision Transformer (PVT) [16], have been created to minimize the model's computational and memory needs, making it more acceptable for deployment in resource-constrained environments. However, reducing model complexity may result in some performance loss. Furthermore, certain variations, such as PVT, include extra modules or architectures to tackle specific issues, such as dealing with objects of varied scales or forms. However, these new modules can increase the model's complexity and necessitate more computational resources. Figure 6 shows the architecture of Pyramid Vision Transformer (PVT). Overall, the choice of the architecture of the vision transformer depends on the specific task, performance requirements, and available resources.

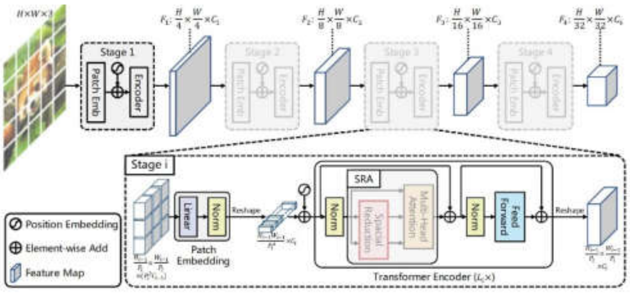
$H \times W \times 3$
$F_1: \frac{H}{4} \times \frac{W}{4} \times C_1$
$F_2: \frac{H}{8} \times \frac{W}{8} \times C_2$
$F_3: \frac{H}{16} \times \frac{W}{16} \times C_3$
$F_4: \frac{H}{32} \times \frac{W}{32} \times C_4$
Stage 1
Patch Emb
Encoder
Stage 2
Stage 3
Stage 4
Stage i
Linear
Norm
Reshape
Patch Embedding
SRA
Norm
Spatial Reduction
Multi-Head Attention
Norm
Feed Forward
Reshape
Transformer Encoder (L_i×)
Position Embedding
Element-wise Add
Feature Map

Figure 6: architecture of Pyramid Vision Transformer (PVT) [16]

IV. Applications and Use Cases

A. *Applications and use cases of vision transformers*

Vision transformers have demonstrated significant potential for a variety of computer vision tasks and applications, including image classification[17], [18], object detection[19], and semantic segmentation[20],[21], Anomaly detection [22]. In image classification, the vision transformer architecture has been used to process numerous datasets, including ImageNet[12] and

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.

CIFAR-10, and achieved state-of-the-art accuracy results. Vision transformers have also been adopted for object detection, which involves recognizing the location and type of objects inside an image. [23] proposes the DETR (DEtection TRansformer) architecture, which employs a vision transformer to create a fixed number of object queries, which are then used to attend to distinct regions of the image and predict the item's class and location. DETR outperformed classic object detection methods that use region proposal networks on the COCO object detection benchmark while being more computationally efficient, Figure 7 shows the architecture of DETR's transformer. Finally, vision transformers have been used to perform semantic segmentation, which includes labeling each pixel in an image. [24] proposes the Semantic Segmentation with Vision Transformers (SegViT), which use the attention mechanism to generate masks for semantic segmentation. On various semantic segmentation benchmarks, SegViT achieved state-of-the-art results, demonstrating the potential of vision transformers for this task. Overall, vision transformers' versatility and potential have led to their extensive use in a variety of computer vision tasks and applications.

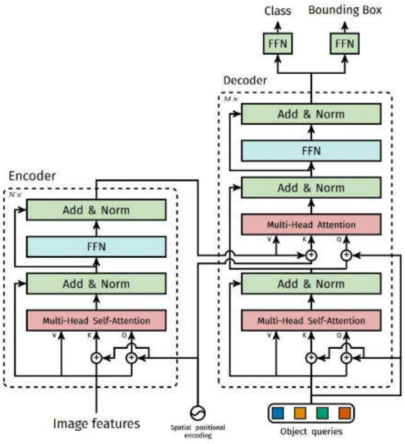

Figure 7: Architecture of DETR's transformer[23]

B. The advantages and the limitations of using ViT

Using vision transformers[8] in various applications can provide multiple benefits, including greater efficiency on large-scale datasets, the capacity to handle long-range dependencies, and improved interpretability. Vision transformers have the ability to recognize complex patterns and relationships in input data, allowing them to excel at object detection, image classification, and semantic segmentation tasks [24], [25], [21]. They can also handle an extensive variety of image sizes and scales, making them suitable for a wide number of applications. In addition, Vision transformers are also more interpretable than other deep learning architectures, allowing researchers to comprehend and observe the model's internal attention mechanism. However, there are several limits to employing vision transformers, such as the model's high computational and memory needs, which make them less practicable for implementation on resource-constrained devices. Due to the enormous number of parameters to train the model, training times and costs can be extended. Finally, the requirement for large amounts of labeled data might be a significant obstacle in some applications because getting labeled data can be time-consuming and costly. Overall, adopting vision transformers can provide multiple benefits in various applications; however, researchers must also examine the constraints and trade-offs involved with using these models.

C. Successful use cases of vision transformers

There have been several successful real-world applications of vision transformers, proving its promise for various computer vision applications. One such use is in medical imaging, where vision transformers have been used to accurately identify and diagnose diseases based on medical images, such as breast cancer and COVID-19[26]. [27] employed a vision transformer architecture to accurately classify benign and malignant breast tumors on mammograms, outperforming typical CNNs. Another application is in autonomous driving, where vision transformers are utilized for object detection and segmentation tasks to assist vehicles in perceiving and navigating their surroundings. For example, [28] demonstrated the promise of these models for real-time applications by using a vision transformer-based object detection model to detect and track pedestrians, automobiles, and other objects in real-time. Vision transformers have also been utilized in natural language processing and computer vision hybrid applications like image captioning and visual question answering, where the model creates natural language descriptions or answers questions about an image's content. for example, [29] introduced a vision transformer-based image captioning model that provides natural language descriptions of images and obtained outstanding results on various benchmarks. Overall, these successful use cases show the potential of vision transformers for a wide range of real-world applications, as well as their advantages over traditional deep learning architectures.

V. Future Directions and Challenges

A. The current state of research in vision transformers

Although the topic of vision transformers is still in its early stages, it has garnered considerable attention and research interest in recent years. While the original vision transformer architecture achieved outstanding results on a variety of computer vision applications, there is still potential for improvement and additional development. Current vision transformers are computationally expensive and require a substantial amount of memory and processing power, making them hard to deploy on resource-constrained devices. One area for further improvement is model efficiency and scalability. Another area of investigation is the adaptation of vision transformers for specific applications and domains like as 3D object recognition, video analysis, and medical imaging. Furthermore, the interpretability and explainability of vision transformers are important fields of research since understanding these model's attention mechanisms can provide insights into how they make decisions and assist establish trust in their outputs. Finally, stronger benchmark datasets and assessment metrics for vision transformers are needed to let researchers compare and evaluate different models and push the state-of-the-art further. In general, the current status of vision transformer research is encouraging, and there are many significant areas for future development that can assist increase performance, efficiency, and interpretability.

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.

B. The challenges of ViT

Although vision transformers have shown potential in providing outstanding performance on a range of computer vision tasks, their adoption faces a number of difficulties. One of the main difficulties is that the self-attention mechanism used in this model takes a large amount of memory and processing power. This can make deploying vision transformers on resource-constrained devices like mobile phones and embedded systems difficult. Furthermore, training vision transformers necessitates a huge amount of data, which might be difficult to get in some fields, such as medical imaging, where there may be little labeled data available. Furthermore, vision transformers can be dependent on the quality and diversity of the training data, and if the training dataset is biased or limited, the model could fail to generalize well to unseen data. Another problem is the interpretability of vision transformers, as the attention process used in these models can be complex and difficult to interpret, making it difficult to grasp how the model makes decisions. At last, there is a need for more study and development in optimizing vision transformers for specific tasks and domains, such as video analysis and 3D object recognition, because these areas have unique challenges and requirements that may not be addressed by current vision transformer architectures. In general, though ViT has demonstrated the ability to deliver outstanding results on a range of computer vision tasks, there are still significant challenges that must be overcome before they can be extensively used in practical applications.

C. Potential future developments in the field of ViT

There is a lot of opportunity for further advancements and discoveries in the field of vision transformers, which is still in the early stages of development. The creation of more effective and scalable vision transformer architectures, which can be executed on devices with limited resources and employed in practical applications, is one field of research that has prospective. Another field of research focuses on the application of vision transformers to various areas, such as robotics and medical imaging, where the difficulties and requirements differ from those of standard computer vision tasks. Furthermore, there is a requirement for further research and advancement in the interpretability and explainability of vision transformers, which can contribute to the growth of confidence in these models and offer information on their decision-making processes. Finally, combining vision transformers with other machine learning methods like reinforcement learning and generative models has a lot of potential for enabling more advanced and complicated computer vision tasks. Overall, the field of vision transformers is still developing, and there is a lot of opportunity for new innovations and advances that may assist to overcome present restrictions and extend the range of uses for these models.

VI. Conclusion

In conclusion, this paper presents a conceptual and architectural overview of vision transformers, which is a novel class of neural network developed to overcome the drawbacks of convolutional neural networks (CNN) in computer vision tasks. The paper gives insights into the applications and use cases of vision transformers, such as image classification, object identification, and semantic segmentation, as well as later variations of the original architecture proposed in other papers. The paper also identifies significant possibilities for further development and advancements in the field of vision transformers, such as the development of more effective and scalable architectures and the adaptation of vision transformers to different areas. Finally, the paper covers the challenges and limitations of using vision transformers, including their high computational necessities and the requirement for large amounts of training data.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, Jun. 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[5] S. Kumar, F. Odone, N. Noceti, and L. Natale, "Object segmentation using independent motion detection," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Seoul, South Korea: IEEE, Nov. 2015, pp. 94–100. doi: 10.1109/HUMANOIDS.2015.7363537.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[7] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks." arXiv, Dec. 02, 2015. Accessed: May 22, 2022. [Online]. Available: http://arxiv.org/abs/1511.08458

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. Accessed: May 22, 2022. [Online]. Available: http://arxiv.org/abs/2010.11929

[9] A. AL Smadi, A. Mehmood, A. Abugabah, E. Almekhlafi, and A. M. Al-smadi, "Deep convolutional neural network-based system for fish classification," *IJECE*, vol. 12, no. 2, p. 2026, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2026-2039.

[10] S. M. Abas, A. M. Abdulazeez, and D. Q. Zeebaree, "A YOLO and convolutional neural network for the detection and classification of leukocytes in leukemia," *IJEECS*, vol. 25, no. 1, p. 200, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp200-213.

[11] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, Accessed: Jul. 24, 2021. [Online]. Available: http://arxiv.org/abs/1706.03762

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database".

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention." arXiv, Jan. 15, 2021. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2012.12877

[14] Y. Cheng and F. Lu, "Gaze Estimation using Transformer." arXiv, May 30, 2021. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2105.14424

[15] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite Transformer with Long-Short Range Attention." arXiv, Apr. 24, 2020. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2004.11886

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.

[16] W. Wang *et al.*, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions." arXiv, Aug. 11, 2021. Accessed: May 18, 2022. [Online]. Available: http://arxiv.org/abs/2102.12122

[17] J. Y.-Y. Lin, S.-M. Liao, H.-J. Huang, W.-T. Kuo, and O. H.-M. Ou, "Galaxy Morphological Classification with Efficient Vision Transformer." arXiv, Feb. 03, 2022. Accessed: Jun. 30, 2022. [Online]. Available: http://arxiv.org/abs/2110.01024

[18] H. Chen *et al.*, "GasHis-Transformer: A Multi-scale Visual Transformer Approach for Gastric Histopathology Image Classification." arXiv, Feb. 17, 2022. Accessed: Jun. 01, 2022. [Online]. Available: http://arxiv.org/abs/2104.14528

[19] Y. Lee and P. Kang, "AnoViT: Unsupervised Anomaly Detection and Localization With Vision Transformer-Based Encoder-Decoder," *IEEE Access*, vol. 10, pp. 46717–46724, 2022, doi: 10.1109/ACCESS.2022.3171559.

[20] S. Zheng *et al.*, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers." arXiv, Jul. 25, 2021. Accessed: May 18, 2022. [Online]. Available: http://arxiv.org/abs/2012.15840

[21] J. Fu *et al.*, "Dual Attention Network for Scene Segmentation." arXiv, Apr. 21, 2019. Accessed: Feb. 24, 2023. [Online]. Available: http://arxiv.org/abs/1809.02983

[22] A. Berroukham, K. Housni, and M. Lahraichi, "Fine-Tuning Pre-trained Vision Transformer Model for Anomaly Detection in Video Sequences," in *Proceedings of the 6th International Conference on Big Data and Internet of Things*, M. Lazaar, E. M. En-Naimi, A. Zouhair, M. Al Achhab, and O. Mahboub, Eds., in Lecture Notes in Networks and Systems, vol. 625. Cham: Springer International Publishing, 2023, pp. 279–289. doi: 10.1007/978-3-031-28387-1_24.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers." arXiv, May 28, 2020. Accessed: May 18, 2022. [Online]. Available: http://arxiv.org/abs/2005.12872

[24] B. Zhang *et al.*, "SegViT: Semantic Segmentation with Plain Vision Transformers." arXiv, Dec. 12, 2022. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2210.05844

[25] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation." arXiv, Sep. 02, 2021. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2105.05633

[26] F. Mehboob *et al.*, "Towards robust diagnosis of COVID-19 using vision self-attention transformer," *Sci Rep*, vol. 12, no. 1, p. 8922, May 2022, doi: 10.1038/s41598-022-13039-x.

[27] G. Ayana *et al.*, "Vision-Transformer-Based Transfer Learning for Mammogram Classification," *Diagnostics*, vol. 13, no. 2, p. 178, Jan. 2023, doi: 10.3390/diagnostics13020178.

[28] Z. Sun, C. Liu, H. Qu, and G. Xie, "PVformer: Pedestrian and Vehicle Detection Algorithm Based on Swin Transformer in Rainy Scenes," *Sensors*, vol. 22, no. 15, p. 5667, Jul. 2022, doi: 10.3390/s22155667.

[29] Y. Wang, J. Xu, and Y. Sun, "End-to-End Transformer Based Model for Image Captioning." arXiv, Mar. 29, 2022. Accessed: May 06, 2023. [Online]. Available: http://arxiv.org/abs/2203.15350

Authorized licensed use limited to: TECH UNIVERSITY OF KOREA. Downloaded on April 02,2025 at 05:43:40 UTC from IEEE Xplore. Restrictions apply.