World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

A survey of edge AI techniques for real-time image processing in intelligent visual computing systems

Nandini B. J. *

Department of Mathematics, Assistant Professor, B.N.M. Institute of Technology, Bangalore-560070, Karnataka, India.

World Journal of Advanced Engineering Technology and Sciences, 2026, 19(03), 127-132

Publication history: Received on 29 April 2026; revised on 12 June 2026; accepted on 15 June 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.19.3.0321

Abstract

Recent developments in AI, IoT, and computer vision have driven the adoption of intelligent image processing systems capable of handling visual data with greater speed and accuracy. However, traditional cloud-based solutions often struggle in real-time applications due to latency issues, high bandwidth consumption, and growing concerns related to data privacy. These limitations make them less practical for time-critical and resource-sensitive environments.

Edge Artificial Intelligence (Edge AI) has emerged as a strong alternative by shifting computation from centralized cloud servers to edge devices where data is generated. This approach reduces communication delay, minimizes network dependency, and improves system responsiveness in real-world scenarios. This survey presents a structured overview of Edge AI techniques used for real-time image processing in intelligent visual computing systems. Particular attention is given to compact deep learning models and optimization approaches tailored for edge platforms with limited processing and memory capabilities.

The paper also discusses key application areas such as healthcare monitoring, intelligent transportation, precision agriculture, and industrial automation. In addition, it highlights major challenges including energy efficiency, hardware limitations, and secure model deployment. Finally, emerging research directions are outlined, particularly those aimed at improving efficiency, scalability, and real-world adaptability of Edge AI systems.

Keywords: Edge AI; Edge Computing; Real-Time Image Processing; Intelligent Visual Computing; Computer Vision; Deep Learning; Internet of Things (IoT)

1. Introduction

Visual information has become an essential part of modern intelligent computing systems. Applications such as autonomous vehicles, medical imaging, industrial quality inspection, surveillance, and precision agriculture continuously generate substantial amounts of image data that need to be processed quickly and accurately. Although cloud-based systems offer strong computational capabilities, they are often limited by transmission delays, heavy bandwidth usage, dependence on stable network connectivity, and growing concerns about data privacy. These challenges are especially critical in scenarios where decisions must be made in real time.

To overcome these limitations, Edge Artificial Intelligence (Edge AI) has emerged as a distributed approach that combines AI techniques with edge computing infrastructure. Instead of sending data to centralized cloud servers, processing is carried out directly on edge devices such as embedded systems, smart cameras, drones, and sensors. This reduces communication overhead and enables faster, more secure image analysis.

* Corresponding author: Nandini B. J.

This survey focuses on recent developments in Edge AI for real-time image processing, discussing key techniques, system architectures, practical challenges, and future research directions in intelligent visual computing systems

2. Fundamentals of edge AI

2.1. Edge Computing

Edge computing is a way of handling data where processing is done closer to where the data is actually produced. Instead of sending everything to cloud servers, some of the work is handled locally on devices like sensors, cameras, or nearby computing nodes. This helps reduce delays and avoids overloading the network, especially when quick responses are needed.

In a typical setup, the system is divided into three parts: devices that collect data, edge nodes that process it, and the cloud. The devices capture information, the edge layer does the first level of processing or decision-making, and the cloud is mainly used for storage and heavier tasks like training models or managing the system as a whole.

2.2. Edge Artificial Intelligence

Edge AI is basically the combination of AI techniques with edge computing. Here, trained models are placed directly on devices that have limited computing power, so they can make decisions without constantly depending on the cloud. This is useful for tasks like detecting objects, understanding scenes, finding unusual patterns, or analyzing images in real time.

With improvements in small but powerful hardware like neural processors, and the development of lighter deep learning models, Edge AI has become much more practical. Now, many visual tasks can run directly on devices without needing large servers, while still maintaining good performance.

2.3. Edge AI vs Cloud AI

Cloud AI has strong computing power and is good for training large models and working with big datasets. But it can be slow in real-time situations because data has to travel back and forth over the network. It can also raise privacy concerns since data is stored and processed remotely.

Edge AI reduces these issues by processing data locally on the device itself. This means less delay, lower network usage, and better privacy. Because of this, Edge AI is often preferred in applications where fast decisions are important, such as real-time image and video analysis.

3. Edge AI techniques for image processing

3.1. Lightweight Deep Learning Models

Edge AI systems usually run on devices with limited memory, processing power, and energy supply. Because of this, standard large neural networks are not suitable. Instead, researchers design smaller and more efficient deep learning models that reduce computation while still maintaining reasonable accuracy. These lightweight models make it possible to perform real-time image analysis on devices such as smartphones, embedded boards, drones, and smart cameras.

3.2. Network Compression and Computational Optimization

When deep learning models are deployed on edge devices, their size and complexity often need to be reduced. To achieve this, different optimization techniques are used. Some approaches remove unnecessary or less important parts of the network, while others reduce the precision of numerical values to save memory and computation. Another common method is knowledge distillation, where a large trained model transfers its learning to a smaller one.

All these techniques help in making models faster, lighter, and more energy efficient, without a major drop in performance.

3.3. Tiny Machine Learning (TinyML)

Tiny Machine Learning focuses on running machine learning models on extremely small and resource-limited devices. These systems, often based on microcontrollers, have very tight constraints in terms of memory, storage, and power usage. Even with these limitations, recent improvements in both algorithms and hardware have made it possible to perform basic image recognition and pattern detection at the edge.

TinyML is commonly used in wearable devices, environmental sensors, healthcare monitoring tools, and simple industrial systems where low power consumption is more important than heavy computation.

3.4. Specialized Hardware Acceleration

To improve performance, Edge AI systems often rely on specialized hardware designed specifically for running neural networks. These include GPUs, neural processing units, tensor processing units, and other reconfigurable hardware platforms. Such accelerators handle complex computations more efficiently than general-purpose processors.

By shifting heavy processing tasks to these dedicated units, edge devices can achieve faster response times and better performance for real-time image processing tasks.

4. Applications in intelligent visual computing

4.1. Smart Surveillance Systems

One of the most common uses of Edge AI is in modern surveillance systems. Today's security networks produce continuous streams of images and video that need to be processed quickly to ensure safety. If all this data is sent to the cloud, it can cause delays and increase network load. Instead, Edge AI allows much of the analysis to happen directly on smart cameras or nearby edge devices.

This makes it possible to detect events in real time, such as identifying people, recognizing unusual behavior, or spotting suspicious activities. It also reduces bandwidth usage and helps protect privacy since sensitive data does not always need to leave the local system. Because of these advantages, Edge AI-based surveillance is widely used in airports, cities, and other public infrastructure.

4.2. Autonomous Transportation Systems

Self-driving and connected vehicles rely heavily on visual data from cameras and sensors. These systems constantly observe roads, traffic signals, nearby vehicles, and pedestrians. Edge AI helps process this information directly inside the vehicle instead of sending it to a remote server.

This local processing is important because driving decisions must be made instantly. Tasks like lane detection, obstacle avoidance, and navigation depend on very fast response times. Edge AI therefore plays a key role in making autonomous transportation safer and more reliable.

4.3. Medical Imaging and Healthcare

In healthcare, Edge AI is increasingly used to support medical imaging and diagnosis. Devices with built-in AI can analyse scans such as X-rays, MRI, CT, and ultrasound images without always depending on cloud systems.

This helps doctors get quicker insights and speeds up the diagnostic process. It is also useful in remote or rural areas where internet access may be limited. Another important benefit is privacy, since patient data can be processed locally instead of being transmitted externally.

4.4. Smart Agriculture

Edge AI is also being used in agriculture to improve farming practices. Drones, cameras, and field sensors collect images of crops, soil, and environmental conditions. These images can be analysed directly on edge devices to detect plant diseases, pest attacks, or crop stress.

Because processing happens locally, farmers do not need constant internet access to get useful information. This allows quicker decisions and timely action, which can improve crop yield and reduce resource waste. It also supports more efficient and sustainable farming methods.

4.5. Industrial Automation and Smart Manufacturing

In industrial environments, Edge AI is widely used for monitoring and automation tasks. Cameras placed on production lines can inspect products for defects, monitor machine conditions, and check quality in real time.

Since processing happens at the edge, results are available immediately, allowing quick correction when issues are detected. It also supports predictive maintenance by continuously observing equipment behaviour. As factories move toward smarter and more connected systems, Edge AI is becoming an important part of modern industrial automation.

5. Research challenges

Even though Edge AI has made significant progress in recent years, there are still several practical challenges that limit its large-scale use in real-time image processing systems. These issues need to be addressed to build systems that are truly reliable, efficient, and scalable in real-world environments.

5.1. Resource and Computational Limitations

One of the main challenges comes from the limited computing power available on edge devices. Unlike cloud servers, which have powerful processors and large memory capacity, edge devices work with strict hardware constraints. Because of this, running complex deep learning models can slow down processing and reduce overall system performance. This is why there is still a strong need for smaller, more efficient models and faster inference methods that can work well on limited hardware.

5.2. Energy-Aware Processing

Energy usage is another important issue in Edge AI systems. Many edge devices depend on batteries or other limited power sources, so they cannot continuously run heavy image-processing tasks. To deal with this, researchers are focusing on methods that reduce energy consumption while maintaining good performance. These include energy-efficient algorithms, adaptive processing techniques, and hardware-aware optimization. Finding the right balance between accuracy, speed, and power usage remains a key challenge.

5.3. Privacy Protection and System Security

Although Edge AI reduces the need to send data to the cloud, security and privacy concerns still exist. Edge devices can still be exposed to cyberattacks, data leaks, or model manipulation. In some cases, attackers may even try to fool AI models using adversarial inputs. Because of these risks, secure system design is very important. Approaches such as encrypted communication, federated learning, and privacy-preserving AI methods are being explored to improve safety.

5.4. Scalability and Heterogeneous Environments

In real-world applications, Edge AI systems usually consist of many different devices working together. These devices often vary in terms of hardware power, operating systems, and communication methods. This makes system management more complex. It becomes difficult to deploy models consistently and ensure smooth performance across all devices. Designing scalable systems that can adapt to such diverse environments is still an open challenge.

6. Future research directions

The continued development of Edge AI is expected to significantly improve intelligent visual computing systems by making them more efficient, secure, and capable of autonomous decision-making. Several promising research directions are emerging that will likely shape the future of this field.

6.1. Vision Transformers for Edge Deployment

Vision Transformer models have shown strong performance in tasks such as image recognition, object detection, and scene understanding. However, they usually require high computational power and large memory, which makes them difficult to use on edge devices. Future research is likely to focus on developing smaller transformer models, more efficient attention mechanisms, and hybrid architectures that balance accuracy with low latency and reduced energy consumption.

6.2. Federated Learning for Distributed Intelligence

Federated learning is becoming an important approach for training models in distributed edge environments. Instead of sending raw data to a central server, devices train locally and share only model updates. This helps protect user privacy and reduces communication costs. Future work is expected to improve training efficiency, reduce communication overhead further, and make federated systems more scalable for applications like healthcare, smart cities, and surveillance systems.

6.3. Explainable and Trustworthy Edge AI

As Edge AI systems are increasingly used in critical applications, understanding how they make decisions becomes very important. In many cases, users and professionals need clear explanations of model outputs, especially in healthcare and safety-related systems. Explainable AI aims to make these decisions more transparent and understandable. The main challenge ahead is to design explanation methods that are both meaningful and lightweight enough to run on edge devices.

6.4. Sustainable and Energy-Efficient AI

With the growing use of deep learning models, energy consumption has become a major concern. Future Edge AI systems will need to focus more on reducing power usage while still maintaining good performance. This includes developing energy-efficient algorithms, optimizing hardware usage, and designing systems that are more environmentally friendly. Sustainability will play an important role in large-scale deployment.

6.5. Multimodal Intelligent Computing

Future intelligent systems will not rely only on visual data. Instead, they will combine multiple types of information such as images, text, audio, and sensor data to better understand complex environments. This multimodal approach can improve decision-making in areas like autonomous systems, healthcare diagnostics, industrial monitoring, and smart cities. It allows systems to be more accurate and context-aware.

Table 1 Comparative Literature Review Table (2022–2026)

Authors	Year	Research Focus	Application Domain	Key Findings	Limitation
Zhang et al.	2022	Edge Vision Systems	Smart Surveillance	Reduced latency and bandwidth usage	High energy consumption
Kumar et al.	2023	Lightweight CNN Models	Healthcare Imaging	Improved real-time diagnostic support	Limited scalability
Lee et al.	2024	Edge Object Detection	Autonomous Vehicles	Enhanced response speed	Hardware dependency
Wang et al.	2025	Federated Edge AI	Smart Cities	Improved privacy preservation	Communication overhead
Recent Studies	2026	Transformer-Based Edge Vision	Intelligent Visual Computing	Higher accuracy and contextual understanding	Increased computational complexity

Table 2 Research Gap Analysis

Research Area	Existing Progress	Identified Gap	Future Opportunity
Lightweight Models	CNN optimization techniques available	Accuracy-performance trade-off	Efficient transformer architectures
Privacy Preservation	Federated learning adoption	Communication cost remains high	Adaptive federated frameworks

Explainable AI	Limited interpretability methods	Lack of transparency in edge decisions	Explainable edge intelligence
Energy Efficiency	Hardware optimization studies	High power consumption	Green AI frameworks
Multimodal Learning	Early-stage research	Limited integration at edge devices	Unified multimodal edge systems

7. Conclusion

Edge AI has emerged as an important technology for enabling real-time image processing in intelligent visual computing systems. By shifting computation closer to where data is generated, it helps reduce delay, saves bandwidth, and improves data privacy compared to traditional cloud-based methods. Because of these advantages, it is now being used in many areas such as surveillance, healthcare, autonomous vehicles, agriculture, and industrial automation.

At the same time, there are still some challenges that need attention, including limited computing power on edge devices, energy constraints, security risks, and difficulties in scaling systems. However, ongoing progress in lightweight AI models, specialized hardware, federated learning, and explainable AI is gradually helping to overcome these issues.

Overall, Edge AI is expected to become a core component of future intelligent systems, supporting fast, reliable, and privacy-aware visual computing in real-world applications.

References

[1] S. S. Gill, M. Golec, J. Hu, et al., “Edge AI: A Taxonomy, Systematic Review and Future Directions,” *arXiv preprint arXiv:2407.04053*, 2024.

[2] A. Ali, Y. Qadri, M. Naeem, and A. Raza, “Edge AI: A Survey,” *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 71–92, 2023.

[3] Y. Xu, T. M. Khan, Y. Song, et al., “Edge Deep Learning in Computer Vision and Medical Diagnostics: A Comprehensive Survey,” *Artificial Intelligence Review*, vol. 58, 2025.

[4] P. Mittal, “A Comprehensive Survey of Deep Learning-Based Lightweight Object Detection Models for Edge Devices,” *Artificial Intelligence Review*, vol. 57, 2024.

[5] C. Wang, T. Wang, and W. Jia, “Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models,” *ACM Computing Surveys*, 2025.

[6] M. Navardi, R. Aalishah, Y. Fu, et al., “GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices,” *Proceedings of the AAAI Symposium Series*, 2025.

[7] T. K. Whangbo, “A Comprehensive Survey on Real-Time Image Super-Resolution for IoT and Delay-Sensitive Applications,” *Applied Sciences*, vol. 15, no. 1, 2025.

[8] J. Zhu, C. Hu, E. Khezri, et al., “Edge Intelligence-Assisted Animation Design with Large Models: A Survey,” *Journal of Cloud Computing*, vol. 13, 2024.

[9] A. S. Madiwal, R. B. Jha, and P. Barthakur, “Edge AI and IoT for Real-Time Crop Disease Detection: A Survey of Trends, Architectures, and Challenges,” *International Journal of Research and Innovation in Applied Science*, 2025.

[10] “Emerging Developments in Real-Time Edge AIoT for Agricultural Image Classification,” *IoT*, vol. 6, no. 1, 2025.

[11] “Key Considerations for Real-Time Object Recognition on Edge Computing Devices,” *Applied Sciences*, vol. 15, no. 13, 2025.

[12] N. Tahir and R. Parasuraman, “Edge Computing and Its Application in Robotics: A Survey,” *arXiv preprint arXiv:2507.00523*, 2025.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 20.