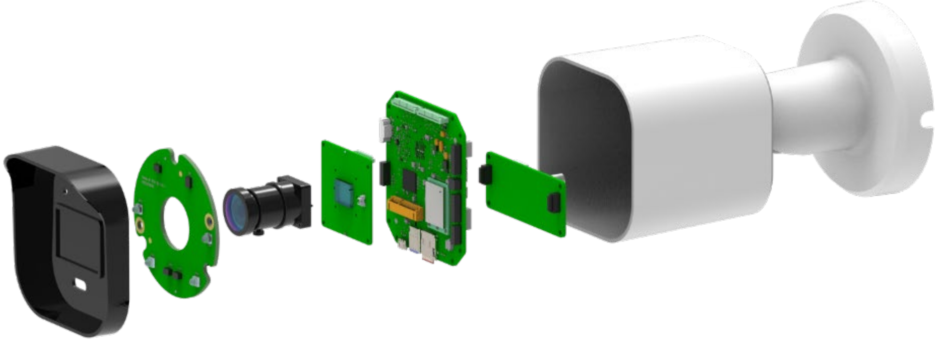
Edge and Gen AI for Embedded Vision: Breaking Through Performance and Cost Barriers

Sebastien Dignard, President, Macnica Americas, Inc.

Manufacturers whose products feature embedded vision are seeking a step change in performance and capability but are encountering mixed results. The performance and cost boundaries of traditional computer vision techniques limit the ability to plan product roadmaps beyond the status quo. These challenges encompass various aspects, from raw computing bandwidth to cloud deployment constraints, creating interrelated issues that cause manufacturers to hesitate.

If 2023 was the year the world discovered generative AI (gen AI), like ChatGPT and Midjourney, this year is when embedded teams can use it to augment computer vision-based results. Advancements in gen AI-based vision language models (VLM) and system-on-chip (SoC) hardware make deploying edge-based embedded vision a reality. Manufacturers are seeing material benefits already, from security cameras answering natural-language questions (“Did Amazon deliver a package today?”) to smart construction camera systems that improve security and safety as well as perform real-time assessment of assets (“Confirm everyone wore hardhats onsite today”).



This paper explores how combining gen AI with advanced hardware features overcomes the limitations of traditional embedded vision, enabling enhanced capabilities on edge devices with lower costs, reduced power consumption, and decreased reliance on cloud services. It discusses the key challenges faced by manufacturers, the technologies that enable next-generation embedded vision, and the real-world applications that demonstrate the benefits of edge AI.

Challenges with edge-based embedded vision

Edge-based embedded vision refers to computer or machine vision algorithms running on edge devices rather than relying on remote servers or cloud infrastructure. It enables the analysis of visual data to be closer to where it is captured, on devices like cameras, industrial sensors, drones, and cars.

Getting embedded vision right isn't easy, as it requires knowledge and skills in the following areas:

- **Hardware design** to perform complex computations locally without exceeding the power and thermal requirements of the overall system. This typically requires some combination of CPU, signal processor, memory, and video interfaces to perform operations like on-chip video encoding and application-specific tasks.
- **Computer vision algorithms** to extract meaningful information from visual data efficiently. Common tasks include object detection, image classification, facial recognition, and anomaly detection.
- **AI image processing** to perform advanced computer vision tasks within the constraints of modern compute and storage capabilities. Examples include real-time image stabilization, low-light processing, and temporal filtering.
- **Integrating and optimizing models** to work in the ecosystem of the embedded system. This includes quantization, pruning, and distillation to improve model performance and efficiency.

The more efficiently an embedded vision application can capture, process, and act on visual data, the better it can support decision-making. It's the milliseconds difference between an industrial robot hitting a moving obstacle and avoiding it.

Developing embedded vision solutions at the edge comes with the following four challenges.

1. Real-time image processing demands

Real-time image processing requirements grow with every new product, driven by the capabilities of high-resolution sensors and the market demand for advanced features. A single 4K camera stream operating at 30 frames per second (4Kp30) generates approximately 746 MB/s of raw data – before the computational overhead of computer vision algorithms and AI image processing.

AI model computations often contain millions of parameters that require substantial resources and software optimizations. This challenge is compounded when dealing with multiple video streams, common in applications like industrial automation and security systems. Each stream must undergo multiple processing stages, from initial image enhancement and AI inference to final encoding, while maintaining strict latency requirements.

What is AI inference?

AI inference is the process of a previously trained artificial intelligence model analyzing new data to make real-time decisions. At runtime, this trained model continuously analyzes incoming video frames to detect and classify activities. For example, a smart security camera determining what constitutes normal versus suspicious behavior based on being trained by millions of images.

For embedded vision systems, inference must not introduce undesirable latency or jitter into the processing pipeline. This computational workload becomes challenging when running on traditional edge devices that do not have specialized hardware or algorithms to process data efficiently.

Input quality also affects computational needs, as degraded sensor data leads to poor AI model performance and unreliable decision-making. Image processing pipelines must be designed to handle challenging environmental conditions, varying lighting situations, and complex scene dynamics.

The choice of sensor plays a critical role in computational efficiency. For example, selecting a Sony CMOS sensor with built-in HDR and noise reduction capabilities reduces resource requirements in an already demanding processing pipeline.

2. Implementation challenges for edge devices

The available hardware ecosystem has traditionally been a limiting factor in deploying edge-based embedded vision solutions. Implementing AI-intensive workloads can push power consumption and thermal management beyond the limits of many edge devices. These constraints are particularly problematic in applications like outdoor surveillance systems and drones, where systems must operate reliably on battery power under varying scene conditions.

Form factor constraints further complicate implementation. Many applications require compact cameras and processing components that can be deployed in space-constrained environments. This impacts the physical size of the vision system hardware and the design options for heat management and power delivery.

Overcoming these challenges lies in balancing the physical constraints with the need for powerful processing capabilities.

3. Considering cloud deployments as the default

Despite advances in edge device designs, many engineers and product leaders believe that AI-intensive workloads can only run on cloud infrastructure. While cloud computing offers virtually unlimited processing power, it's worthwhile considering edge-based embedded vision devices to overcome these constraints:

- Latency issues can introduce delays in round-trip time for data transmission, processing, and response. Variable and unpredictable latencies can be problematic for image processing applications requiring deterministic response times, such as robotic control systems or safety-critical monitoring applications.
- Cloud network bandwidth and throughput may not support application needs or create unpredictable bottlenecks, particularly when dealing with high-resolution video streams. This also impacts determinism and overall performance.
- Security and privacy concerns may limit options for processing sensitive data and maintaining strict control over how that data is handled. Additionally, many products fall under data sovereignty rules, constraining where cloud workloads can run.
- The operating costs of cloud-based workloads can scale unpredictably with deployment size. Bandwidth costs, storage fees, and compute resources contribute to a total cost of ownership that may exceed what users want to pay.
- Reliance on specific cloud providers can lead to vendor lock-in, limiting flexibility and potentially increasing long-term product maintenance costs.

4. Underestimating the roles of strategic and technical partnerships

The history of computer vision shows that the successful implementation of advanced systems requires deep expertise across multiple domains. AI-based embedded vision is no different. The complexity of these systems makes it easy to underestimate the level of cooperation and intellectual property knowledge required for successful deployment.

General-purpose processors struggle to meet performance requirements while staying within power, thermal, and space constraints. Custom hardware solutions from a single vendor often come with prohibitive costs or development timelines that make them impractical for product roadmaps.

Realistic and achievable product development requires a multidisciplinary approach, with skills in hardware design, AI algorithms, image processing, VLM integration, and application-specific requirements. No single organization possesses all the necessary knowledge in hardware architecture, AI algorithms, image processing, and application-specific requirements – it's critical to choose the right multi-vendor expertise that best fits the application.

Enter Macnica: Advanced embedded vision at the edge

The breakthrough in low-power, highly efficient embedded vision on edge devices comes from converging two technologies: gen AI-based VLMs and purpose-built System-on-Chip (SoC) hardware. To achieve this, Macnica, the North American leader in imaging systems design and development, has partnered with Sony Semiconductor, iENSO, Infineon, and Ambarella. This multidisciplinary collaboration ensures manufacturers get a range of video encoding, computer vision, and AI image processing options to achieve these benefits:

- Improved image quality and reliability.
- Reduced bandwidth requirements and power consumption.
- Enhanced data privacy and security.
- Sophisticated real-time decision-making capabilities.
- Longer device battery life compared to cloud-based alternatives.
- Smaller form factor to fit various edge device use cases.
- Lower total cost of ownership from reduced cloud processing costs.

What is a Vision Language Model?

A Vision Language Model is an artificial intelligence system that processes and combines visual (image) and textual (language) data. Unlike traditional computer vision techniques that simply detect objects, VLMs understand scene complexities to support tasks like language translation, content creation, and chatbot interactions. These capabilities make VLMs ideally suited for applications like surveillance systems, smart home appliances, and robotics that interact naturally with their environments.

For example, given a street scene, a product using classical convolutional neural networks (CNN) could detect cars as objects. A VLM could take the same scene and determine, “There is a red car parked next to a fire hydrant on a busy street.”

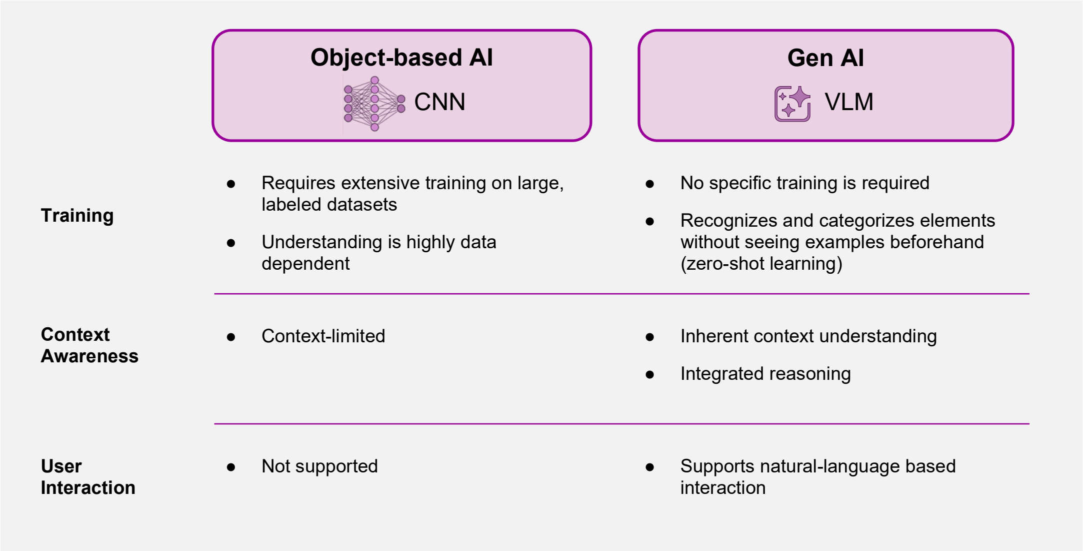
The following sections describe the key technologies behind these edge-based embedded vision products.

The evolution from object-based CNN to gen-AI VLM

Traditional object-based vision systems, using convolutional neural networks (CNNs), can only detect and classify predefined objects according to their training data. This approach is valuable for basic vision applications but falls short when products require a more sophisticated understanding of scenes or encounter new situations.

Modern VLMs allow devices to understand and describe complex scenes, recognize actions, and generate natural-language descriptions of visual content. Table 1 describes the differences between the two methods.

Since CNNs excel at object detection with relatively lower latency, some companies like iENSO combine them with VLMs to help meet the performance requirements of edge devices.



	Object-based AI CNN	Gen AI VLM
Training	• Requires extensive training on large, labeled datasets • Understanding is highly data dependent	• No specific training is required • Recognizes and categorizes elements without seeing examples beforehand (zero-shot learning)
Context Awareness	• Context-limited	• Inherent context understanding • Integrated reasoning
User Interaction	• Not supported	• Supports natural-language based interaction

Table 1: Comparison between object-based AI and generative AI

This combination also enables a new class of applications that can adapt to changing conditions and provide richer insights from visual data.

Advanced SoC designs to power the AI pipeline

The Sony Semiconductor, Macnica, iENSO, and Ambarella partnership enables hardware, firmware, and software designs that ensure the AI imaging processing pipeline has the resources to run efficiently. This brings value in the following ways.

SoC design and integration expertise

iENSO has systematically transitioned its camera product portfolio from FPGA-based designs to SoC platforms, driven by the increasing sophistication and efficiency of SoC solutions for vision processing workloads.

The advantages of SoCs for embedded vision are:

- Specialized processing cores that can deliver highly efficient performance per watt compared to traditional FPGA designs.

- Accelerators for computer vision and AI computations, optimized for CNNs and VLMs, further enhancing their capabilities
- Support through extensive vendor ecosystems for low-risk development and customization.
- Versatility in vision solutions for a broad range of feature, performance, and cost requirements.

AI-based image signal processing

Each embedded vision platform is based on an Ambarella SoC, providing 4Kp30+ image processing, video encoding and decoding, and CVflow® computer vision processing in a single, low-power design. This includes Ambarella's Artificial Intelligence Image Signal Processor (AISP) which uses neural networks to augment image processing to enable features such as color imaging with low light and high dynamic range (HDR) processing.

Hardware-accelerated computer vision

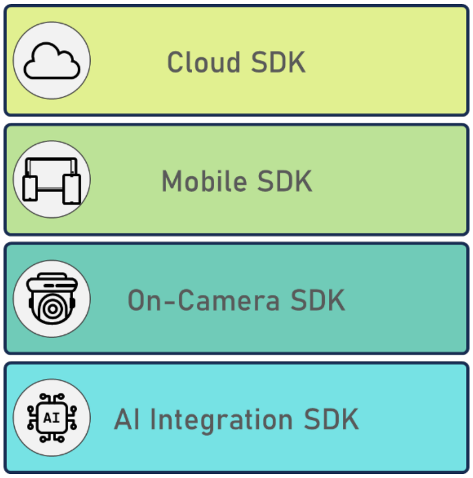
Specialized hardware accelerators complement these processing units to run specific computer vision operations with high performance within strict power and thermal constraints. These operations include:

- Increasing AI model accuracy under low light and nighttime conditions
- Reducing video bandwidth to lower resource and power consumption
- Improving contrast and rendering vivid colors
- Reducing motion artifacts to improve analysis

Developer-friendly tooling

The iENSO Embedded Vision Platform as a Service (EVPaaS) Development Framework provides SDKs and tools for building and deploying custom computer vision applications on Ambarella SoCs. This framework enables users and value-added resellers to create sophisticated vision solutions without deep expertise in the underlying hardware architecture.

With EVPaaS, developers can create custom applications that run directly on the camera hardware, leveraging its advanced processing capabilities for real-time computer vision tasks. The SDK provides robust support for integrating custom AI/ML algorithms, allowing teams to deploy their proprietary models or modify existing ones for specific use cases. It also extends beyond edge processing to support the development of mobile applications and cloud dashboards that interface directly with the camera system.



iENSO EVPaaS Platform

Example use cases

Multiple industries benefit from edge-based AI vision systems, each demonstrating how different capabilities can address specific challenges.

AI-enabled security cameras

Modern security applications require sophisticated real-time analysis capabilities that go beyond simple motion detection. An edge AI camera system monitoring a street-facing entrance can perform complex object classification and behavior analysis directly on the device and answer users' natural language questions.

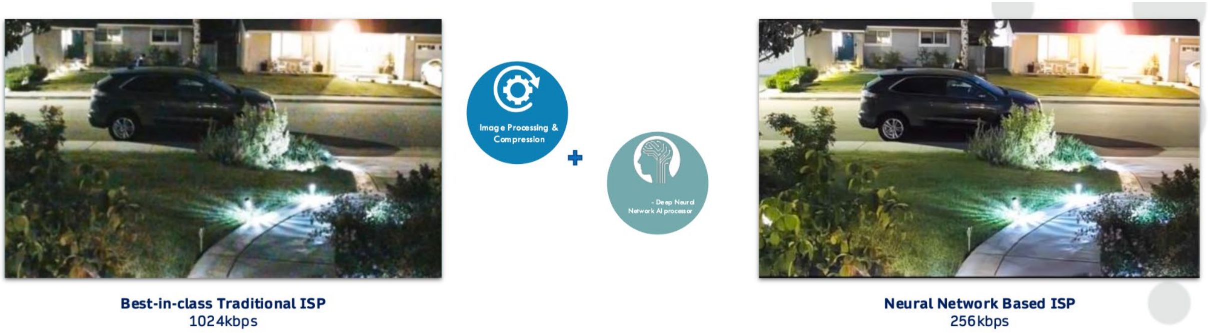


Image Processing & Compression
+
- Deep Neural Network AI processor

Best-in-class Traditional ISP
1024kbps

Neural Network Based ISP
256kbps

Enhanced night video quality with 75% bitrate reduction

By processing data locally, this system can distinguish between routine activities and genuinely suspicious behavior, significantly reducing false positives and maintaining data privacy by keeping sensitive information on the device.

With AISP built-in, the system can perform advanced optimizations to ensure reliable operation in low-light and nighttime lighting conditions. Through sophisticated image signal processing and AI-enhanced noise reduction, the camera can also improve image quality, reduce motion artifacts, and decrease video bandwidth usage.

Industrial and consumer robotics

Real-time path mapping and localization require processing speeds that cloud-based solutions cannot match. Edge-based embedded vision enables robots to make immediate decisions about obstacle avoidance and path planning – critical for safe and efficient operation in dynamic manufacturing environments or everyday obstacles.

Smart farming applications

Computer vision applications in agriculture often operate under challenging outdoor environments with reduced bandwidth between remote devices and cloud servers. By deploying edge-based embedded vision, remote in-field systems can process high-resolution imagery on-device, enabling continuous crop monitoring, pest detection, and optimizing irrigation patterns without requiring constant network connectivity.

Smart home applications

Market studies show that many AI and cloud-enabled smart home applications do not give consumers and manufacturers confidence about their security and privacy.

The benefits of VLMs at the edge – with local data processing and no network traffic – align with most households' needs for privacy. From smart refrigerators that capture content to identify food items, estimate freshness, and detect missing items to context-aware automation that creates intuitive and responsive smart home environments, consumers can get sophisticated experiences without compromising security.

Bringing edge-based embedded vision to reality

The combination of on-device Vision Language Models and innovative SoC hardware overcomes the limitations of cloud-based image processing and reduces the resource barriers to more sophisticated applications. Manufacturers implementing edge-based embedded vision solutions can significantly improve processing efficiency, cost-effectiveness, and operational capability while maintaining high standards for security and privacy.

As these technologies enter the market, SaaS and cloud-based subscription solutions favored over the last decade must adapt to survive.

To learn more about the Macnica imaging ecosystem, including partnerships with Sony, Ambarella, Infineon, and iENSO embedded vision solutions, visit macnica.com/americas.