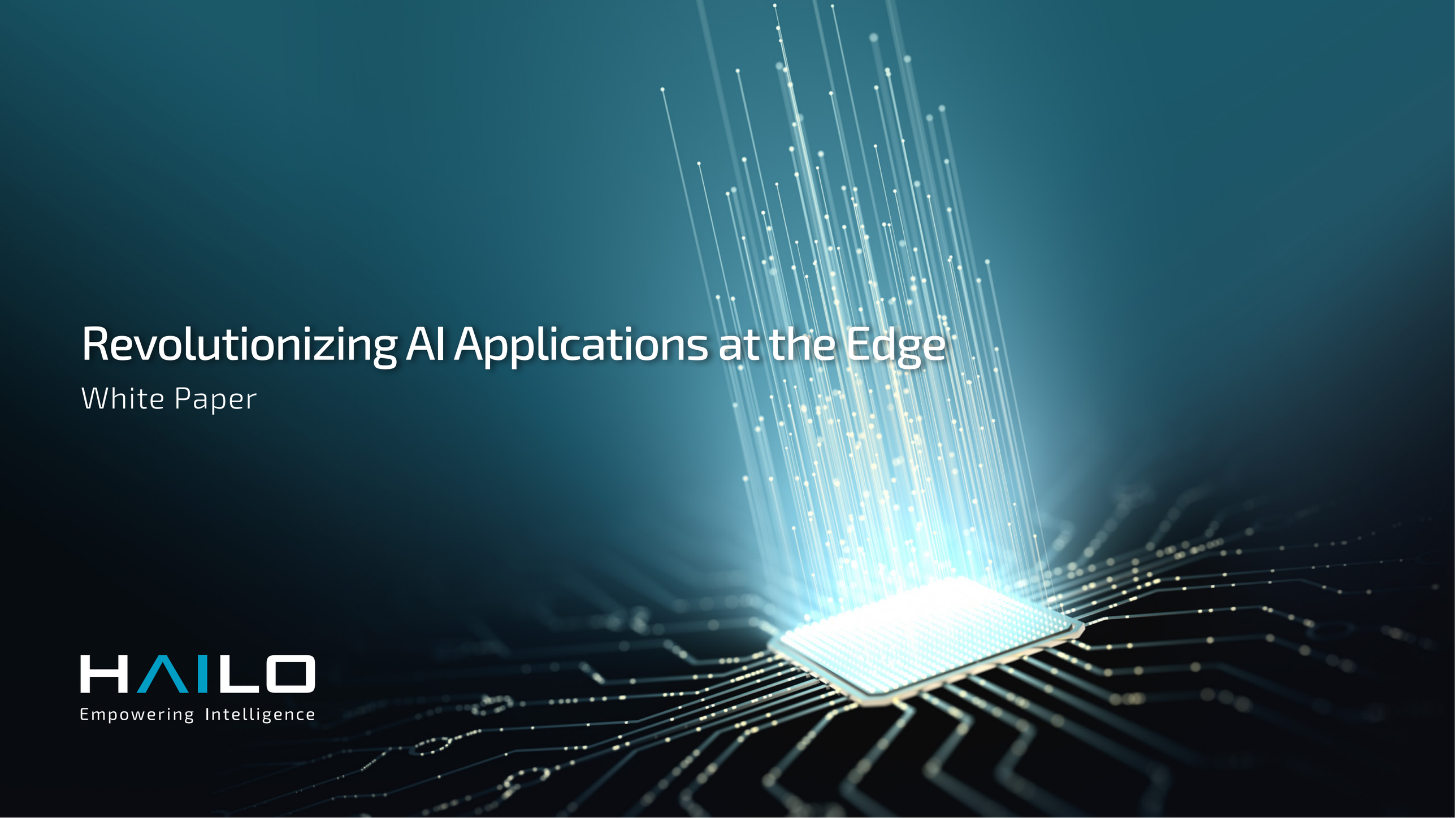
Revolutionizing AI Applications at the Edge

White Paper



HAILO

Empowering Intelligence

Revolutionizing AI applications at the edge

Digitalization processes across applications and domains have introduced an ever-growing number of devices and sensors. The vast amounts of data generated and the advancement in computing power give rise to machine learning technologies such as neural networks, intended to process the unstructured and abundant data[1]. Machine learning requires significant computational power and storage, so it initially became available only in the data center. However, such centralized processing is facilitated by multitudinous data transfers that create exorbitant bandwidth requirements. This limits real-world feasibility, especially in applications that require low latency or data privacy.

Edge AI first employed existing general-purpose processor architectures – CPUs and GPUs. Though evolved significantly, these have come short in their ability to deliver the required compute capacity for a given power budget (that meets product limitations) and cost. General-purpose processors are not efficient or simply do not fit applications where deep learning inference capabilities are embedded into small industrial or consumer devices. They require more power and produce more heat than many types of devices allow, limiting possible applications at the edge[2].

Due to recent advances in domain-specific computing architectures, a new generation of dedicated AI processors has emerged[2]. It is redefining what is possible at the edge – under size, power, heat dissipation and cost constraints – lifting previous limitations and enabling new applications. These processors are designed for deep learning operations at the edge, and though they vary in concept and actual performance, they are generally more efficient and cost-effective than traditional processors.

Data center AI vs. Edge AI

	Data center AI	Edge AI	
		Traditional computing architectures	Domain-specific architectures
Compute capacity	Very high	Low – very low	Moderate – high
Cost	Very high	Low – moderate	Low – moderate
Power consumption	Very high and growing	Low – moderate	Very low – low
Solution size	Very high	Low	Very low
Range of real-time applications	High	Low	Moderate

One of the most computationally demanding tasks that deep learning enables is intelligent machine vision. Smart video analytics are no longer reliant on or limited by available bandwidth, storage capacity or traditional processing architectures. High-resolution, high-frame rate and multi-sensor intelligent video processing is becoming a reality at the edge: on premise and on-device.

Edge AI implementation

The amazing capabilities of domain-specific processors can be introduced to the edge in a custom or standard form factor. The first allows maximum customization and best fit into a given system. The AI processor is designed into a device by its manufacturer in accordance with its mechanical, environmental and functional requirements. Within the

device, the AI processor is on a custom PCB, as a standalone processor or together with a traditional host processor, as well as other supporting elements (e.g. power, memory and storage and interfaces). In this way, cameras, industrial controllers, smart assistants, delivery robots, mobile phones and many other types of devices are empowered with AI to sense, analyze and respond.

The AI processor can also be implemented via industry-standard form factors that can be used in computing devices with appropriate expansion slots. Some AI processors are available on a standard M.2 or mini PCIe module. These are 22 x 30-110 mm and 30 x 51 mm card that house the AI accelerator, usually in a standalone configuration and with the minimum required supporting elements. Such a card plugs into the computing device, where it usually works with a transitional processor, offloading the neural processing task that the latter is not built for.

This is a turnkey method of empowering computing devices such as industrial computers, edge servers and commercial and personal computing devices with advanced AI capabilities for video analytics, NLP, HMI and so on. There are also dedicated edge AI solutions such as the Edge AI Box – a compact and low-power, video and/or voice analytics device designed to deliver AI capabilities by connecting to regular sensors and devices connected via standard interfaces such as Ethernet or USB.

Whatever the implementation, a good AI processor should allow the device manufacturer full programmability and flexibility in terms of neural network architectures and thus type of applications it supports, all while adhering to system requirements such as physical size, operating temperature range, appropriate qualifications etc. Moreover, it should provide the most efficient and cost-effective solution within these rigid limitations.

The value of high-throughput Edge AI across industries

Smart city

Cameras that monitor and perform analytics are superabundant in today's bustling cities. They collect petabytes of data from roads, squares and other public spaces in service of law enforcement, emergency services, health and sanitation and utilities maintenance. To make sense of this vast and unstructured data, AI is increasingly employed[3]. Today, a growing part of the AI processing can and should be performed at the edge, outside of large and costly central data centers.

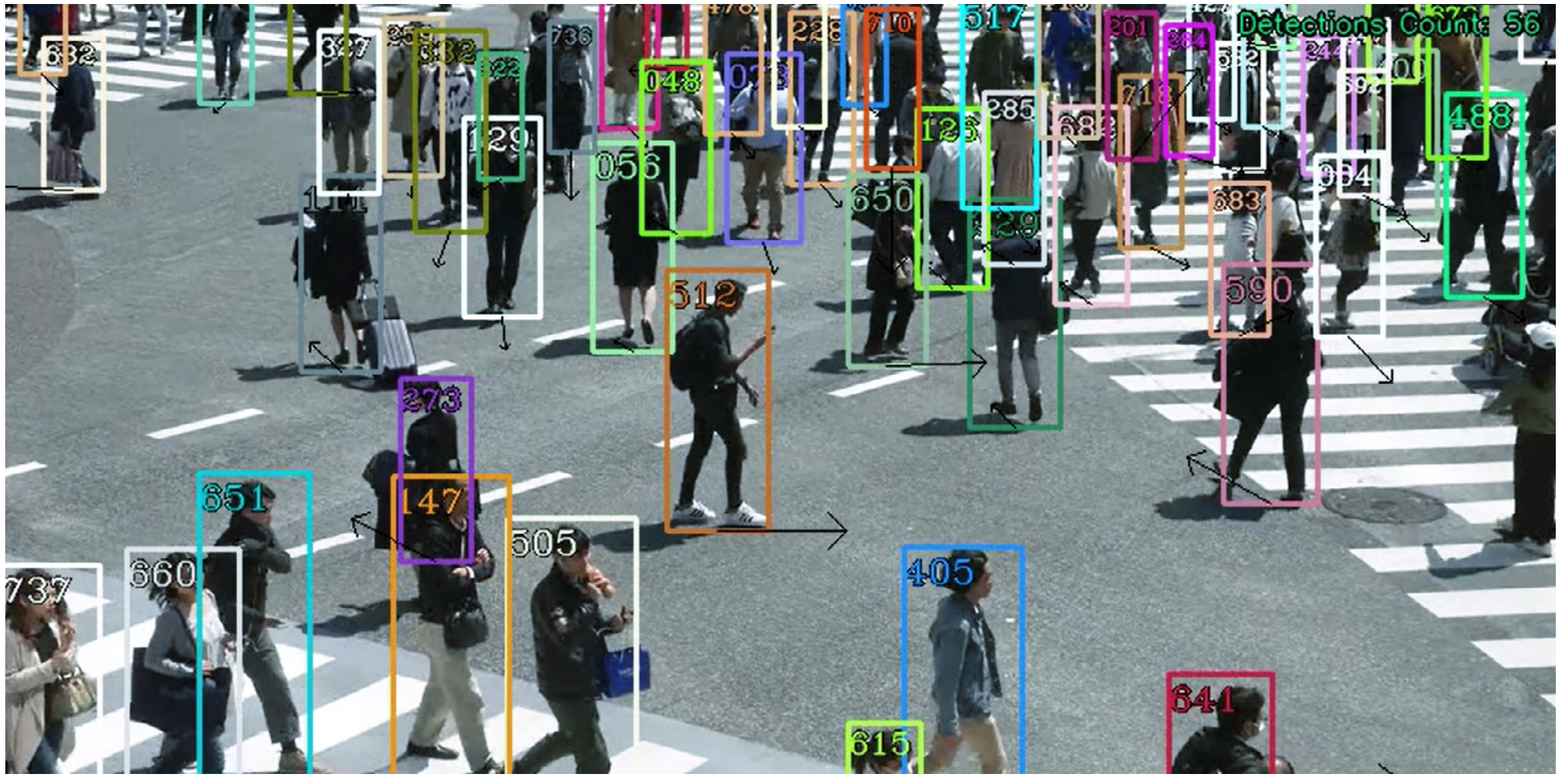
Edge AI processing allows cities to do more with less. Footage that is analyzed on the spot and in real time, is not only more valuable in time-critical applications, it also does not burden municipal budgets with transmission bandwidth and storage costs. When just the insights and analytics can be transmitted or stored, covering tens and hundreds of square kilometers is less cumbersome and costly to do. Local processing and decentralized networks are also easier to scale and secure than centralized ones.

The more efficient the computing power that is deployed at the end point, the more the user gets within a given budget or system. Powerful domain-specific processors deployed in a cost-effective system are designed to maximize end customer ROI (return on investment) by providing more capabilities and applications for less power and fewer resources (cost, power supply, space and deployment conditions). A high-throughput Edge AI solution should:

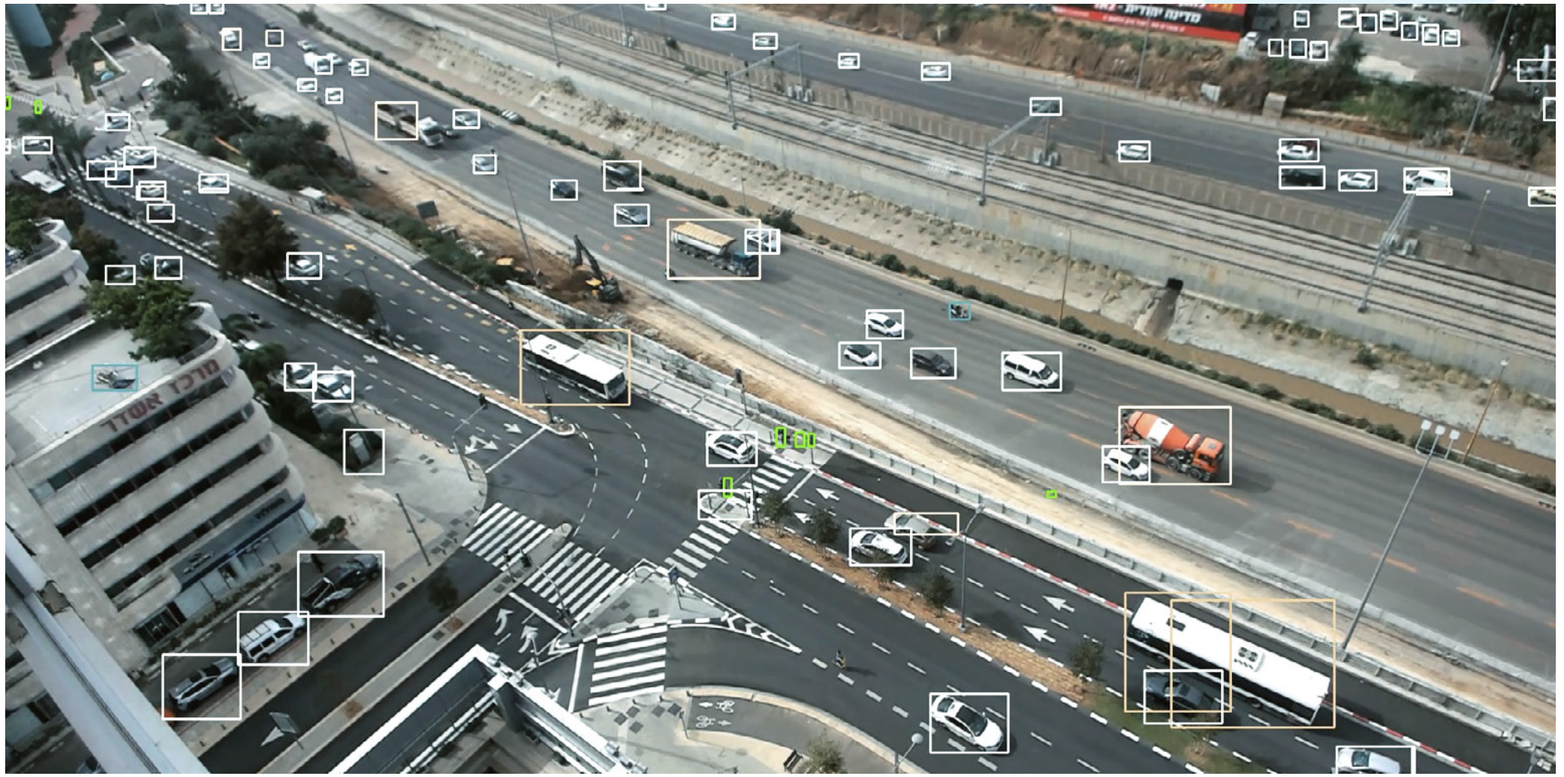
- **Maximize coverage –** high-resolution processing allows each camera to have a wider field of view and still detect the smallest objects. A large busy intersection can be covered by 1 or 2 cameras instead of 6 (or even 12) and still capture vehicle and people, and even hand-held objects and license plates. Moreover, one powerful AI processor

can process all the video streams on the intersection or any other multi-FOV (Field of View), multi-sensor outdoor or indoor space (e.g. square, street, building lobby, market, stadium, tunnel, multi-lane automated toll check post etc).

- **Enhance variety of applications –** provide enough computing power to run more advanced and compute-heavy neural models, previously barred from edge devices.



Pedestrian detection in high-resolution video streams for traffic management, safety and city planning analytics



Vehicle detection in high-resolution video streams for traffic management, safety and city planning analytics

- **Multitask –** a single AI processor can run several neural network models at a time, doing the most in complex environments. The intersection cameras can simultaneously run plate recognition for red light and speed enforcement, analyze traffic patterns on the road and crosswalks to regulate lights and monitor the use of face masks or bike helmets.

Smart retail

AI is helping brick-and-mortar retail compete in a digital world, through automation, optimization and intelligent marketing. Using advanced AI for video analytics, retailers can create the right shopping experience for better business outcomes[4].

Due to the nature of physical retail operations (individual stores or store chains), AI performed locally often makes more sense. An automated checkout, for instance, needs to perform perception operations with minimal latency. Sending the video input for processing to the cloud or to a remote server at company HQ and back may add hundreds of milliseconds, impacting user experience. Security alerts may also be time-sensitive and need quick local action. Another major consideration in support of local AI processing is customer privacy. Local closed-loop systems that do not need to transmit or store footage are the most compliant and secure.

AI can be introduced with the deployment of intelligent cameras that perform AI inference on-device or by implementing a single computing device, such as an Edge AI Box, and connecting it to multiple regular cameras. Whether embedded in a smart device or deployed on a store network, the more powerful and efficient the AI processor inside, the greater the possibilities and eventual ROI for the business.

High throughput and domain-specific efficiency (within the tight constraints of device size, power and heat dissipation) allow retailers to:

- **Cover the entire trading floor** with as little as a single dedicated chip – several high-resolution video streams of multiple camera angles (which are usually required, even in a small store) are processed simultaneously by a single device.
- **Detect small objects and movements –** the high-resolution processing enables detection and recognition of individual products on the shelf, small movements by individuals, a spill in a distant aisle and more. Product detection, for instance, can be used for loss prevention and for inventory management, with the AI alerting about empty shelf

space that needs to be restocked. It can also inform product-level business analytics.

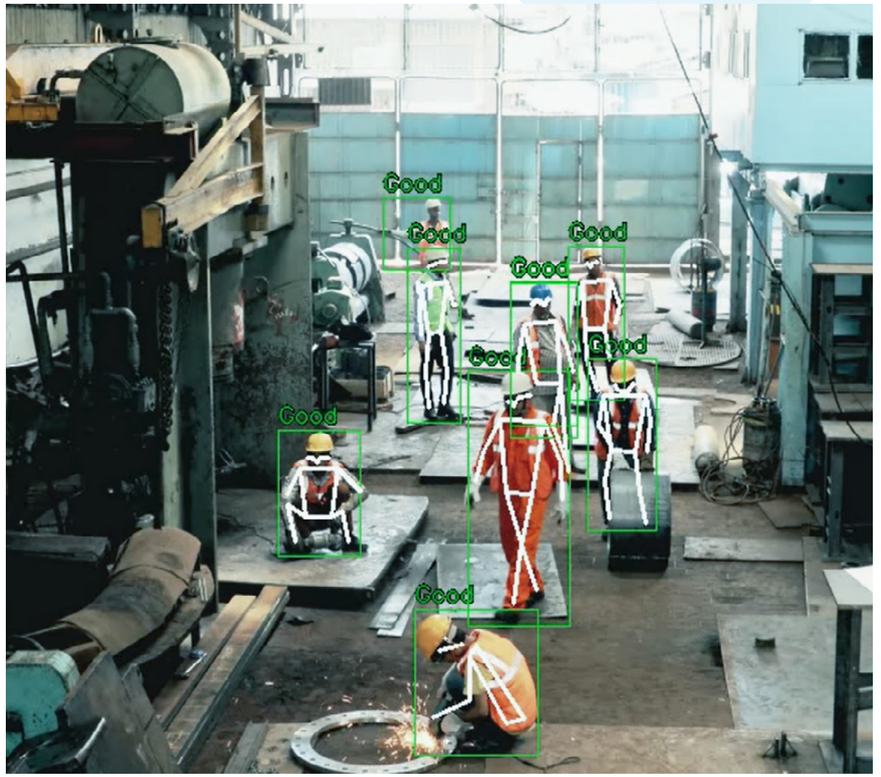
- **Automate operations –** processing that is real-time and low-latency is used in customer-facing and back-of-the-house robotic devices that need to make a decision and take action in milliseconds to function safely or provide a seamless user experience.
- **Multitask –** to maximize efficiency and ROI, multiple NNs can run on a single video stream for different purposes. For instance, one video stream can be used to both track customer movement to create detailed heat maps for sales analytics and perform object detection and recognition of products on the shelves or in customers' hands.
- **Solve data security and customer privacy concerns** by keeping the data processing they require in-house. When video footage is not transmitted to the cloud for analysis and is not sent to a remote server for storage, there are fewer possible points of failure and compliance concerns.

In-store people recognition and counting in high-resolution video for customer traffic management and retail optimization

Industry 4.0

Automation of manufacturing and warehousing has revolutionized the economics of production, with vision-based AI adopted for robotic automation, quality control, process and equipment monitoring, continuous improvement strategies and facility safety & security[5]. Deep learning is used in defect detection in pharmaceutical and electronic manufacturing, smart animal farming and agricultural yield sorting, monitoring safety measures on construction and mining sites, leak detection and facility security in energy production, to name a few.

The speed of manufacturing processes and operational efficiency, as well as security considerations demand processing be done locally. AI-capable processors are designed into industrial devices such as high-speed line cameras, manufacturing robots (sorting, assembly, paint shop etc.) and autonomous



Running multiple neural networks on high-resolution video in tough conditions – detecting protective gear on workers and worker pose estimation for on-site safety monitoring

logistic vehicles that need to navigate the facility safely. There is also demand for on-site closed-loop monitoring and alerting, especially for safety and security applications. These include monitoring the site or facility for safety hazards, personnel's use of PPE (personal protective equipment) and compliance with regulations, unauthorized personnel and suspicious activity, among others.

Powerful domain-specific processing means higher ROI for manufacturers. The high throughput enables the high-FPS (Frame Per Second) processing that a high-paced manufacturing lines requires, supporting high-resolution and complex applications for tasks such as defect detection. It also performs high-resolution, multi-stream, multi-neural network video processing, which maximizes camera coverage, while still allowing to detect small objects like hardhats and safety goggles on workers, alert about unauthorized personnel and non-safety compliant worker behavior. All these are possible within a strict power budget, limited space and TDP within a compact and rugged industrial-grade device.

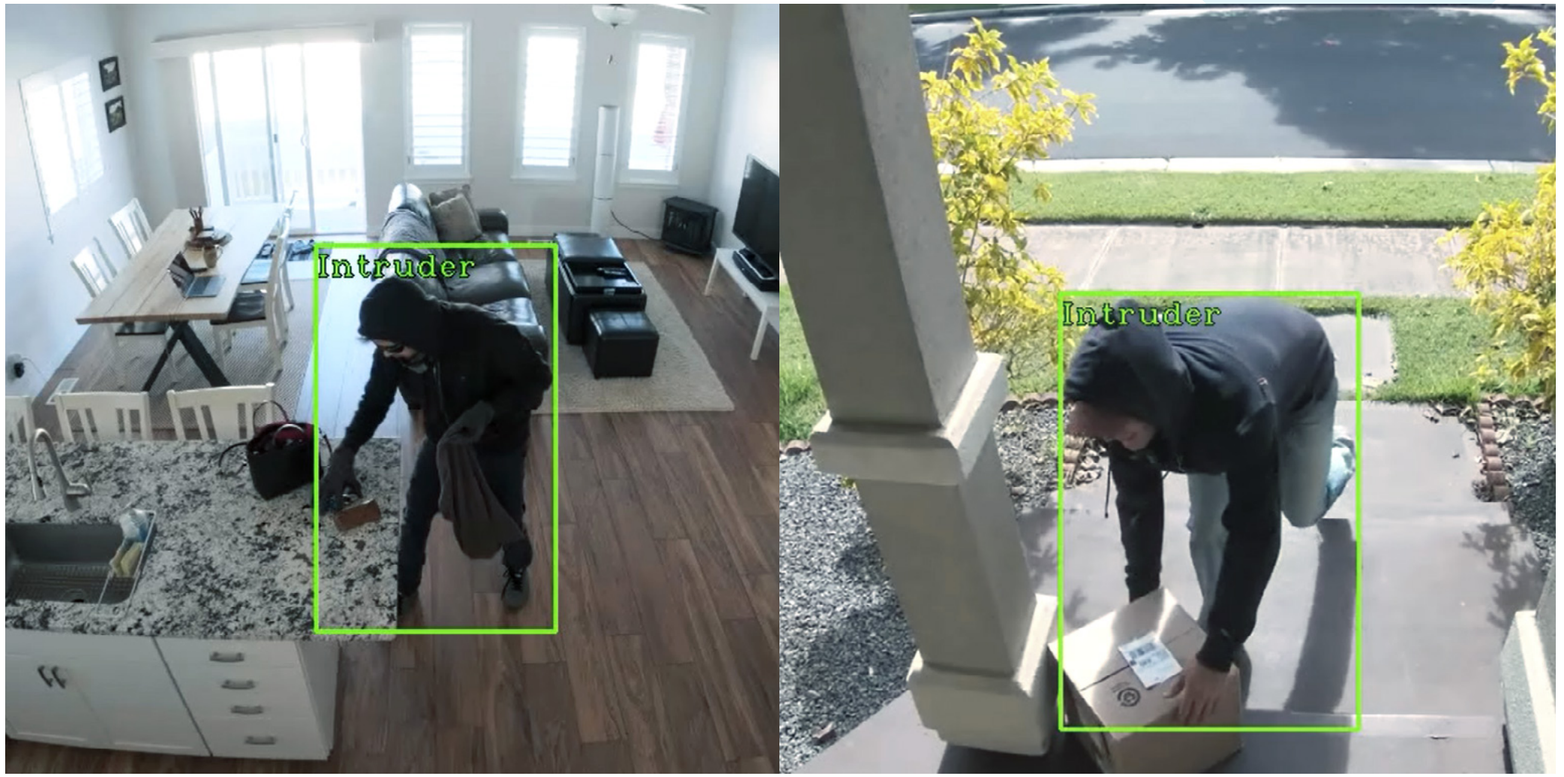
Smart home

Adoption of smart vision sensors and connected home devices for safety, security and comfort is growing[6]. In the private home, more than in other environments, data security and privacy concerns create a strong preference for devices that do not need to send data outside the residence and that don't have to record or store a lot of footage, if any. Moreover, some smart home applications may not want to rely on network connectivity, bandwidth and latency, such as emergency response or security alerts.

To fit into smart consumer devices, the AI processor needs to be small size, low power and cost-effective, as well as comply with consumer safety standards. If the dedicated AI processor is high-throughput and high-efficiency, it will provide the best value under such constraints, especially in such demanding, unstructured data-rich applications as intelligent machine vision.

The high-performance domain-specific processor is able to process multiple high-resolution inputs, running one or multiple neural networks with high accuracy. Thus, it enables intelligent vision that has wider coverage and lower latency and benefit such advanced everyday applications as:

- **Automated assisted living** or monitoring babies and pets – several neural models running continuously and simultaneously to recognize the person or animal and monitor its pose, movement, gestures and sounds to alert about needed assistance or an emergency situation.
- **Remote control** of appliances or home features (doors, lighting, environmental controls) through gestures and/or voice. Such capabilities are especially valuable for homes with disabled individuals and small children.
- **Face detection and recognition** for security purposes, as part of a home security system or smart door lock system.
- **Centralized Smart Home processing –** a single computing device, empowered with high-throughput AI, can securely connect and provide AI processing for all other connected devices and sensors across the home.



Intruder detection for home security cameras and networks. Streams from multiple cameras around the house can be processed simultaneously on one device.

Smart workplace

Enterprises and commercial building owners and operators are adopting AI to bring efficiency and cost-savings to their operations, increase competitiveness (variety and quality of features offered to tenants) and reach sustainability goals. Intelligent vision is used in workplaces for security and safety monitoring and AI-based gesture control of devices. Monitoring spaces helps:

- Direct human **traffic**, e.g. avoid crowded lobbies by controlling elevators, use smart analytics to generate spatial heatmaps
- Comply with public **safety regulations**, like mask & social distancing monitoring
- Manage communal **parking** using vehicle and vacant lot recognition
- **Conserve energy** for sustainability or budgetary reasons – e.g. regulation of ventilation and lighting according to occupancy)
- Maintain **physical and information security –** monitoring for unauthorized personnel or the general workspace of in-office and remote employees.

Edge AI allows to keep the sensitive data of businesses and their customers and employees on-premises. As cloud IT infrastructure transitions raise data security concerns, edge AI computing allows some functions to stay local, at a fraction of traditional deployment cost and space.

An edge device powered by a dedicated high-throughput AI processor further increases the ROI for enterprises, taking on greater processing workloads for less. With higher-resolution processing, fewer cameras and devices are needed to perform security operations in the lobby and across the building. Multiple applications do not mean more computing devices – one on-camera chip can provide processing for both security, social distancing and traffic monitoring in a busy lobby or monitor occupancy and control environmental setting and teleconferencing inside the boardroom.

Such capacity and multi-tasking benefits reduce cost and complexity for enterprises. Additional cost-savings can possibly be achieved by deploying AI via standard form factors and capitalizing on existing computing and sensor devices.

Smart transportation

AI-powered fleet management, analytics and in-cabin monitoring helps authorities and service providers to optimize public transport and cargo and delivery operations. Used in buses, truck, taxis and even trains, in-vehicle AI boxes are replacing traditional NVRs (network video recorder) and fleet management and tracking systems, offering powerful capabilities for driver and passenger safety monitoring, vehicle and valuable cargo security and even better infotainment content suggestion and targeted advertising to passengers.

An in-vehicle Edge AI Box is designed for in-vehicle, on-the-move use, complying with size, power and TDP standards. Powered by a high-throughput dedicated AI processor, it becomes an all-in-one powerhouse, able to process inputs from multiple cameras and sensors and perform several AI tasks simultaneously, while also providing connectivity and positioning functionality.

Conclusion

AI inference at the edge offers unique benefits across industry verticals and deployment environments. Consumers, businesses and operations of various sizes are increasingly recognizing its benefits and adopting intelligent edge devices.

The edge processors that power these devices should not force customers into a tradeoff between AI performance and size, power and price, as the latter are usually constrained. High-throughput, high-efficiency dedicated AI processors help avoid this pitfall, allowing businesses and consumers to enjoy advanced and robust deep learning capabilities where they need them – on device or on premise.

The various implementation options and their flexibility in supporting a wide range of applications make domain-specific AI processors major vehicles for introducing AI into more diverse environments and applications. These revolutionary edge AI processors are empowering new customers and industries with AI every day, solving problems and introducing efficiencies and cost-savings.

Desirable attributes for a best-in-class edge AI solution

Function & performance	System attributes & requirements
High neural processing throughput	Low power consumption and high power efficiency
High processing efficiency (throughput per watt of power)	Small size
Low latency	Passively cooled
High-resolution video processing	Standard interfaces and form factors
High accuracy of output	For industrial applications – ruggedized, extended operating temperature range
Support multiple video streams	
Support multiple neural models running simultaneously Support multiple types of sensors	

1. Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, Amit P. Sheth, Machine learning for internet of things data analysis: a survey, Digital Communications and Networks, Volume 4, Issue 3, 2018, Pages 161-175 (link)
2. Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59 (link)
3. Bhattacharya S, Somayaji SRK, Gadekallu TR, Alazab M, Maddikunta PKR. A review on deep learning for future smart cities. Internet Technology Letters. 2020;e187 (link)
4. Book, A. Complete guide to machine learning and deep learning in retail. The Pourquoi Pas. 08/05/20 (link)
5. Versace, M. From Computer Vision To Deep Learning: How AI Is Augmenting Manufacturing. 16/06/20 (link)
6. Marr, B. The 5 Biggest Smart Home Trends In 2020. Forbes. 13/01/20 (link)

HAILO

Empowering Intelligence

www.hailo.ai