Journal of Hunan University

(Natural Sciences)

Vol. 52 No. 2

February 2025

Available online at

https://jonuns.com

Open Access Article

https://doi.org/10.55463/issn.1674-2974.52.2.11

Comparative Performance Analysis of Transformer and Convolutional Networks for Machine Vision-Oriented Mobile Robots

Robinson Jiménez-Moreno, Anny Astrid Espitia-Cubillos*

(Associated Professors of Engineering Faculty, Universidad Militar Nueva Granada, Bogotá, Colombia)

* Corresponding author: anny.espitia@unimilitar.edu.co

Article History:

Received: January 21, 2025
Reviewed: February 17, 2025
Revised: February 27, 2025
Accepted: March 17, 2025
Published: March 31, 2025

Abstract: This study compares the performance of three deep neural network architectures with the goal of searching for robotic navigation by semantic means of recognizing the environment in which it is located. The first is a pre-designed vision transformer network, the second is also a pre-designed convolutional neural network, and the third is a custom-designed convolutional network. These architectures are oriented to machine vision for mobile robots, enabling the recognition of global environments. The novelty exposed in this development consists in being able to identify a place based on its environment, as a human being does, so that a robot can address a place described by name and not by spatial coordinates, as usual. Comparison metrics include the level of recognition accuracy of the network, its size in kilobytes, and identification time. In addition to being able to operate in real time, each network is intended to be at least 90% accurate as an initial design parameter. The proprietary CCN network proved to be the most suitable for use in a mobile robot because it has a size of 22.5 KB, a response time of 0.07 seconds and an accuracy of 95.8%.

Keywords: Convolutional networks, transformer networks, deep learning, pre-trained network architecture, mobile robotics, transfer learning.

面向机器视觉的移动机器人的变压器与卷积网络的比较性能分析

摘要：本研究比较了三种深度神经网络架构的性能，目的是通过语义识别机器人所处环境的方式寻找机器人导航。第一种是预先设计的视觉变换网络，第二种也是预先设计的卷积神经网络，第三种是定制设计的卷积网络。这些架构面向移动机器人的机器视觉，能够识别全球环境。这一发展中展现的新颖之处在于能够像人类一样根据环境识别一个地方，这样机器人就可以像往常一样通过名称而不是空间坐标来定位一个地方。比较指标包括网络的识别准确度、网络大小（以千字节为单位）和识别时间。除了能够实时运行外，每个网络的初始设计参数都旨在至少达到 90% 的准确率。专有 CCN 网络被证明最适合用于移动机器人，因为它的大小为 22.5 KB，响应时间为 0.07 秒，准确率为 95.8%。

关键词：卷积网络；变压器网络；深度学习；预训练网络架构；移动机器人；迁移学习

1. Introduction

Since the rise of convolutional networks (CNN) and the success of architectures such as AlexNet [1], their use has become widespread in image processing [2] and machine vision systems [3]. However, since the publication of transform networks [4], their architecture has ventured into different fields of application [5], where in this case its use in computer vision systems stands out [6].

The applications of transform networks with images range from the biomedical field [7], medicine [8]-[10], and image segmentation [11], among many others. With applications such as the detection of visual relationships between instances, as well as the trajectories of subjects and objects [12], traffic sign recognition [13], gait recognition [14], eye disease detection [15], glacier identification [16], fingerprint authentication systems [17], and others based on object detection [18].

Therefore, the use of transformer networks in vision systems has been optimized [19] considering the requirements and use of specialized hardware [20]. Moreover, they improved their performance with combinations with convolutional networks [21]-[22] and variant architectures based on CNN such as Yolo [23], ResNet [24]-[25] and VGG [26].

However, given its recent boom, there has been no research and development of transformer networks based on robotic systems, where convolutional networks have shown great success [27]-[28]. Therefore, this study evaluates machine vision models based on transformer networks and convolutional networks in the recognition of global environments instead of specific objects. For this case, a residential environment is taken as a reference, in which the networks must identify four environments corresponding to the living room, the dining room, the bedroom and the bathroom.

The evaluation metrics are based on the level of accuracy of the network recognition, the size in bytes of the network and the identification time, thinking in mobile robotics applications for residential assistants, which navigate in space based not on coordinates but on the recognition of the site.

The article is structured in four sections. The first section corresponds to the state of the art and contextualization of the work developed. The second section corresponds to the methodology used. The third section presents the results obtained and the fourth section presents the conclusions.

2. Methodology

To evaluate and compare the performance of three network architectures based on deep learning between transformer networks and convolutional networks, two pre-trained network architectures reported in the state of the art for pattern recognition in images, such as vision transformer (ViT) [29] and AlexNet [1], are established and used. In turn, a CNN network architecture is designed to generate a comparative under a customized network with the objective of training based on the database, i.e., to recognize not point objects but global environments based on the identification of sets of elements.

For this purpose, the established procedure can be seen in the flowchart in Figure 1. After establishing the architecture of each network, the database to be used, the selection of training parameters and a selection of comparison metrics must be prepared. As we seek to evaluate the learning of a structured environment within a residential environment useful for the navigation machine vision system on a mobile robot, the metrics chosen are network size, to reduce the hardware requirements in memory and classification time, which is vital for the operation of mobile moving and can operate in real time.

For the implementation of the network architectures, including preprocessing, training, and metrics calculation, an Intel Core i9 Laptop with 16 Gigabytes of RAM, 1TeraByte SSD, Windows 11 NVIDIA, GeForce RTX 4080 12 Gigabyte processor was used.

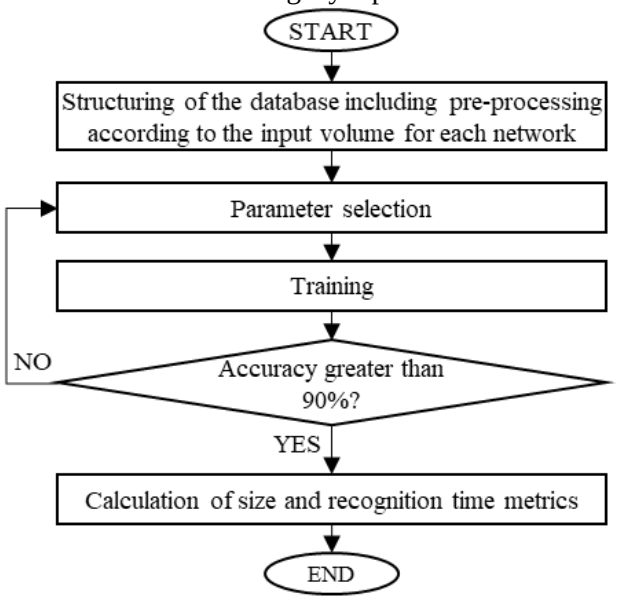

Figure 1. Network testing methodology flowchart (Source: developed by the authors)

In this case, the database used consisted of 400 training images divided into four classes of global environments: bedroom, bathroom, living room and bedroom. Figure 2 shows a sample of the database in the four environments of the chosen residential environment.

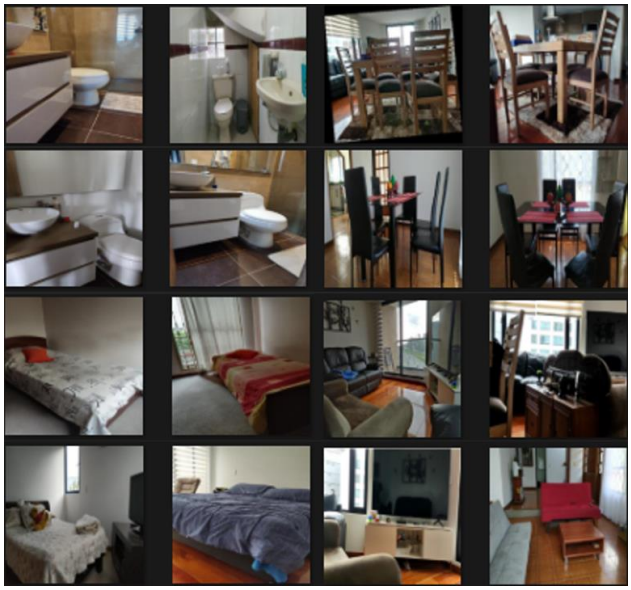

Figure 2. Example of the training database (Source: developed by the authors)

Since the ViT and Alexnet networks are predefined and are used by means of learning transfer, the input volumes are already determined, while for the CNN architecture designed, a volume close to that of the transformer network is chosen, which allows greater pixel recognition, resulting in higher image quality. The input volumes corresponding to the dimensions of the image database for each network, with their characterization, are shown in Table 1.

Table 1. Input volume for each network (Source: developed by the authors)

Net	Network Type	Architecture	Input volume
1	Transformer	ViT	384x384x3
2	CNN	AlexNet	227x227x3
3	CNN	Own	350x350x3

3. Results

Initially, the same parameters were established for all the networks. Given that a level of accuracy greater than 90% was sought, the iteration yielded the final value of epochs and the learning rate required. The options of the parameters selected for each training are shown in Table 2.

Table 2. Training options by network architecture (Source: developed by the authors)

Net	Optimizer	Number of epochs	Learning rate	Mini Batch size
1	SGDM	20	0.0001	2
2	SGDM	20	0.00001	2
3	SGDM	100	0.00001	2

Based on the predefined architectures chosen and the network architecture developed, the following results were obtained after their respective training.

Net 1.

The ViT network architecture has 88.6 million learning parameters with 143 layers. Figure 3 shows the learning curve, where it is evident that the required training time is 24 min and 11 s, for a total of 20 epochs, where learning converges in 10 epochs for an accuracy of 100%.

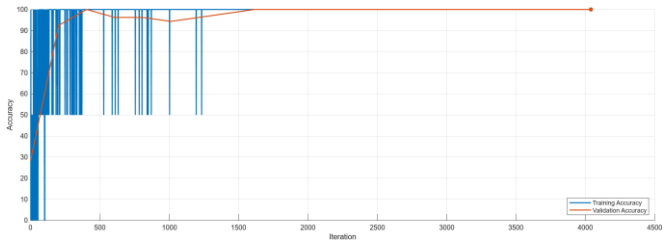

Figure 3. Transformer network training (Source: developed by the authors)

Figure 4 illustrates the confusion matrix of the transformer network where it is evident that there is no confusion between classes. At the bottom, an image with the class prediction label is observed, which clearly corresponds to the environment presented. This image, which can be seen similarly within the samples in Figure

2, corresponds to a reflection used as a data augmentation technique in the test database.

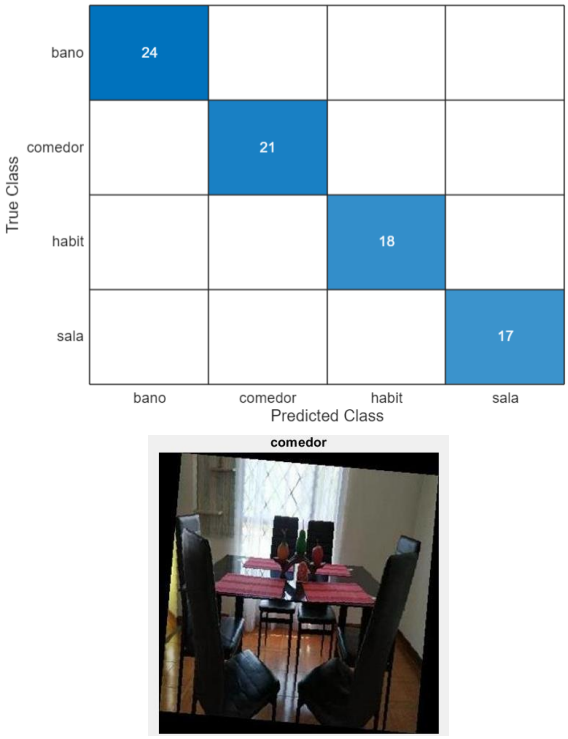

Figure 4. Confusion matrix and example of transformer network prediction (Source: developed by the authors)

Net 2.

The Alexnet network architecture has 56.8 million learning parameters with 24 layers. Figure 5 shows the learning performance, where it is evident that the required training time is 1 min and 19 s, for a total of 20 epochs, where learning converges near epoch 19 with an accuracy of 98.08%.

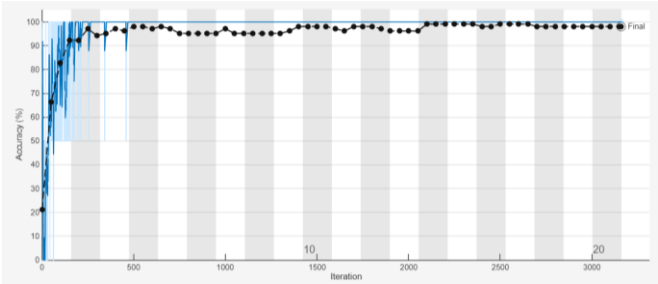

Figure 5. Alexnet training (Source: developed by the authors)

Figure 6 illustrates the AlexNet confusion matrix where it is evident that there is only one class confusion between room and room, which is attributed to an image of the room entrance that does not exhibit the whole environment. The same validation image is observed for each network with the class prediction label, which clearly concerns the corresponding environment.

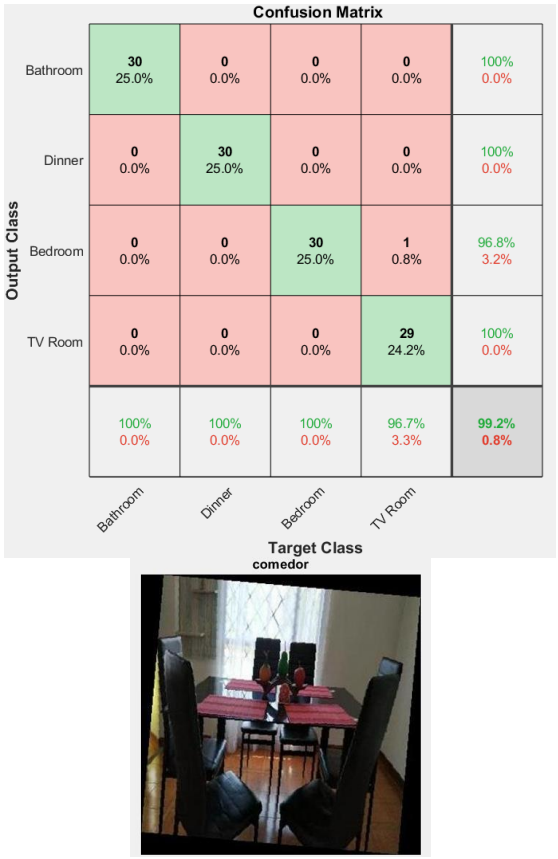

Figure 6. Confusion matrix and example of AlexNet prediction (Source: developed by the authors)

Net 3.

The architecture of the own CNN network has 6.2 million learning parameters, with 36 layers. Figure 7 shows the learning performance, where it is evident that the required training time is 15 min and 48 s, for a total of 100 epochs, where learning converges near epoch 65 with an accuracy of 95.8%.

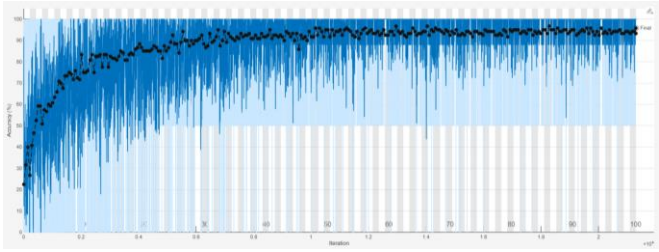

Figure 7. Own CNN network training (Source: developed by the authors)

Figure 8 illustrates the designed CNN confusion matrix where several class confusions are evidenced mainly between the bathroom and room. Below is the same validation image used for each network with the class prediction label, which clearly corresponds to the provided environment.

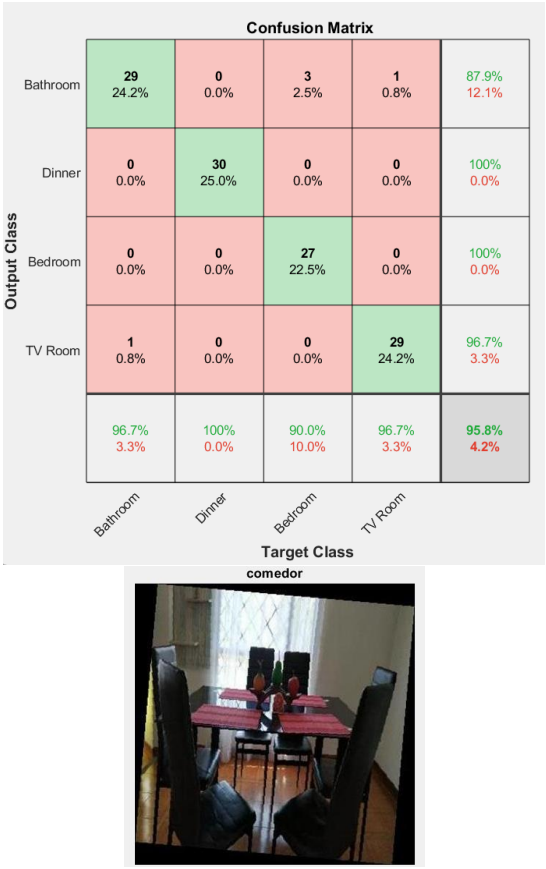

Figure 8. Confusion matrix and example of designed CNN prediction (Source: developed by the authors)

Table 3 summarizes the results presented by adding the digital size in kilobytes of each network and the approximate classification time per image. It is evident that the designed architecture, although with more layers of depth, has fewer filters and therefore fewer learning parameters than ALexNet and Transformer, which makes it digitally lighter and ideal to be embedded in a portable system. At the same time, the classification time is very little longer than AlexNet and small enough to operate in a mobile system in real time.

Table 3. Metrics of each network (Source: developed by the authors)

Net	1	2	3
Size (KB)	313.497	206.787	22.533
Time (seconds)	0.445300	0.059073	0.069771
Learning parameters (millions)	88.6	56.8	6.2
Layers	143	24	36
Training time	24 min 11 seg	1 min 19 seg	15 min 48 seg
Initial epochs	20	20	100
Convergence epochs	10	19	65
Accuracy	100%.	98,08%.	95,8%.

In [30], a DAG-CNN network was used for the semantic identification of residential spaces with an accuracy of 98.3%, a training time of 32.4 min, a size of 36.3KB, 9.9 M of total parameters, 60 layers and an average classification time of 0.074 s. Being a two-branch convolutional network architecture, it increases the number of parameters and the size of the network with no additional benefit to a 2.5% increase in accuracy. While this is still comparable to the learning transfer architectures evaluated, it does not significantly improve the use of one branch. The decrease in the classification time and digital weight obtained in the present study is preferable to that reported in [30]. Figure 9 illustrates the comparative in learning activations of the two DAG-CNN networks (left) and the proposed CNN (right), where the relevance of the activations in the toilet is exposed but in the proposed network is stronger activation than that reported in [30].

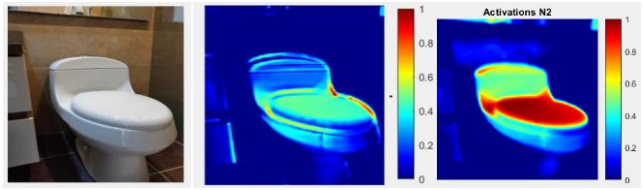

Figure 9. Learning Activation DAG-CNN and designed CNN (N2) (Source: developed by the authors)

4. Conclusion

The main findings of the present study show the rapid convergence in learning the transformer network and its high level of accuracy, which, although in the case of an embedded network for a mobile robot it does not have a favorable size or time, it is ideal for pattern detection in images with more robust computing systems. Compared to previous studies, the designed CNN is faster, smaller and lighter, optimal for mobile robot applications, and oriented to navigate under place identification. The strengths and limitations obtained show that, as mentioned in the state of the art, there are several robust architectures derived from CNNs; however, simple self-designed architectures offer reduced sorting times and lower digital weight, which are optimal for robotic navigation systems, but require effort in their design and parameter selection. Although learning transfer facilitates and speeds up the design and implementation processes, its results do not necessarily turn out to be the most favorable in all scenarios. In the tests performed, it is observed that the self-designed CNN network is the most convenient for the case of a mobile robotic system in terms of network size, which is almost 14 times smaller than the ViT network and more than 9 times smaller than that of the Alexnet network.

Declarations

Author Contributions

Conceptualization, formal analysis and writing—review and editing J.-M.R., and E.-C.A.; methodology, validation, investigation supervision, project administration, funding acquisition, and data curation, J.-M.R.; writing—original draft preparation and visualization, E.-C.A. All authors have read and agreed to the published version of the manuscript.

Funding

The product was derived from the research project titled "Diseño de un modelo de interacción humano robot mediante algoritmos de aprendizaje profundo" INV-ING-3971 financed by the vice-rector for research of the Universidad Militar Nueva Granada, year 2024.

Acknowledgements

The authors thank the Universidad Militar Nueva Granada for the time and resources available for the development of this article.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Universidad Militar Nueva Granada (project INV-ING-3971, date of approval: 18/01/2024).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and falsification, double publication and submission, and redundancies have been completely observed by the authors.

References

[1] KRIZHEVSKY A., SUTSKEVER I., and HINTON G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, 60(6): 84–90. https://doi.org/10.1145/3065386. 2017

[2] LVYANG Y., JIANKANG Z., HUAIQIANG L., LONGFEI R., CHEN Y., JINGYU W., and DONGYUAN S. A comprehensive end-to-end computer vision framework for restoration and recognition of low-quality engineering drawings. *Engineering Applications of Artificial Intelligence*, 2024, 133(Part E): 108524. https://doi.org/10.1016/j.engappai.2024.108524

[3] TAO R., PENG R., JIN Y., GONG F. and LI B., Automatic Detection of Asphalt Pavement Crack Width Based on Machine Vision. *IEEE Transactions on Intelligent Transportation Systems*, 2025, 26(1): 484-496. https://doi.org/10.1109/TITS.2024.3492731

[4] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., and POLOSUKHIN I. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. New York, United States of America, 2017, 6000–6010. Curran Associates Inc., Red Hook. https://doi.org/10.48550/arXiv.1706.03762

[5] BERROUKHAM A., HOUSNI K., and LAHRAICHI M., Vision Transformers: A Review of Architecture, Applications, and Future Directions. *Proceedings of the 7th IEEE Congress on Information Science and Technology (CiSt)*, Agadir - Essaouira, Morocco, 2023, 205-210. https://doi.org/10.1109/CiSt56084.2023.10410015

[6] WU C., and HE T. A Survey of Applications of Vision Transformer and its Variants. *Proceedings of the 10th IEEE International Conference on Intelligent Data and Security (IDS)*, New York, United States of America, 2024, 21-25. https://doi.org/10.1109/IDS62739.2024.00011

[7] TEH S., SIVAKUMAR S., and MOTALEBI F. Vision Transformers for Biomedical Applications. *Proceedings of the 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, Miri Sarawak, Malaysia, 2024, 195-201. https://doi.org/10.1109/GECOST60902.2024.10474871

[8] BALDEON-CALISTO M., RIVERA-VELASTEGUI F., LAI-YUEN S. K., RIOFRÍO D., PÉREZ-PÉREZ N., BENÍTEZ D., and FLORES-MOYANO R. DistilIQA: Distilling Vision Transformers for no-reference perceptual CT image quality assessment. *Computers in Biology and Medicine*, 2024, 177: 108670. https://doi.org/10.1016/j.compbiomed.2024.108670

[9] AL FAHIM M., RAMANARAYANAN S., RAHUL G.S., GAYATHRI M. N., SARKAR A., RAM K., and M. SIVAPRAKASAM. OCUC Former: An Over-Complete Under-Complete Transformer Network for accelerated MRI reconstruction. *Image and Vision Computing*, 2024, 150: 105228. https://doi.org/10.1016/j.imavis.2024.105228

[10] LI J., CHEN N., ZHOU H., LAI T., DONG H., FENG C., CHEN R., YANG C., CAI F., and WEI L. MCRformer: Morphological constraint reticular transformer for 3D medical image segmentation. *Expert Systems with Applications*, 2023, 232: 120877. https://doi.org/10.1016/j.eswa.2023.120877

[11] GONG Z., CHANMEAN M. and GU W. Multi-Scale Hybrid Attention Integrated with Vision Transformers for Enhanced Image Segmentation. *Proceedings of the 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV)*, Zhenjiang, China, 2024: 180-184. https://doi.org/10.1109/AIPMV62663.2024.10691911

[12] QU M., DENG G., DI D., CUI J., and SU T. Dual attentional transformer for video visual relation prediction. *Neurocomputing*, 2023, 550: 126372. https://doi.org/10.1016/j.neucom.2023.126372

[13] FARZIPOUR A., MANZARI O. N. and SHOKOUHI S. B. Traffic Sign Recognition Using Local Vision Transformer. *Proceedings of the 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Islamic Republic of Iran, 2023, 191-196. https://doi.org/10.1109/ICCKE60553.2023.10326288

[14] LI K. and MENG S. TransGait: Vision Transformer Based Gait Recognition Network. *Proceedings of the 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Chengdu, China, 2023, 339-343. https://doi.org/10.1109/ICICML60161.2023.10424880

[15] PURBA S. O., KHAIRINA N., MUHATHIR, MULIONO R. and LUBIS A. H. Classification of Eye Diseases in Humans Using Vision Transformer Architecture

Model. *Proceedings of the 2024 International Conference on Information Technology Research and Innovation (ICITRI)*, Jakarta, Indonesia, 2024: 71-75, https://doi.org/10.1109/ICITRI62858.2024.10699068
[16] NADACHOWSKI P., ŁUBNIEWSKI Z. and TĘGOWSKI J. Glacial Landform Classification with Vision Transformer and Digital Elevation Model. *Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium*, Athens, Greece, 2024, 7254-7258. https://doi.org/10.1109/igarss53475.2024.10641509
[17] GURUNATHAN V., SUDHAKAR R., SATHIYAPRIYA T., and SOUNDAPPAN, J. Finger Vein Authentication Using Vision Transformer. *Proceedings of the 2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, Coimbatore, India, 2024: 1-5. https://doi.org/10.1109/ICSTEM61137.2024.10560933
[18] SURYA S. S., KALAISELVI S., BERNICE T. and GUNASEKRAN V. Cursor Movement Based on Object Detection Using Vision Transformers. *Proceedings of the 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, Vellore, India, 2023: 1-5. https://doi.org/10.1109/ViTECoN58111.2023.10157042
[19] IBRAHIMOVIC E. Optimizing Vision Transformer Performance with Customizable Parameters, *Proceedings of the 46th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, 2023: 1721-1726, https://doi.org/10.23919/MIPRO57284.2023.10159761
[20] PULLAKANDAM M., SONI S., GUPTA S., REDDY YANAMALA R. M. and THOTA G. K. Vision Transformer Implementation on Edge GPU (AGX Orin) for Image Classification. *Proceedings of the First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)*, Delhi, India, 2024: 551-556, https://doi.org/10.1109/IC2SDT62152.2024.10696661
[21] WANG Y., QIAN X. and ZHOU W. Transformer-Prompted Network: Efficient Audio–Visual Segmentation via Transformer and Prompt Learning. *IEEE Signal Processing Letters*, 2025, 32: 516-520. https://doi.org/10.1109/LSP.2024.3524120
[22] GAO Y., SHI S., SUN Z. and LING C. The combination of transformer and CNN in computer vision. *Proceedings of the IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, Dali, China, 2022: 321-325. https://doi.org/10.1109/ICCASIT55263.2022.9987025
[23] WU P., WENG H., LUO W., ZHAN Y., XIONG L., ZHANG H., and YAN H. An improved Yolov5s based on transformer backbone network for detection and classification of bronchoalveolar lavage cells. *Computational and Structural Biotechnology Journal*, 2023, 21: 2985-3001. https://doi.org/10.1016/j.csbj.2023.05.008
[24] DEEPA P. L., PONRAI D. N., and SREENA V. G. A Hybrid Vision Transformer model using ResNet152 for Brain Tumor Classification. *Proceedings of the 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)*, Rourkela, India, 2024: 1-5. https://doi.org/10.1109/ICSPCRE62303.2024.10675105
[25] ÇELEBI A., IMAK A., ÜZEN H., BUDAK Ü., TÜRKOĞLU M., HANBAY D., and ŞENGÜR A. Maxillary sinus detection on cone beam computed tomography images using ResNet and Swin Transformer-based UNet. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 2024, 138(1): 149-161, https://doi.org/10.1016/j.oooo.2023.06.001
[26] HU Z., WANG Z., JIN Y., and HOU W. VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction. *Computer Methods and Programs in Biomedicine*, 2023, 229: 107291. https://doi.org/10.1016/j.cmpb.2022.107291
[27] BALAPAN A., YERALKHAN R., ARYSLANOV A., KALIMULDINA G. and YESHMUKHAMETOV A. A Novel Pattern Recognition Method for Self-Powered TENG Sensor Embedded to the Robotic Hand, *IEEE Access*, 2023, 11: 1-11. https://doi.org/10.1109/ACCESS.2025.3530465
[28] ALHARTHI A. S., TOKATLI O., LOPEZ E. and HERRMANN G. Toward Semi-Autonomous Robotic Arm Manipulation Operator Intention Detection from Force Data. *IEEE Access*, 2025, 13: 664-680, https://doi.org/10.1109/ACCESS.2024.3523325
[29] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., and DEHGHANI M. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. https://doi.org/10.48550/arXiv.2010.11929
[30] JIMÉNEZ R., CASTILLO R. and JARAMILLO J. Machine Vision System for Robotic Navigation in a Residential Environment. *Journal of Intelligent & Fuzzy Systems*, 2024, 47(5-6): 427–437. https://doi.org/10.3233/JIFS-238028

参考文:

[1] KRIZHEVSKY A.、SUTSKEVER I. 和 HINTON G. E. 使用深度卷积神经网络进行 ImageNet 分类。美国计算机协会通讯，2017，60(6)：84–90。https://doi.org/10.1145/3065386。2017
[2] LVYANG Y.、JIANKANG Z.、HUAIQIANG L.、LONGFEI R.、CHEN Y.、JINGYU W. 和 DONGYUAN S. 用于低质量工程图修复和识别的综合端到端计算机视觉框架。人工智能工程应用，2024，133（E 部分）：108524。https://doi.org/10.1016/j.engappai.2024.108524
[3] TAO R.、PENG R.、JIN Y.、GONG F. 和 LI B.，基于机器视觉的沥青路面裂缝宽度自动检测。IEEE 智能交通系统学报，2025，26(1)：484-496。https://doi.org/10.1109/TITS.2024.3492731
[4] VASWANI A.、SHAZEER N.、PARMAR N.、USZKOREIT J.、JONES L.、GOMEZ A. N.、KAISER L. 和 POLOSUKHIN I. 注意力就是你所需要的一切。第 31 届神经信息处理系统国际会议 (NIPS'17) 论文集。美国纽约，2017，6000–6010。Curran Associates Inc.，Red Hook。https://doi.org/10.48550/arXiv.1706.03762
[5] BERROUKHAM A.、HOUSNI K. 和 LAHRAICHI M.，视觉变形金刚：架构、应用和未来方向的回顾。第 7 届

IEEE 信息科学与技术大会 (CiSt) 论文集，摩洛哥阿加迪尔 - 索维拉，2023，205-210。https://doi.org/10.1109/CiSt56084.2023.10410015

[6] WU C. 和 HE T. 视觉变形金刚及其变体的应用调查。第 10 届 IEEE 智能数据和安全国际会议 (IDS) 论文集，美国纽约，2024，21-25 页。https://doi.org/10.1109/IDS62739.2024.00011

[7] TEH S.、SIVAKUMAR S. 和 MOTALEBI F. 用于生物医学应用的视觉变压器。2024 年绿色能源、计算和可持续技术国际会议 (GECOST) 论文集，马来西亚砂拉越美里，2024，195-201 页。https://doi.org/10.1109/GECOST60902.2024.10474871

[8] BALDEON-CALISTO M.、RIVERA-VELASTEGUI F.、LAI-YUEN S.K.、RIOFRÍO D.、PÉREZ-PÉREZ N.、BENÍTEZ D. 和 FLORES-MOYANO R. DistilIQA：用于无参考感知 CT 图像质量评估的蒸馏视觉变压器。生物和医学计算机，2024，177：108670。https://doi.org/10.1016/j.compbiomed.2024.108670

[9] AL FAHIM M.、RAMANARAYANAN S.、RAHUL G.S.、GAYATHRI M. N.、SARKAR A.、RAM K. 和 M. SIVAPRAKASAM。OCUC Former：用于加速 MRI 重建的过完备欠完备变压器网络。图像与视觉计算，2024，150：105228。https://doi.org/10.1016/j.imavis.2024.105228

[10] LI J.、CHEN N.、ZHOU H.、LAI T.、DONG H.、FENG C.、CHEN R.、YANG C.、CAI F. 和 WEI L. MCRformer：用于 3D 医学图像分割的形态约束网状变换器。应用专家系统，2023，232：120877。https://doi.org/10.1016/j.eswa.2023.120877

[11] GONG Z.、CHANMEAN M. 和 GU W. 多尺度混合注意力与视觉变换器集成用于增强图像分割。第二届算法、图像处理和机器视觉国际会议 (AIPMV) 论文集，中国镇江，2024：180-184。https://doi.org/10.1109/AIPMV62663.2024.10691911

[12] QU M.、DENG G.、DI D.、CUI J. 和 SU T. 用于视频视觉关系预测的双注意变换器。神经计算，2023，550：126372。https://doi.org/10.1016/j.neucom.2023.126372

[13] FARZIPOUR A.、MANZARI O. N. 和 SHOKOUHI S. B. 使用局部视觉变换器进行交通标志识别。第 13 届国际计算机与知识工程会议 (ICCKE) 论文集，伊朗伊斯兰共和国马什哈德，2023，191-196。https://doi.org/10.1109/ICCKE60553.2023.10326288

[14] LI K. 和 MENG S. TransGait：基于视觉变换器的步态识别网络。2023 年国际图像处理、计算机视觉和机器学习会议 (ICICML) 论文集，中国成都，2023，339-343。https://doi.org/10.1109/ICICML60161.2023.10424880

[15] PURBA S. O.、KHAIRINA N.、MUHATHIR、MULIONO R. 和 LUBIS A. H. 使用视觉变换器架构模型对人类眼部疾病进行分类。2024 年国际信息技术研究与创新会议 (ICITRI) 论文集，印度尼西亚雅加达，2024：71-75，https://doi.org/10.1109/ICITRI62858.2024.10699068

[16] NADACHOWSKI P.、ŁUBNIEWSKI Z. 和 TĘGOWSKI J. 使用视觉变换器和数字高程模型对冰川地貌进行分类。2024 年 IEEE 国际地球科学与遥感研讨会论文集，希腊雅典，2024，7254-7258。https://doi.org/10.1109/igarss53475.2024.10641509

[17] GURUNATHAN V.、SUDHAKAR R.、SATHIYAPRIYA T. 和 SOUNDAPPAN, J. 使用视觉变换器进行手指静脉身份验证。2024 年国际科学技术工程与管理会议 (ICSTEM) 论文集，印度哥印拜陀，2024：1-5。https://doi.org/10.1109/ICSTEM61137.2024.10560933

[18] SURYA S. S.、KALAISELVI S.、BERNICE T. 和 GUNASEKRAN V. 基于使用视觉变换器进行物体检测的光标移动。第二届面向通信和网络技术新兴趋势的视觉国际会议 (ViTECoN) 论文集，印度韦洛尔，2023：1-5。https://doi.org/10.1109/ViTECoN58111.2023.10157042

[19] IBRAHIMOVIC E. 使用可定制参数优化 Vision Transformer 性能，第 46 届 MIPRO ICT 和电子大会 (MIPRO) 论文集，克罗地亚奥帕蒂亚，2023：1721-1726，https://doi.org/10.23919/MIPRO57284.2023.10159761

[20] PULLAKANDAM M.、SONI S.、GUPTA S.、REDDY YANAMALA R. M. 和 THOTA G. K. 在边缘 GPU（AGX Orin）上实现 Vision Transformer 用于图像分类。第一届计算机科学与数字技术先锋发展国际会议 (IC2SDT) 论文集，印度德里，2024：551-556，https://doi.org/10.1109/IC2SDT62152.2024.10696661

[21] WANG Y.、QIAN X. 和 ZHOU W. Transformer 提示网络：通过 Transformer 和提示学习实现高效的音频-视频分割。IEEE 信号处理快报，2025，32：516-520。https://doi.org/10.1109/LSP.2024.3524120

[22] GAO Y.、SHI S.、SUN Z. 和 LING C. Transformer 与 CNN 在计算机视觉中的结合。IEEE 第四届民航安全与信息技术国际会议论文集（ICCASIT），中国大理，2022：321-325。https://doi.org/10.1109/ICCASIT55263.2022.9987025

[23] WU P.、WENG H.、LUO W.、ZHAN Y.、XIONG L.、ZHANG H. 和 YAN H. 一种基于 transformer 主干网络的改进 Yolov5s，用于检测和分类支气管肺泡灌洗细胞。计算与结构生物技术杂志，2023，21：2985-3001。https://doi.org/10.1016/j.csbj.2023.05.008

[24] DEEPA P. L.、PONRAI D. N. 和 SREENA V. G. 一种使用 ResNet152 进行脑肿瘤分类的混合视觉变换器模型。2024 年 IEEE 智能电源控制和可再生能源国际会议 (ICSPCRE) 论文集，印度 Rourkela，2024：1-5。https://doi.org/10.1109/ICSPCRE62303.2024.10675105

[25] ÇELEBI A.、IMAK A.、ÜZEN H.、BUDAK Ü.、TÜRKOĞLU M.、HANBAY D. 和 ŞENGÜR A. 使用

ResNet 和基于 Swin Transformer 的 UNet 对锥形束计算机断层扫描图像进行上颌窦检测。口腔外科、口腔医学、口腔病理学和口腔放射学，2024，138(1): 149-161，https://doi.org/10.1016/j.oooo.2023.06.001
[26] HU Z.、WANG Z.、JIN Y. 和 HOU W. VGG-TSwinformer：基于 Transformer 的深度学习模型，用于早期阿尔茨海默病预测。生物医学中的计算机方法和程序，2023，229：107291。https://doi.org/10.1016/j.cmpb.2022.107291
[27] BALAPAN A.、YERALKHAN R.、ARYSLANOV A.、KALIMULDINA G. 和 YESHMUKHAMETOV A. 一种用于嵌入机械手的自供电 TENG 传感器的新型模式识别方法，IEEE 访问，2023，11：1-11。https://doi.org/10.1109/ACCESS.2025.3530465
[28] ALHARTHI A. S.、TOKATLI O.、LOPEZ E. 和 HERRMANN G. 从力数据实现半自主机械臂操作操作员意图检测。 IEEE Access，2025，13：664-680，https://doi.org/10.1109/ACCESS.2024.3523325
[29] DOSOVITSKIY A.、BEYER L.、KOLESNIKOV A.、WEISSENBORN D.、ZHAI X.、UNTERTHINER T. 和 DEHGHANI M. 一张图片胜过 16x16 个字：用于大规模图像识别的变压器。2021。https://doi.org/10.48550/arXiv.2010.11929
[30] JIMÉNEZ R.、CASTILLO R. 和 JARAMILLO J. 用于住宅环境中机器人导航的机器视觉系统。《智能与模糊系统杂志》，2024，47(5-6): 427–437。https://doi.org/10.3233/JIFS-238028

Word count: 2,597 words, excluding references.

Peer review information:

Whether the manuscript was fast tracked? - No
Number of reviewer report submitted in first round: 3 reports
Number of revision rounds: 2 rounds
Final revised version submitted: March 17, 2025