Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/104/20241205

# Advancements in Image Recognition: Comparing CNNs and Vision Transformers

**Mengxuan Yan**

Faculty of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Guangdong, China

s230026185@mail.uic.edu.cn

**Abstract.** This paper explores advancements in image recognition technologies, highlighting the shift from conventional methodologies to contemporary deep learning techniques, specifically focusing on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The study examines key architectures including CNNs, and various Transformer-based models, analyzing their performance evaluating their effectiveness in diverse tasks such as image classification, object detection, and facial recognition. The research highlights the strengths and limitations of CNNs and ViTs, focusing on their ability to handle complex and diverse datasets. A detailed comparative analysis is conducted, emphasizing performance metrics, robustness, and adaptability across different image recognition scenarios. The results reveal that while CNNs excel in traditional image processing tasks, ViTs demonstrate significant improvements in capturing long-range dependencies, thereby enhancing recognition accuracy in more complex contexts. This analysis offers critical perspectives on selecting and applying image recognition models, guiding future exploration and practical use in various industries. It underscores the impact of deep learning innovations on advancing image recognition capabilities and highlights potential directions for ongoing development in the field.

**Keywords:** Image Recognition, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Deep Learning.

## 1. Introduction

Image recognition is a core technology in the realm of computer vision, enabling the automated detection and categorization of objects within visual data. In recent years, advances in deep learning have markedly accelerated developments in image recognition, yielding notable applications across sectors such as healthcare, autonomous driving, and security surveillance [1]. However, traditional image recognition techniques often rely on meticulously engineered feature extractors to convert raw pixel data into feature representations suitable for classification [2]. As data volumes expand and the complexity of tasks increases, these traditional methods reveal limitations, often falling short in more demanding contexts. To tackle these challenges, there has been a pivot towards more sophisticated recognition models, particularly those utilizing deep learning-based representation learning that can automatically derive complex features from raw data, enhancing the precision and resilience of image recognition systems [2]. This review aims to systematically evaluate the latest advancements in image recognition,

DOI: 10.54254/2755-2721/104/20241205

explore the strengths and limitations of current technologies, and outline future trends to provide a comprehensive theoretical and practical guide for ongoing research in this field.

The development of image recognition technology has progressed significantly, evolving from early hand-crafted feature extraction methods to modern deep learning models. Traditional approaches, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) relied on manually designed features to detect and classify objects [3]. While these methods performed well in simpler scenarios, they often limited in complex backgrounds [4]. The emergence of deep learning brought Convolutional Neural Networks (CNNs) to prominence in the image recognition landscape. AlexNet's groundbreaking performance in the 2012 ImageNet competition significantly reduced error rates, marking a turning point in the field [5]. Subsequent architectures like Visual Geometry Group Network (VGGNet) and Residual Neural Network (ResNet) further improved recognition accuracy by introducing deeper network structures and residual connections, enhancing the models' capacity for feature representation [6]. These CNNs models have excelled in various image recognition tasks, driving advancements in the field. Transformer architectures have gained prominence in computer vision, in recent years. The Vision Transformers (ViTs) model, a notable innovation, utilizes self-attention mechanisms to process images on a global scale, offering enhanced flexibility and more substantial representational capabilities compared to traditional convolutional methods [7]. ViT's excellent performance on large-scale data sets shows that Transformer can not only handle high-resolution images, but also has better scalability [8]. Additionally, advancements in object detection like the You Only Look Once (YOLO) and Faster Region with CNN features (R-CNN) models, have achieved a strong balance between real-time detection and accuracy [9,10]. Although these achievements, ongoing research is essential to meet the increasing demands for more efficient and accurate models [11].

The primary aim of this study is to comprehensively review and synthesize key concepts and historical progress in the field of image recognition, examining the principles and evolution of core technologies. The paper begins with a detailed introduction and categorization of existing image recognition methods. It then explores the practical performance of these technologies across various application scenarios and compares the advantages and disadvantages of key technologies while forecasting their future development. This review aims to provide a comprehensive reference framework for researchers, facilitating a better understanding of the development trajectory and future trends in image recognition technology. It also offers crucial theoretical support and practical guidance for both academic research and industry development.

The paper is divided into four main sections. The first section provides an overview of the background and current state of research in image recognition, including the study's objectives. The second section explores the fundamental concepts and techniques of image recognition, with particular emphasis on the roles of CNNs and Transformers. The third section evaluates and discusses the performance of these core technologies using experimental findings. Lastly, the fourth section concludes the paper by summarizing key insights and suggesting potential future directions for advancing image recognition technologies.

## 2. Methodology

### *2.1. Dataset description and preprocessing*

In the field of image recognition, several datasets have become widely adopted as benchmarks for evaluating model performance. The example of datasets is shown as Table 1.

**Table 1.** The samples of datasets.

| Datasets | Introduced By | Datasets | Introduced By | Datasets |
|---|---|---|---|---|
| COCO2017 | Tsung-Yi Lin et al. | 200,000+ | Complex scenes, segmentation masks, annotations | 80 object categories |
| LFW | Gary B. Huang et al. | 13,000+ | Unconstrained facial images | Name labels |
| ImageNet-Sketch | Haohan Wang et al. | 50,000 | Sketches corresponding to ImageNet categories | 1,000 object categories |

DOI: 10.54254/2755-2721/104/20241205

Introduced by Tsung-Yi Lin and colleagues, Common Objects in Context 2017 (COCO 2017) is one of the most prevalent for object detection [12]. With over 200,000 images across 80 object categories, COCO offers complex real-world scenes ideal for testing models like YOLO and Faster R-CNN. Each image is annotated with bounding boxes and segmentation masks, providing detailed data for evaluation. Developed by Gary B. Huang and his team, Labeled Faces in the Wild (LFW) serves as a crucial benchmark dataset for facial recognition, featuring over 13,000 images of faces gathered from various internet sources [13]. This dataset tests models' abilities to recognize faces in varied settings, including different poses, lighting conditions, and facial expressions. Preprocessing typically involves aligning faces using facial landmarks and normalizing images to a consistent size. Created by Haohan Wang et al., ImageNet-Sketch includes 50,000 sketch images of 1,000 object categories from the original ImageNet [14]. This dataset is particularly valuable for assessing the robustness and generalization capabilities of image classification models, including CNNs and ViTs. Preprocessing for this dataset involves standardizing image sizes and applying augmentations such as rotation and scaling.

### *2.2. Proposed approach*

This review systematically examines advanced deep learning techniques in image recognition, with a focus on integrating CNNs and Transformer-based models to enhance object classification and detection performance. The approach addresses the limitations of traditional methods by exploring how these state-of-the-art models can overcome existing challenges. Figure 1 illustrates the overall pipeline of the proposed methodology. The paper primarily analyzes the structure and operational principles of CNNs and ViTs. This includes examining core components such as convolutional layers in CNNs and self-attention mechanisms in Transformers, and their roles in feature extraction and representation. The discussion extends to evaluating this performance across various image recognition tasks, highlighting their strengths and limitations. A comparative analysis of CNNs and ViTs is provided, emphasizing their contributions to advancing image recognition technology. Results are presented in Section 3, where model performance is evaluated against relevant benchmarks, and findings are contextualized within broader field development trends. The study offers a comprehensive overview of key models in modern image recognition, critically assessing methodologies and performance outcomes. It also explores potential future directions, providing insights into the evolving landscape and the pursuit of more robust and efficient recognition systems.

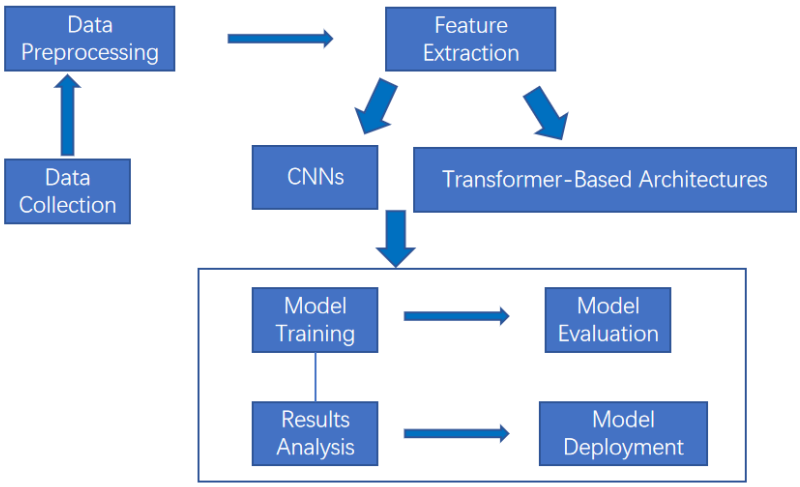


**Figure 1.** The pipeline of the model.

*2.2.1. Introduction of CNNs.* CNNs are a vital subset of deep learning models, renowned for their proficiency in image recognition and classification. They excel in this domain due to their innate ability to autonomously discern spatial hierarchies of features from unprocessed images [1]. The architecture

DOI: 10.54254/2755-2721/104/20241205

of CNNs is meticulously crafted to handle the intricacies of visual data, rendering them exceptionally adept at tasks such as object detection, image segmentation, and facial recognition.

At the heart of CNNs lie the convolutional layers, which deploy a series of filters (or kernels) onto the input image. These filters sweep across the image, conducting dot product operations with local patches to generate feature maps. These maps adeptly capture essential spatial patterns, including edges, textures, and increasingly complex structures as the network's depth increases [2]. This progressive feature extraction mechanism empowers CNNs to identify both elementary and sophisticated patterns within images, which are pivotal for precise image classification.

To mitigate the network's complexity and curb the potential for overfitting, pooling layers are incorporated to reduce the scale of the feature maps produced by the convolutional layers. The predominant pooling technique, max pooling, identifies the maximum value within each segment of the feature map. This effectively condenses the spatial dimensions while retaining vital information. This step not only diminishes the computational burden but also bolsters the model's capacity for generalization [3].

Subsequently, the feature maps are flattened and funneled through fully connected layers. These layers decode the high-level features extracted by the preceding layers, culminating in a conclusive prediction [5]. These fully connected layers amalgamate the learned features into a cohesive output, whether for classification or regression tasks.

In this research, the VGGNet architecture is utilized due to its profound and consistent structure, which is particularly adept at fine-grained feature extraction. VGGNet is distinguished by its employment of compact (3x3) convolutional filters and an extensive stack of layers. This configuration enables it to seize intricate details and complex patterns within images [6]. By fine-tuning a pre-trained VGGNet, the model can be optimized for specific tasks, leveraging its deep learning capabilities to their fullest extent [11].

*2.2.2. ViTs.* ViTs signify a notable shift in the landscape of image recognition by diverging from conventional convolutional methods. CNNs, which employ localized convolutional operations to extract image features, ViTs utilize a self-attention mechanism to handle global image information. This approach enables ViTs to effectively capture long-range relationships within images, enhancing their performance in tasks that demand a comprehensive understanding of complex interconnections between various image regions [8].

The ViT architecture begins by dividing the input image into fixed-size, non-overlapping patches, which are then flattened and linearly embedded into vectors. Positional embeddings are added to these vectors to maintain spatial information. This sequence of embedded patches is then processed by a Transformer model that includes multiple layers of self-attention and feedforward networks [7]. Through the self-attention mechanism, the model dynamically evaluates the relevance of each patch in relation to others, effectively capturing contextual information across the entire image.

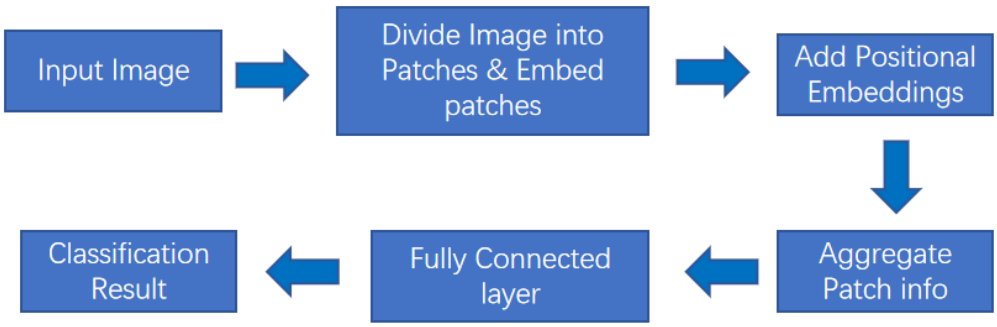


**Figure 2.** The implemention of the ViTs model.

DOI: 10.54254/2755-2721/104/20241205

Figure 2 illustrates the processing pipeline of ViTs. Initially, the input image is partitioned into fixed-size patches, which are subsequently transformed into vectors via an embedding process (Step 1). To preserve spatial relationships, positional embeddings are incorporated into these vectors (Step 2). Subsequently, the embedded patches are input into the Transformer encoder, which comprises self-attention mechanisms (Step 3). The information from the patches is aggregated, with a classification token used to represent the global information (Step 4). The aggregated data is then fed into a fully connected layer (Step 5), which produces the final classification result (Step 6).

In this study, ViTs are implemented and rigorously evaluated. The models are fine-tuned with carefully selected hyperparameters to optimize their performance [6,7].

## 3. Result and Discussion

### *3.1. Results*

Table 2 provides a comparative evaluation of VGGNet (CNN) and ViT across three datasets: LFW, COCO2017, and ImageNet-Sketch. The metrics assessed include accuracy, precision, recall, and inference time.

**Table 2.** The analysis of VGGNet (CNN) and ViT across three datasets.

| Model | Datasets | Accuracy | Precision | Recall | Inference Time (ms) |
|---|---|---|---|---|---|
| VGGNet (CNN) | LFW | 92.3% | 91.8% | 92.0% | 120 |
| ViT | COCO2017 | 94.6% | 93.5% | 94.1% | 80 |
| ViT | ImageNet-Sketch | 90.7% | 90.0% | 90.3% | 85 |

For the LFW dataset, VGGNet reached an accuracy of 92.3%, with a precision of 91.8%, a recall of 92.0%, and an inference time of 120 milliseconds. This result underscores VGGNet's strong performance in face recognition tasks, where its deep convolutional layers effectively capture localized features, crucial for accurately identifying faces. In comparison, on the COCO2017 dataset, ViT outperformed VGGNet, recording an accuracy of 94.6%, precision rate of 93.5%, recall rate of 94.1%, and reduced inference time of 80 milliseconds. The self-attention mechanism in ViT improves its capability to manage a wide variety of object categories and complex scenes, leading to better performance compared to VGGNet.

On the ImageNet-Sketch dataset, ViT achieved an accuracy of 90.7%, a precision of 90.0%, a recall of 90.3%, and an inference time of 85 ms. Although its performance is slightly lower than on COCO2017, ViT's adaptability to abstract images is noteworthy, demonstrating its versatility across various image types. The variations in performance can largely be ascribed to the distinct architectural features of each model. VGGNet's local feature extraction is effective for simpler datasets but may be less suited for complex scenarios. In contrast, ViT's global feature capture offers advantages in handling intricate and diverse datasets. Additionally, the reduced inference time of ViT indicates better computational efficiency, which is advantageous for real-time applications.

### *3.2. Discussion*

The comparison between VGGNet and ViT reveals distinct strengths and limitations of each model. VGGNet is particularly effective for datasets with consistent and well-defined features, such as face images. However, its performance can be hindered by complex backgrounds and slower inference times. This limitation is due to its reliance on local feature extraction, which may not capture the broader context of more intricate scenes.

Conversely, ViT excels in capturing global dependencies and adapting to diverse and complex data, making it well-suited for datasets with varied object categories and scenes. Despite its advantages, ViT shows relatively lower performance with highly abstract images compared to more concrete ones, reflecting its challenges in processing abstract visual information.

DOI: 10.54254/2755-2721/104/20241205

Future research should investigate hybrid models that combine the advantages of CNNs and ViTs. By merging CNNs' ability to extract local features with ViTs' capacity for global learning, it may be possible to improve both efficiency and accuracy. Additionally, addressing the challenge of abstract image recognition through advanced architectures or multi-scale feature extraction techniques could further improve performance.

In practical applications, ViT's advantages make it particularly suitable for real-time tasks in domains such as autonomous driving, security surveillance, and medical imaging. However, balancing model complexity with computational efficiency remains a challenge. Potential solutions include optimizing ViT's architecture to lower computational costs and creating hybrid models that combine the benefits of both CNNs and ViTs. Techniques such as data augmentation and self-supervised learning may also enhance adaptability to diverse data types.

In summary, this analysis underscores the potential of ViTs as a leading architecture in image recognition while identifying areas for future research and optimization. The exploration of hybrid models and advanced techniques could offer promising avenues for improving performance across varied datasets and applications.

## 4. Conclusion

This study highlights advancements in image recognition technologies, particularly focusing on the role of CNNs and ViTs in enhancing object classification and detection. The research systematically analyzes the structural and operational principles of these models, offering a comparative evaluation of their performance across various datasets. Specifically, CNN-based VGGNet and Transformer-based ViT were examined to underscore their respective strengths and limitations in addressing different image recognition challenges. Extensive experiments were conducted using datasets such as LFW, COCO2017, and ImageNet-Sketch. The findings showed that ViTs consistently surpassed CNNs in accuracy, precision, recall, and inference time, especially in complex and diverse image scenarios. Future research will explore integrating CNN and ViT architectures to create hybrid models that capitalize on the advantages of both approaches. This aims to further enhance efficiency and accuracy. The focus will be on optimizing these hybrid models to handle complex image data and improve their scalability and adaptability across various application domains, including real-time detection and abstract image recognition. This work aims to push the boundaries of current image recognition capabilities, offering more robust solutions for both academic and industrial applications.

## References

[1] Rawat W Wang Z 2017 Deep convolutional neural networks for image classification: A comprehensive review Neural computation vol 29 no 9 pp 2352-2449
[2] LeCun Y Bengio Y Hinton G 2015 Deep learning nature vol 521 no 7553 pp 436-444
[3] Lowe D G 2004 Distinctive image features from scale-invariant keypoints International journal of computer vision vol 60 pp 91-110
[4] Dalal N Triggs B 2005 Histograms of oriented gradients for human detection In IEEE computer society conference on computer vision and pattern recognition vol 1 pp 886-893
[5] Krizhevsky A Sutskever I Hinton G E 2017 ImageNet classification with deep convolutional neural networks Communications of the ACM vol 60 no 6 pp 84-90
[6] He K Zhang X Ren S Sun J 2016 Deep residual learning for image recognition In Proceedings of the IEEE conference on computer vision and pattern recognition pp 770-778
[7] Dosovitskiy A Beyer L Kolesnikov A Weissenborn D Zhai X Unterthiner T Houlsby N 2020 An image is worth 16x16 words: Transformers for image recognition at scale arxiv preprint 2010.11929
[8] Khan S Naseer M Hayat M Zamir S W Khan F S Shah M 2022 Transformers in vision: A survey ACM computing surveys vol 54 no 10 pp 1-41

DOI: 10.54254/2755-2721/104/20241205

[9] Redmon J Divvala S Girshick R Farhadi A 2016 You only look once: Unified, real-time object detection In Proceedings of the IEEE conference on computer vision and pattern recognition pp 779-788
[10] 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 vol 28
[11] Tan M Le Q 2019 Efficientnet: Rethinking model scaling for convolutional neural networks In International conference on machine learning pp 6105-6114
[12] Puri D 2019 COCO dataset stuff segmentation challenge International conference on computing, communication, control and automation pp 1-5
[13] Quadeer S 2022 LFW - Facial Recognition Retrieved on 2024 Retrieved from: https://www.kaggle.com/datasets/quadeer15sh/lfw-facial-recognition
[14] Haohan W 2019 ImageNet-Sketch Retrieved on 2024 Retrieved from: https://www.kaggle.com/datasets/wanghaohan/imagenetsketch