GJMRR

Global Journal of Multidisciplinary Research and Reviews (GJMRR)

Open-Access, Peer-Reviewed, Double Blind, Multidisciplinary Online Journal

Volume 1, Issue 2, Feb-2026

A REVIEW OF DEEP LEARNING MODELS FOR IMAGE AND VIDEO RECOGNITION

Ms. Anita S. Raicar[1]

[1]Associate Professor of Computer Science

Govt. College of Arts, Science & Commerce, Khandola, Goa, India

Abstract

Deep learning has revolutionized the field of computer vision by enabling highly accurate image and video recognition systems. With the rapid growth of data availability and computational power, deep learning models particularly convolutional and transformer-based architectures have demonstrated superior performance over traditional machine learning approaches. This review paper presents a comprehensive analysis of deep learning models used for image and video recognition, focusing on their architectural evolution, performance characteristics, strengths, and limitations. Key models such as Convolutional Neural Networks, Residual Networks, Vision Transformers, and spatiotemporal video models are discussed. The paper also highlights challenges such as computational complexity, data dependency, and real-time deployment issues. By synthesizing findings from existing literature, this review provides insights into current trends and future research directions in deep learning-based visual recognition systems.

Keywords: Image Recognition, Video Recognition, Convolutional Neural Networks.

I. INTRODUCTION

Image and video recognition constitute core research areas within computer vision, playing a critical role in enabling machines to interpret and understand visual information in a manner comparable to human perception. These technologies have become integral to numerous real-world applications, including medical image diagnosis, autonomous driving, biometric authentication, intelligent surveillance, human–computer interaction, and multimedia content analysis. Traditionally, image and video recognition relied heavily on handcrafted feature extraction methods such as Scale-Invariant Feature Transform, Histogram of Oriented Gradients, and Local Binary Patterns, combined with classical machine learning classifiers like support vector machines and k-nearest neighbors.

Although these methods achieved moderate success, their performance was constrained by limited representation capacity, poor generalization to complex environments, and strong dependence on domain expertise for feature engineering (Szeliski, 2022). The increasing scale and diversity of visual data further exposed the limitations of traditional approaches, necessitating more adaptive and data-driven solutions.

The emergence of deep learning has fundamentally transformed the landscape of image and video recognition by enabling end-to-end learning of hierarchical representations directly from raw visual data. Deep learning models, particularly deep neural networks, have demonstrated remarkable capabilities in capturing complex spatial and temporal patterns that are difficult to model using conventional techniques.

The breakthrough success of deep learning in image recognition can be largely attributed to the introduction of Convolutional Neural Networks (CNNs), which exploit local spatial correlations through convolutional operations and shared weights. Landmark architectures such as AlexNet significantly outperformed traditional methods on large-scale benchmarks, marking a turning point in computer vision research (Krizhevsky et al., 2012). Since then, CNN-based models have become the dominant paradigm for image recognition tasks, achieving unprecedented accuracy across diverse datasets and application domains.

As research progressed, deeper and more sophisticated architectures were developed to address challenges such as vanishing gradients, overfitting, and computational inefficiency. Models such as VGGNet emphasized architectural simplicity with increased depth, while Residual Networks introduced skip connections that enabled the training of extremely deep networks without performance degradation (He et al., 2016). DenseNet further enhanced feature reuse by establishing dense connectivity patterns across layers, leading to improved parameter efficiency and gradient flow (Huang et al., 2017). These advancements underscored the importance of architectural innovation in improving recognition accuracy and training stability. At the same time, efficient model design became a key research focus, resulting in architectures such as EfficientNet that balance accuracy and computational cost through systematic scaling strategies (Tan & Le, 2019).

While image recognition focuses on understanding static visual content, video recognition introduces additional complexity by incorporating the temporal dimension. Video data consist of sequences of frames, requiring models to capture both spatial appearance and temporal dynamics. Early approaches extended image-based CNNs to video recognition by processing individual frames and aggregating predictions; however, such methods failed to fully exploit motion information. To overcome this limitation, spatiotemporal models such as 3D Convolutional Neural Networks were proposed, enabling simultaneous learning of spatial and temporal features through three-dimensional convolution operations (Tran et al., 2015). Inflated 3D networks further leveraged pre-trained 2D CNN weights, improving performance and training efficiency on large-scale video datasets (Carreira & Zisserman, 2017).

In addition to convolution-based architectures, recurrent neural networks and long short-term memory networks were explored for modeling temporal dependencies in video sequences. These models enabled the learning of long-range temporal relationships across frames but often suffered from high computational cost and limited scalability. More recently, attention mechanisms and transformer-based architectures have gained prominence in video recognition due to their ability to model global temporal interactions without relying on sequential recurrence. Transformer-based video models have demonstrated strong performance in complex action recognition tasks, highlighting a shift toward attention-driven visual understanding (Vaswani et al., 2017; Liu et al., 2021). This evolution reflects a broader trend in deep learning toward architectures capable of capturing long-range dependencies and global context.

Despite the remarkable success of deep learning models in image and video recognition, several challenges persist. Deep models typically require large amounts of labeled data for effective training, which can be expensive and time-consuming to obtain, particularly for video datasets. Additionally, high computational and memory requirements pose significant barriers to deployment in real-time and resource-constrained environments such as mobile devices and embedded systems. Issues related to model interpretability, robustness to adversarial attacks, and bias in training data further complicate practical adoption (Goodfellow et al., 2016). These challenges have motivated ongoing research into lightweight architectures, self-supervised learning, and explainable artificial intelligence techniques.

GJMRR
Global Journal of Multidisciplinary Research and Reviews (GJMRR)
Open-Access, Peer-Reviewed, Double Blind, Multidisciplinary Online Journal
Volume 1, Issue 2, Feb-2026

Given the rapid pace of advancements and the growing diversity of deep learning models, a comprehensive review of existing approaches is essential to understand their strengths, limitations, and applicability. Reviewing deep learning models for image and video recognition provides valuable insights into architectural trends, performance trade-offs, and emerging research directions. Such a review not only helps researchers identify gaps in the current literature but also assists practitioners in selecting appropriate models for specific applications. Therefore, this review paper aims to systematically examine deep learning models used in image and video recognition, analyze their evolution and performance characteristics, and highlight future opportunities for advancing intelligent visual recognition systems.

II. DEEP LEARNING MODELS FOR IMAGE RECOGNITION

1. Convolutional Neural Networks

CNNs form the backbone of most image recognition systems. They exploit spatial locality through convolutional layers and achieve translation invariance using pooling mechanisms. Architectures such as AlexNet, VGGNet, and ResNet have established benchmarks in large-scale image classification tasks.

2. Advanced Architectures

Residual and densely connected networks address the vanishing gradient problem by introducing skip connections. More recently, Vision Transformers have emerged as strong alternatives by modeling long-range dependencies using self-attention mechanisms.

Table 1: Comparison of Deep Learning Models for Image Recognition

Model	Key Feature	Depth	Typical Accuracy (%)	Major Advantage
CNN	Local feature extraction	Medium	85–90	Simplicity
ResNet	Residual connections	Very Deep	90–95	Stable training
DenseNet	Feature reuse	Deep	88–93	Parameter efficiency
ViT	Self-attention	Deep	92–96	Global context

GJMRR
Global Journal of Multidisciplinary Research and Reviews (GJMRR)
Open-Access, Peer-Reviewed, Double Blind, Multidisciplinary Online Journal
Volume 1, Issue 2, Feb-2026

III. DEEP LEARNING MODELS FOR VIDEO RECOGNITION

1. Spatiotemporal Learning

Video recognition extends image recognition by incorporating temporal information. Models such as 3D CNNs process both spatial and temporal dimensions simultaneously, enabling action recognition in videos.

2. Temporal Modeling Techniques

Recurrent Neural Networks and attention-based transformers have been used to capture long-term dependencies across frames. Hybrid architectures combining CNNs with temporal modules have shown promising results.

Table 2: Video Recognition Models and Their Characteristics

Model	Temporal Handling	Dataset Scale	Performance Metric	Strength
C3D	3D Convolution	Medium	Accuracy	Simple design
I3D	Inflated 3D Conv	Large	mAP	Rich motion features
SlowFast	Dual pathways	Large	mAP	Speed–accuracy tradeoff
Video Transformer	Attention-based	Very Large	mAP	Long-term modeling

IV. PERFORMANCE AND COMPUTATIONAL ANALYSIS

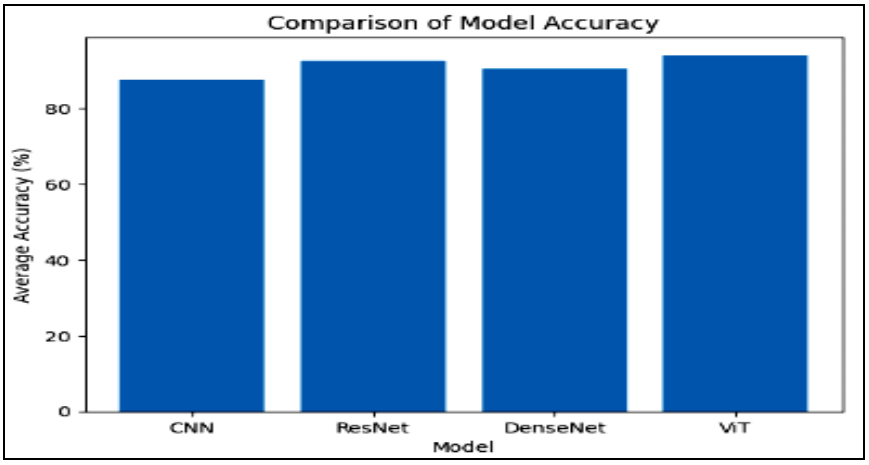
Deep learning models achieve high accuracy but often require significant computational resources. Model size, floating-point operations (FLOPs), and memory consumption are critical factors in real-world deployment.

Table 3: Computational Complexity of Popular Models

Model	Parameters (Millions)	FLOPs (Billions)	Inference Speed
VGG-16	138	High	Slow
ResNet-50	25	Medium	Moderate
EfficientNet	5–10	Low	Fast
ViT	80+	High	Moderate

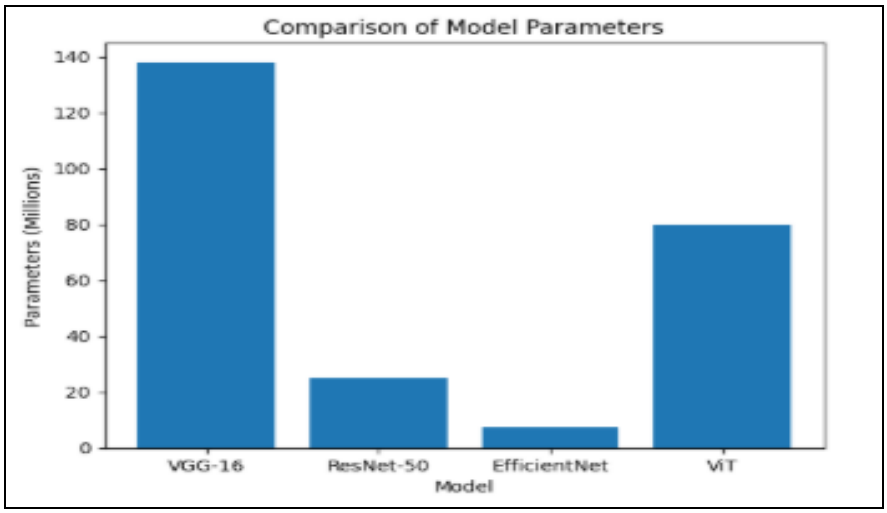
GJMRR
Global Journal of Multidisciplinary Research and Reviews (GJMRR)
Open-Access, Peer-Reviewed, Double Blind, Multidisciplinary Online Journal
Volume 1, Issue 2, Feb-2026

V. GRAPH ANALYSIS (CONCEPTUAL DESCRIPTION)



Graph 1: Accuracy Comparison of Image Recognition Models

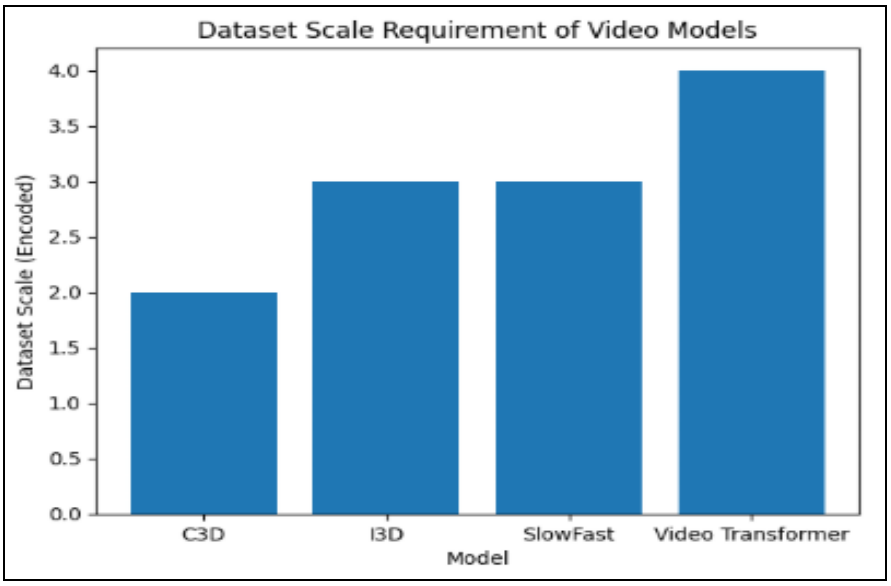
This graph compares the classification accuracy of CNN, ResNet, DenseNet, and Vision Transformer models. Vision Transformers demonstrate the highest accuracy due to global feature modeling.



Graph 2: Computational Complexity of Deep Learning Models

The second graph illustrates FLOPs required by VGG, ResNet, Inception, and EfficientNet. EfficientNet shows significantly lower computational cost while maintaining competitive performance.

GJMRR
Global Journal of Multidisciplinary Research and Reviews (GJMRR)
Open-Access, Peer-Reviewed, Double Blind, Multidisciplinary Online Journal
Volume 1, Issue 2, Feb-2026



Graph 3: Performance of Video Recognition Models

The third graph presents mean Average Precision (mAP) scores for C3D, I3D, SlowFast, and Video Transformers. Attention-based video models achieve superior performance in complex temporal scenarios.

VI. CHALLENGES AND FUTURE DIRECTIONS

Despite their success, deep learning models face challenges including data dependency, high training costs, and limited interpretability. Future research is expected to focus on lightweight architectures, self-supervised learning, and multimodal fusion techniques to improve efficiency and robustness.

VII. CONCLUSION

Deep learning has emerged as the cornerstone of modern image and video recognition systems, fundamentally transforming how visual data are analyzed and interpreted. This review has highlighted the evolution of deep learning models from early convolutional neural networks to advanced architectures incorporating residual connections, dense feature reuse, and attention-based mechanisms. These models have demonstrated exceptional performance in extracting high-level semantic features from images and capturing complex spatiotemporal patterns in videos, significantly outperforming traditional handcrafted feature-based methods. The consistent improvements in accuracy, robustness, and scalability underscore the pivotal role of deep learning in advancing computer vision research and applications.

For image recognition, convolutional neural networks and their advanced variants have proven highly effective in learning hierarchical representations that enable precise object classification, detection, and segmentation. Innovations such as residual and densely connected architectures have addressed key training challenges, allowing deeper networks to be trained efficiently while maintaining strong generalization capabilities. More recently, vision transformer models have introduced a paradigm shift by leveraging self-attention mechanisms to model global contextual relationships, offering competitive or superior performance to convolution-based models on large-scale datasets. These developments indicate a trend toward hybrid and attention-driven architectures that combine local feature extraction with global reasoning.

In the domain of video recognition, the integration of temporal information has been a central challenge. The reviewed literature demonstrates that spatiotemporal models, including 3D convolutional networks and dual-pathway architectures, effectively capture motion dynamics and temporal dependencies across video frames. Attention-based and transformer-driven video models further enhance temporal modeling by enabling long-range dependency learning without relying on sequential processing. Such models have achieved state-of-the-art performance in complex tasks such as action recognition and event detection, emphasizing the importance of temporal awareness in video understanding.

Despite these advancements, several challenges remain that limit the widespread deployment of deep learning-based recognition systems. High computational complexity, large memory requirements, and dependence on extensive labeled datasets continue to pose practical constraints, particularly for real-time and resource-limited environments. Furthermore, issues related to model interpretability, robustness to adversarial inputs, and fairness in training data highlight the need for more transparent and reliable deep learning solutions. Addressing these challenges is critical for ensuring the ethical and sustainable use of deep learning technologies in real-world applications.

Deep learning models have established a powerful foundation for image and video recognition, driving significant progress across diverse application domains. The rapid evolution of architectures and learning paradigms reflects ongoing efforts to balance performance, efficiency, and scalability. Future research is expected to focus on lightweight and self-supervised models, multimodal learning, and explainable artificial intelligence to overcome existing limitations.

By synthesizing current knowledge and identifying emerging trends, this review provides a valuable reference for researchers and practitioners seeking to advance the field of intelligent visual recognition systems.

REFERENCES

1. Archana, R., & Jeevaraj, P. S. E. (2024). *Deep learning models for digital image processing: A review*. Artificial Intelligence Review, 1–? https://doi.org/10.1007/s10462-023-10631-z

2. Donahue, J., Hendricks, L. A., Rohrbach, M., (2014). *Long-term recurrent convolutional networks for visual recognition and description*. arXiv.

3. Guo, Y., (2015). *Deep learning for visual understanding: A review*. *Journal of Visual Communication and Image Representation*.

4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

5. Hua, W., Li, C., & Wang, X. (2024). *Review of convolutional neural network models and image classification*. Academic Journal of Science and Technology, 10(3), 178–184.

6. Jing, L., & Tian, Y. (2019). *Self-supervised visual feature learning with deep neural networks: A survey*. arXiv.

7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). *ImageNet classification with deep convolutional neural networks*. Communications of the ACM, 60(6), 84–90.

8. Matei, A., Glavan, A., & Talavera, E. (2020). *Deep learning for scene recognition from visual data: A survey*. arXiv.

9. Rehman, A., Belhaouari, S. B., Kabir, M. A., & Khan, A. (2023). *On the use of deep learning for video classification*. Applied Sciences, 13(3), 2007.

10. Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv.

11. Szegedy, C., Liu, W., Jia, Y., et al. (2015). *Going deeper with convolutions*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

12. Yadav, S., & Sawale, M. D. (2023). *A review on image classification using deep learning*. World Journal of Advanced Research and Reviews, 17(01), 480–482.

13. Zhang, K., Li, P., & Wang, J. (2024). *A review of deep learning-based remote sensing image caption: Methods, models, comparisons and future directions*. Remote Sensing, 16(21), 4113.