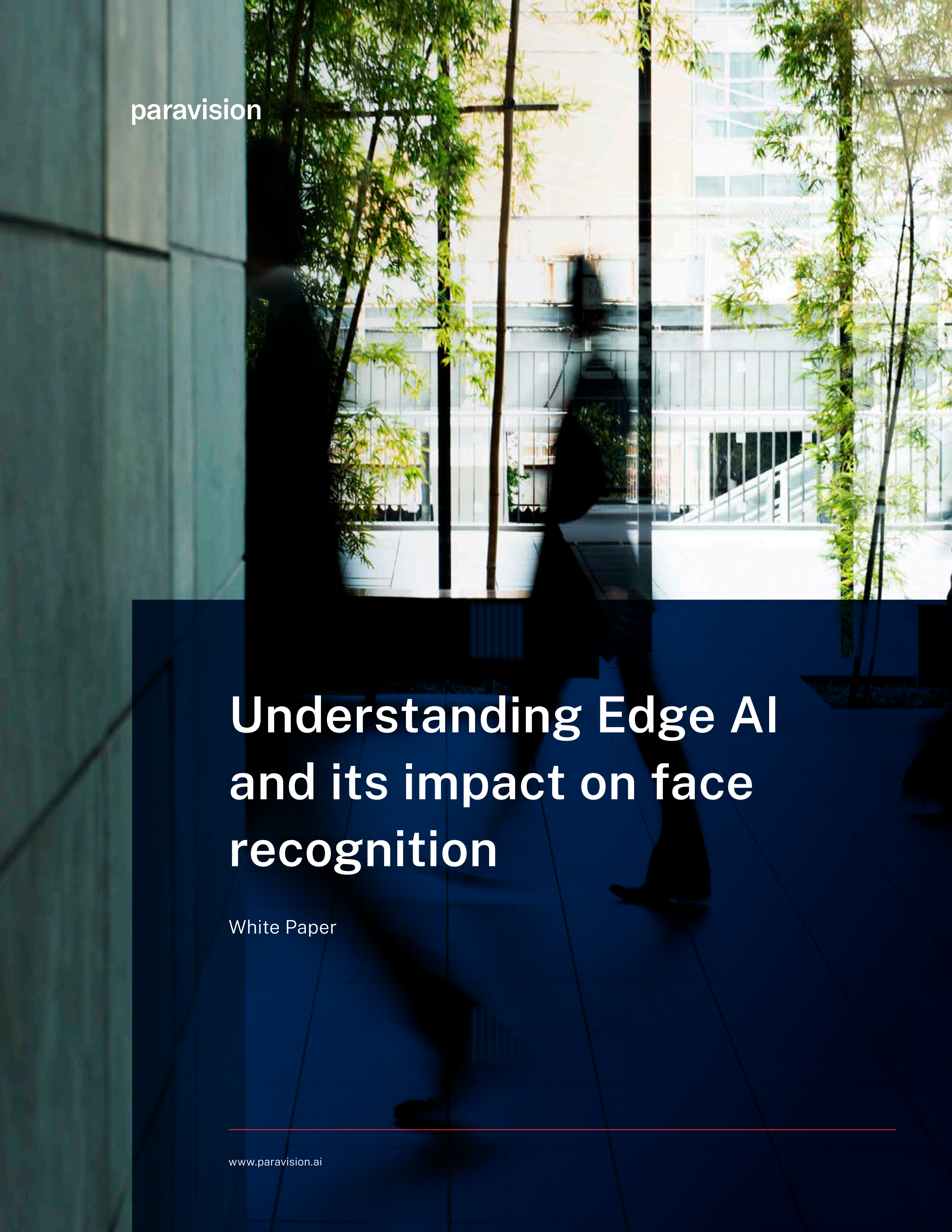
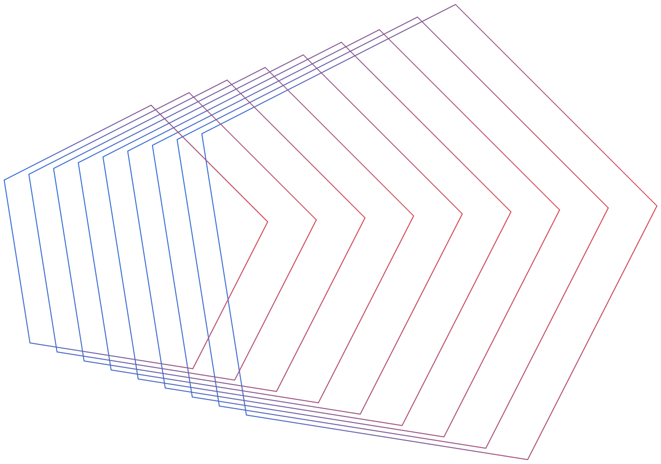


paravision

Understanding Edge AI and its impact on face recognition

White Paper

www.paravision.ai



Introduction

Already redefining the way we work, connect, create and live, edge computing and artificial intelligence have rapidly become indispensable technologies. Edge computing brings data processing to the edge of the network where data is being created, reducing reliance on cloud computing and, by extension, triggering a cascade of compelling benefits for a wide range of applications. Artificial intelligence, meanwhile, promises the ability to make faster, more accurate decisions from highly complex data sources. Now, edge computing and AI have been brought together to form “Edge AI,” harnessing their respective benefits while creating wholly new opportunities as a result of their combination.

These enormous benefits are connecting with the market, resulting in massive predicted growth. According to market research firm IDC, the worldwide edge computing market will reach $250.6 billion in 2024 with a compound annual growth rate (CAGR) of 12.5% over the 2019–2024 forecast period. In addition to professional and provisioned services, hardware will account for $80.7 billion in revenue and software will reach $54 billion in 2024.

Part of the market growth is attributable to the benefits of edge computing, such as the reduction or elimination of delays for data transport, enhanced privacy and more. The other part will be the enablement of applications such as Edge AI that either weren’t possible before or were only partially effective.

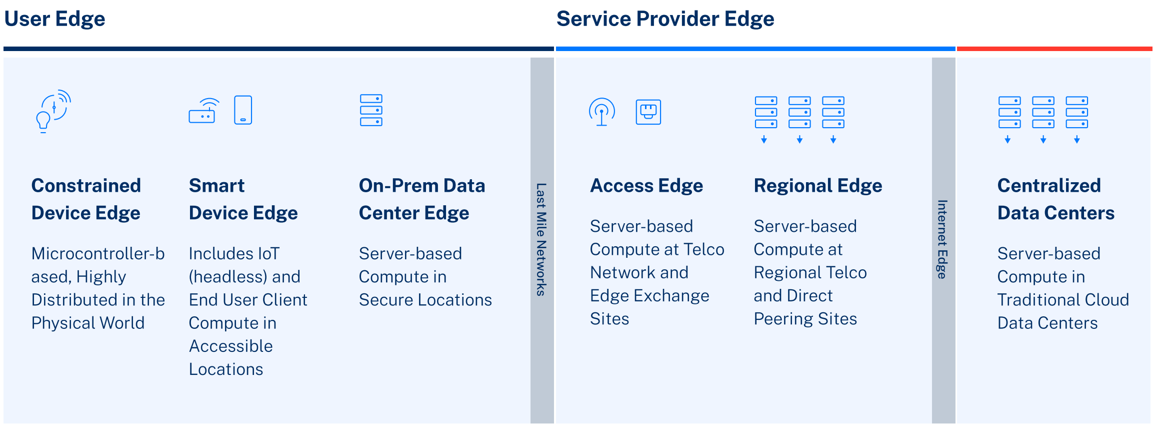
Edge AI itself will have enormous implications for face recognition systems, as it allows for drastically increased efficiency and usability, lower operational costs, and enhanced security and privacy protections. As we dive deeper into the strengths, utilities, and importance of edge computing and Edge AI, we will come to understand how these technologies can pair up and boost the potential of applications such as face recognition.

What is Edge Computing?

Understanding Edge AI requires starting with defining edge computing. Edge computing is easy to understand, but hard to define because the definition can vary based on the application. In the most basic terms, edge computing places high-performance compute, storage and network resources as close as possible to end-users and devices. In practice, there are many different places on a network that can serve as an edge. A widely used model of edge computing comes from the Linux Foundation's LF Edge group shows this as a continuum from cloud to edge.

The LF Edge model focuses on the two main edge tiers that straddle last-mile networks that connect to devices and people: the "Service Provider Edge" and the "User Edge." Those tiers are further broken down into subcategories. According to this model, the user edge—the main area of relevance to our discussion of Edge AI—consists of:

- Self-contained edge devices, such as smartphones, wearables and automobiles;
- Gateway devices such as IoT aggregators, switching and routing devices;
- On-premises server platforms.



User Edge

Constrained Device Edge

Microcontroller-based, Highly Distributed in the Physical World

Smart Device Edge

Includes IoT (headless) and End User Client Compute in Accessible Locations

On-Prem Data Center Edge

Server-based Compute in Secure Locations

Last Mile Networks

Service Provider Edge

Access Edge

Server-based Compute at Telco Network and Edge Exchange Sites

Regional Edge

Server-based Compute at Regional Telco and Direct Peering Sites

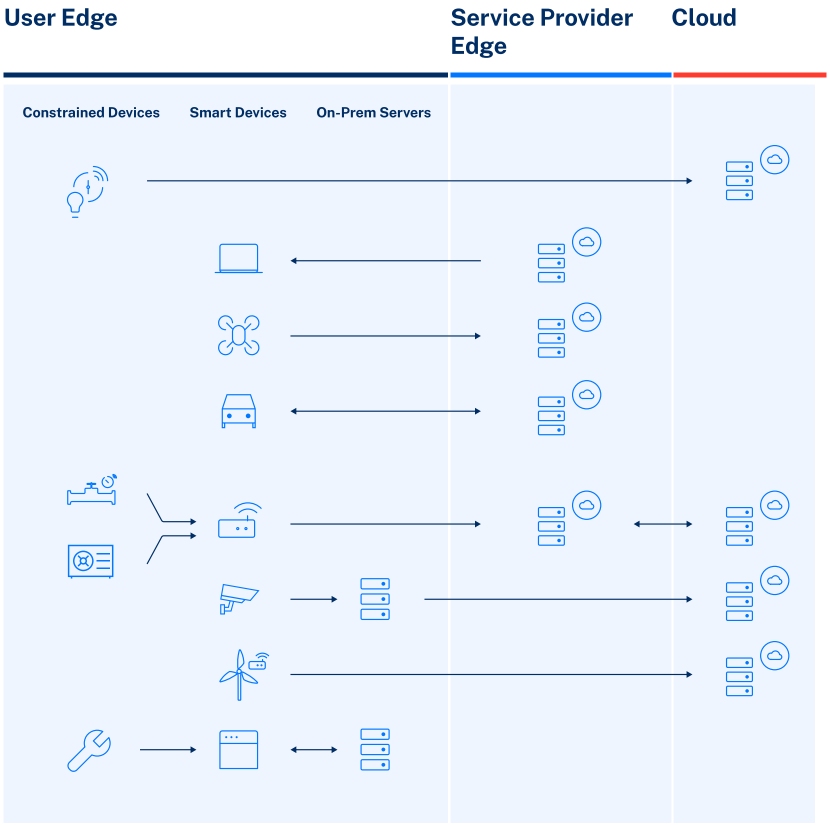
Internet Edge

Centralized Data Centers

Server-based Compute in Traditional Cloud Data Centers

The Internet of Things (IoT) refers to all devices connected to the internet that can communicate with each other; this is the smart device edge in LF Edge parlance. There are also devices in the "constrained device edge" which may be connected via private networks or protocols other than TCP/IP and may even have intermittent connectivity.

Older generations of edge devices were usually proprietary and closed, meaning that they weren't programmable by the customer. They also tended to lack the processing power to run complex applications such as image and data processing. To compensate for this lack of compute horsepower, data would be fed to a server, where it would then processed before returning to the local device—a process that takes extra time and can add cost in the form of bandwidth fees, for instance.



Constrained device communicating directly with cloud via router for all processing

Service Provider streaming video content to end user from regional DC for low latency

Service Provider streaming video content from drone to regional DC to process images

Service Provider coordinating Cellular Vehicle-to-Everything (C-V2X) for assisted driving

Local IoT Edge Gateway used to normalize protocols and perform analytics and control

On-prem edge servers process video and trigger events for cloud-based services

Embedded IoT Edge Gateway in field for analytics to reduce network cellular bandwidth

On-prem edge compute (no cloud) due to security concerns

The figure above provides an abstracted view of how edge devices deliver data to the cloud for processing. In some cases, such as video security, on-premises servers and gateways are used to aggregate data before delivery to a cloud for actions such as face recognition. The time and cost involved in the process are not optimal for many applications where control or instant answers are needed. Performing operations in real-time such as using facial recognition for access control requires low latency, and that means moving the analytics engine as close as possible to the data. In other words, Edge AI is needed.

What is Edge AI?

Edge AI is a rapidly evolving subset of the overall market for artificial intelligence that provides a solution to the shortcomings of cloud-dependent AI. Simply stated, Edge AI refers to running AI models on a device that has the appropriate sensors and processors. Network connectivity is not required for the device to process data and take action. Edge AI utilizes a new generation of specialized hardware and software that runs AI models locally instead of on remote servers, at once taking advantage of the benefits of AI and edge computing.

Training vs inference

Understanding how AI can work at the edge means understanding the two components of AI: training and inference. Model training requires an iterative process of analyzing a large amount of historical data, detecting patterns in that data, and generating an algorithm for that kind of pattern detection. The model is checked for accuracy, and the process is repeated. The second component of AI, inference, takes the algorithm generated by training and analyzes new data to produce insights.

Moving inference to the edge

Edge inference is the form of AI that is most commonly being deployed in the context of Edge AI. Inference has emerged as a key edge computing workload according to market research firm Omdia. Many companies have introduced chipset solutions to accelerate these Edge AI workloads. Instead of using general-purpose CPUs or expensive GPUs, chips are being designed specifically to accommodate the distinct data processing patterns of AI.

> The market for edge AI chipsets is expected to grow from $7.7 billion in 2019 to $51.9 billion by 2025, according to a forecast from market research firm Omdia.

It's not just hardware that's seeing advancements: the software used to build and deploy AI models and the models themselves are being optimized for edge environments. In some cases, tools are used to compress models to run on edge devices. In other cases, approaches such as TinyML are built from the start to accommodate use in edge devices. All told, advancements in Edge AI are resulting in advanced computer vision capabilities in compact, power-efficient edge devices ranging from intelligent video cameras to biometric terminals for automated access control.

Why Edge AI Matters

Edge AI has significant implications for facial recognition specifically, but it is valuable to understand the broader benefits of Edge AI for IoT and related applications.

Bandwidth Efficiency

Processing data at the edge reduces the need to transmit information over a network. This enables more efficient usage of bandwidth, which consequently results in:

- Reduced operating costs
- High performance results even on a constrained network

Latency

Edge AI enables low latency processing of data because data is either being moved a shorter distance or not at all (when processed on device):

- **Fully on-device processing** - Users no longer wait for information to return from a remote server; instead, all processing occurs on a local device, producing faster response times.
- **Hybrid configuration** - By combining edge with nearby server-side processing (on-premises, or at the access edge, in LF Edge parlance), the time required for round trip communications with centralized cloud-based services is reduced. This adds latency compared to on-device processing but offers a compromise in the form of additional processing power and the potential for using on-demand edge cloud services from cloud providers or telcos, for example.

Reduced Size, Weight, and Power

Conventional processor technologies can be costly and consume vast amounts of power, while Edge AI chips can deliver results from a dramatically economized power envelope.

Edge AI chips typically consume between 1 to 5 watts (sometimes less), whereas typical CPUs and GPUs run in the range of 50W and more. This means Edge AI chips reduce the reliance on heat sinks and fans for cooling and consume less power. The finished product's overall size and weight can be greatly condensed, resulting in a smaller, simplified design.

Privacy

Edge AI provides enhanced privacy protection for personal information. Processing data locally is inherently more secure than sending data across networks. Potentially sensitive data can be managed at the source, enabling the enforcement of policies around data storage or masking.

UI/UX (Integration and Responsiveness)

High latency can result in sluggish, non-intuitive user experiences. The low latency brought about by Edge AI allows for immediate feedback, generating a more interactive and compelling UI/UX.

Additionally

Edge AI can drive the use of dynamic, responsive, and interactive UI features, such as LCD or LED elements with reactive colors, shapes, and patterns.

Hardware Costs

Edge AI diminishes the need to rely on cloud services for processing and storage, offering the potential for a lower total cost of ownership. Data points underpinning this opportunity are beginning to emerge: A study from Deloitte notes that an Edge AI chip will cost around the same amount as a smartphone's processor while offering better performance and lower power consumption than traditional processor architectures.

A study from HPE suggested that the total cost of ownership (TCO) of using cloud services for data analytics workloads was 1.7 to 3.4 times higher than comparable on-premises deployments, with the latter offering workload throughput improvements of 45% over the cloud architecture.

Offline Functionality

Due to advances in chip design, Edge AI devices can now deliver advanced functionality without relying on network connectivity. This means that in cases of power or network outages (or intermittent connectivity) devices will continue to process data. Some of the most advanced Edge AI solutions have even been tailored for battery-powered operation, sipping power for months between recharge.

How Edge AI Benefits Face Recognition

Detection/Activation

Traditional video-based face recognition solutions rapidly process video feed to check for visible faces before delivering the results back to the client. However, continuously sending video across a network consumes a large amount of bandwidth. With Edge AI, cameras can selectively record and send footage that only includes faces. Whether or not face recognition is used, this method reduces the amount of data that is recorded and transmitted. That translates into lower bandwidth use and lower operational costs for transmission and storage compared to conventional server-side processing.

Image Quality

Face recognition performance is heavily correlated with the quality of submitted face images. With Edge AI, real-time image quality assessment can ensure the submission of only high-quality face image samples from the edge device. Depending on the use case, if low-quality is detected, the edge device can notify the user and provide instantaneous feedback. Image quality filters can be used in tandem with detection/activation technologies to further improve bandwidth efficiency.

Industrial Design

The enhanced power efficiency of Edge AI—and, by extension, its reduced size and weight—provides product designers with a broader range of innovative industrial design options. With an Edge AI processor, it is possible to construct form factors better suited to specific architectural limitations and environments. For example, next-generation AI-powered access control devices could include video cameras and face recognition capabilities, all in the size of a typical card reader or consumer video doorbell.

Privacy

In applications or environments where the privacy of people in the field of view needs to be preserved, Edge AI can be used to detect faces and then remove, obscure, or encrypt them at the video feed's source. In short, systems would be outfitted with advanced privacy protection and eliminate personally identifiable information at the video source.

Liveness

Edge AI can be used to differentiate between human beings and non-living spoofs in real-time. This security measure—known as Presentation Attack Detection (PAD)—prevents attackers from potentially fooling and bypassing a biometric system. Moving PAD to the edge enables this accurate and powerful threat-detection capability to integrate more advanced multi-sensor technologies while being more readily deployable and responsive.

Market Implications of Edge AI-Based Face Recognition

The Edge AI and face recognition technologies outlined above can be applied to numerous automated identification and authentication scenarios. Here, we examine their use in access control, video security, and payments.

Access Control

In controlled environments, ensuring a user's identity is critical. Bolstering face recognition systems with Edge AI will dramatically improve the speed, security, and reliability of access control devices. Other benefits include:

- Fewer technical failures due to low network-connectivity
- Decreased chance of information theft due to reduced data transmission
- Liveness detection and image quality filters will reinforce threat detection measures while still offering a streamlined user experience

Video Security

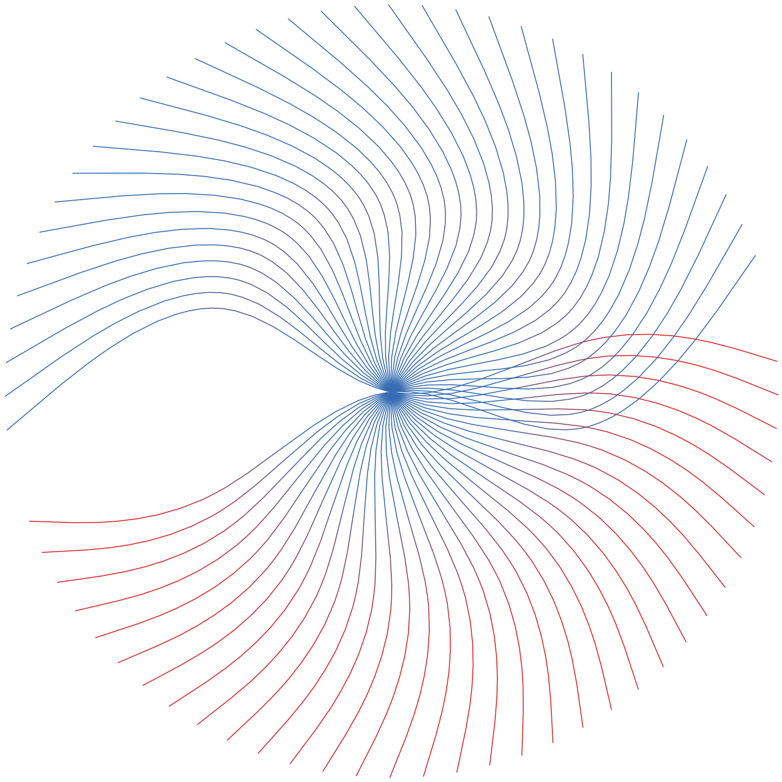
Applying Edge AI face recognition to video security can result in a more cost-effective, accurate, and user-friendly system. More specifically, it can:

- Increase speed of recognition
- When used in conjunction with next-gen processors, offer on-device processing for 4K and 8K resolution video streams for increased detail
- Apply privacy filters when appropriate

Payments

In a world where brick and mortar stores struggle to compete with the seamless experience provided by online retailers, Edge AI-based face recognition can deliver frictionless, touchless payments. For instance, it can:

- Guarantee a productive and satisfying user experience by enabling a quick, interactive checkout process from retail stores
- Rapidly identify potential fraud activity due to advanced liveness detection



In Conclusion

With rapid advances in imaging sensors, embedded computing, and deep learning technologies, Edge AI-powered face recognition can support and empower businesses in a way that is more cost-effective, accurate, and user-friendly.

Edge computing and Edge AI are foundational elements in computer vision-centric IoT applications, and their use will facilitate more widespread adoption and secure, appropriate use of face recognition.

For more detailed documentation, or to schedule a technical introduction to evaluate and integrate Paravision’s computer vision toolset, please contact us at
sales@paravision.ai

paravision