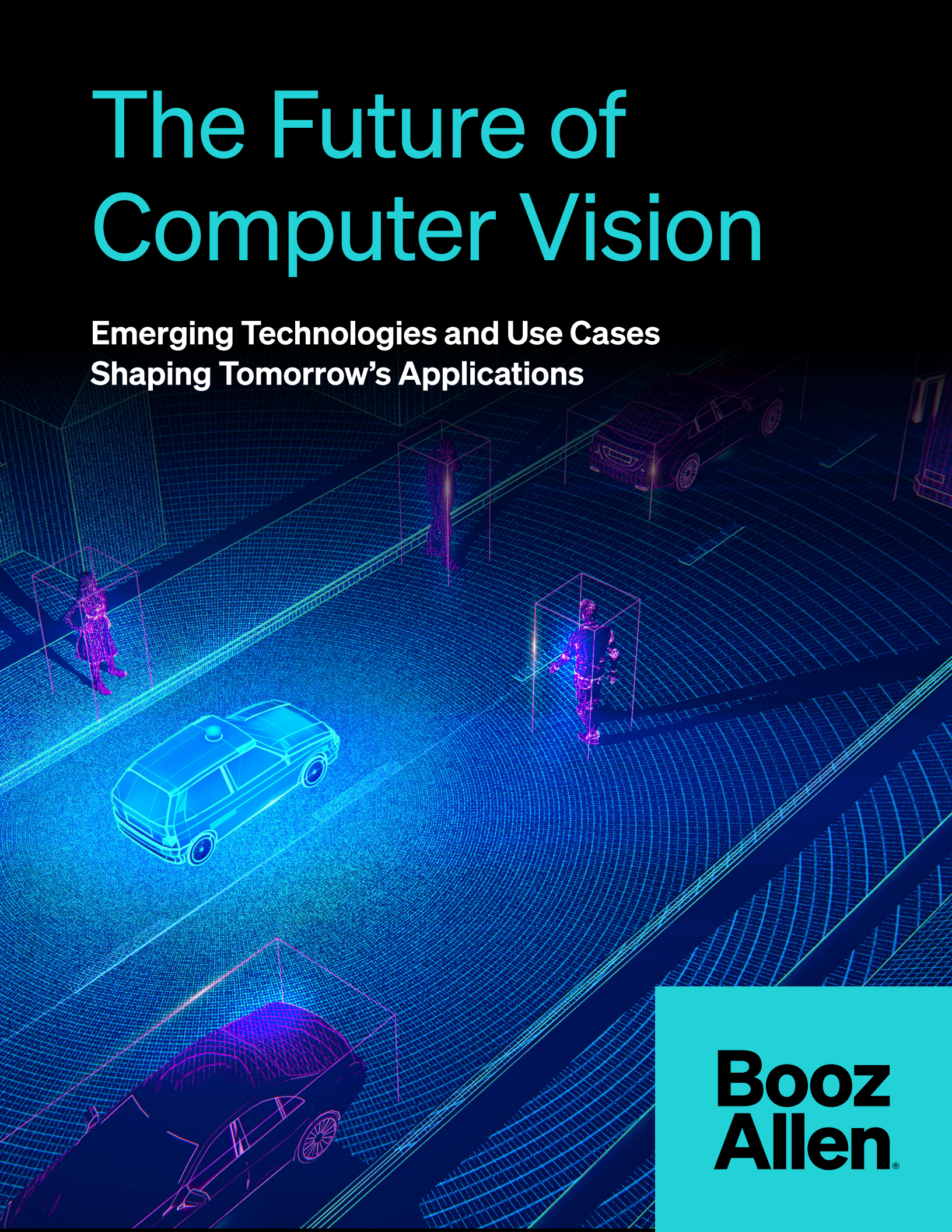
# The Future of Computer Vision

## Emerging Technologies and Use Cases Shaping Tomorrow's Applications



Booz Allen®

# Table of Contents

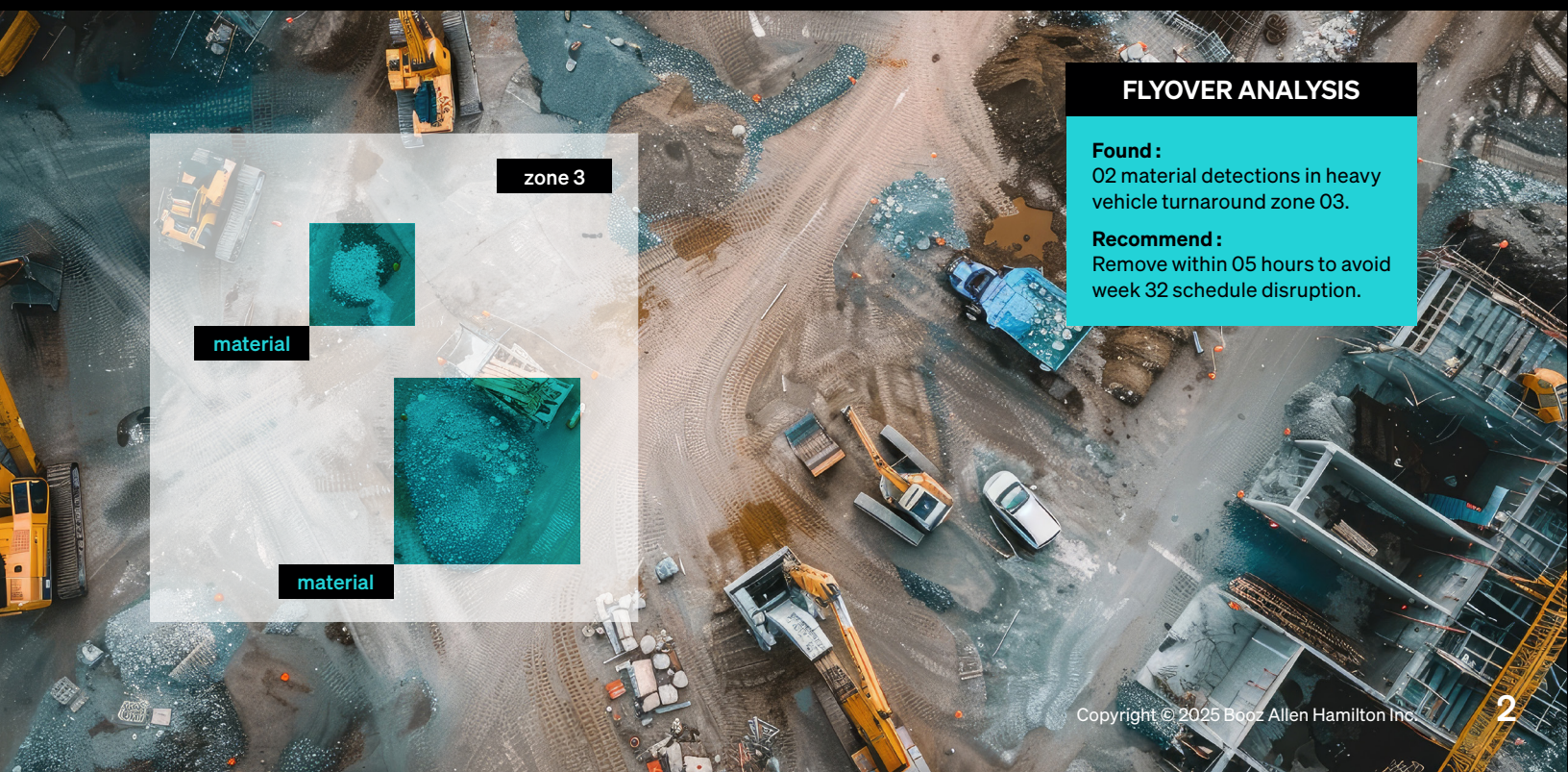
# Overview

Computer vision is easy to take for granted, given how reliable and adaptive it has become. Travel down the highway, and your toll payment account is correctly debited thanks to a computer vision tool that classifies vehicles by size and type. But this day-to-day ubiquity can make it more difficult to recognize the transformative power of this type of artificial intelligence (AI). Computer vision is unique and valuable largely because it enables computers to see and experience the world in much the way humans do. As we will explore in this report, recent technology advances are ushering in an era of enhanced performance and insight—a potential golden age for computer vision—for enterprises prepared to seize the moment.

Computer vision leverages machines to find, decipher, and extract meaning from representations of the rich visual and other electromagnetic information present in the world around us. It is an area of AI that has already demonstrated significant impact while holding potential for new innovations. The "advanced seeing" embodied through computer vision uses machine learning (ML) algorithms and neural networks to recognize and process data rapidly and in large volumes, enabling human operators to focus on high-value tasks. The increasing accuracy, robustness, and diversity of today's computer vision systems are pushing them to the forefront of many mission applications.

Increasingly, computer vision is moving beyond its traditional focus on identification and classification to encompass more complex assessment and analysis tasks. For example, traditional computer vision typically identifies objects within images and puts them into categories, such as distinguishing between dogs and cats in photographs. By contrast, new applications can assess a scene's context, detect subtle anomalies, and even predict future events based on visual data.

As a result, while a traditional application that focused on infrastructure resilience might have helped locate bridges, roads, and buildings in satellite imagery, a new application might detect wear and tear in a bridge span and provide rich data insights that decision makers can use to plan maintenance. The shift to more robust functionality is being driven by many technological changes that are accelerating alongside the rest of the AI universe, meaning that computer vision's importance and ubiquity are poised to grow even further.



# Computer Vision: State of Play

Like human eyes perceiving light, computer vision systems interpret image pixels for further processing and analysis, bolstered by powerful computers and algorithms. But they also supercharge human capabilities by finding and processing visual and other stimuli beyond what human beings can otherwise detect. Incorporating established knowledge about people, environments, and objects as well as the laws of physics, these modern systems are highly accurate, saving time, effort, and focus for human operators by efficiently performing tasks like image recognition, semantic segmentation, and object detection.

Computer vision applications support real-time processing for enhanced situational awareness in the digital battlespace, where low latency and high precision are critical, and for facial recognition tools, which require robust algorithms that enable feature extraction and matching. Applications extend to healthcare diagnostics through close analysis of medical images, precision agriculture with crop monitoring and disease detection, and smart checkout systems for the retail enterprise, among countless other use cases across the public and private sectors.

With ongoing advances in supercomputing, Internet of Things (IoT) systems and edge devices, AI integration, and fast 5G networks, computer vision continues to evolve from automation and identification toward far deeper insight and analysis. This evolution is like moving from a single camera that captures images to a sophisticated, multilayered optical system—one that reads and gains insight from the context of those images much as the human brain understands visual stimuli. As a result, computer vision can now deliver tangible impact whenever organizations need to obtain and use visible and unseen data as quickly as possible.

Influencing this progress is a shift from static ontologies to more dynamic AI systems. Static ontologies are predefined structures that cement relationships between concepts and objects. Often created manually, they lack flexibility when engineers are using novel datasets. By contrast, the latest systems leverage advances in the field of language understanding to become agile and adapt over time, with ML algorithms enabling consistent performance improvements and the delivery of unique outputs that are not tied to predetermined sets of classes.

For example, modern pretrained multimodal models combine language understanding with the ability to capture key optical concepts. They enable capabilities like zero-shot classification, which refers to a system that can recognize and categorize things it has never seen before without explicit training. In other words, these models can apply learned insights to categorize new, unseen classes, as opposed to operators manually constructing this "knowledge." While large language models (LLMs) have demonstrated the ability to learn concepts by ingesting and contextualizing massive amounts of text, visual and multimodal foundation models can similarly interpret and draw relationships between visual data, grounded in linguistic concepts.

No longer a siloed or one-dimensional application, computer vision has matured to become multimodal, multitemporal, and hyperspectral:

- **Multimodal** computer vision, which includes visual foundation models, processes and interprets diverse forms of information, such as text, photographs, video, and audio, providing a deeper understanding of context by integrating sensory data in imitation of human perception, from object and pattern recognition to depth perception and motion tracking.
- **Multitemporal** computer vision tracks changes over time, allowing for a better understanding of dynamic processes and events and improving the prediction of future states and behaviors.
- **Hyperspectral** computer vision captures and analyzes data across the electromagnetic spectrum, providing detailed understanding of informational units the human eye cannot see.

Together, these dynamics underscore the growing power of computer vision to solve real-world challenges in all industries by augmenting the sensing, processing, and interpretation of optical data to generate enterprise insight.

# Data, Software, and Hardware Convergence Drive Breakthrough Performance

Computer vision emerged in the 1960s, with early systems distinguishing shapes like circles, squares, and triangles and identifying simple patterns. However, overall performance was limited by a lack of computing power, minimal access to large datasets, and inflexible learning algorithms. Recently, innovations in data, software, and hardware have revolutionized the capacity of these systems, enhancing their accuracy, efficiency, and usefulness across a wide range of applications.

## Data Is Now Richer and More Ubiquitous

In the realm of computer vision, data is the essential fuel that powers the creation, training, and operation of models. Recent advancements in data management have made the right data more readily available, which has significantly enhanced the performance and utility of computer vision systems. The emergence of data standards and common storage paradigms has paved the way for the computer vision community to innovate faster.

### Development of Prelabeled Datasets

The wider availability of prelabeled datasets has streamlined the training of computer vision models. Labeled datasets provide a vast amount of annotated data that engineers can use to train models more efficiently. These datasets are often curated by experts and contain a wide variety of labeled images, which help in improving the accuracy and robustness of vision models. For instance, datasets like ImageNet and MS-COCO have become benchmarks in the field, enabling model training on diverse and extensive collections of images. The use of these datasets and of more recent examples, such as *LAION-400M*, *Ego4D*, *Objects365*, and *ImageBind*, reduces the time and effort required for manual labeling, allowing for quicker deployment of applications.

### Integration of Synthetic Data

Synthetic data is artificially generated data used to train ML models. It is particularly useful when real-world data is insufficient, expensive, or difficult to obtain. Synthetic data allows organizations to address new scenarios and unseen worlds or objects with computer vision, reducing the cost of developing new models. In defense and intelligence spaces, for example, this capability enables engineers to more accurately predict how current models will respond to new or hypothetical threats through simulating rare events and complex scenarios on the virtual battlefield.

Techniques such as 3D modeling and simulation and generative adversarial networks (GANs) are used to create realistic or stylized images. These methods enable the generation of diverse datasets, simulating edge cases that are difficult to capture with real-world data. Synthetic data also ensures privacy and compliance by avoiding the use of real-world data, which is crucial in sensitive fields like healthcare.

### Automation of Data Pipelines and Labeling

Increasingly, the workflow spanning data ingestion, preprocessing, transformation, and model training is being automated using specialized tools. These integrated data pipelines ensure that data moves efficiently through each stage of a computer vision system, minimizing manual intervention and repetitive tasks and leading to increased efficiency, scalability, and consistency in data processing.

As a key part of the data pipeline, data labeling focuses specifically on annotating images or videos with relevant information that identifies the objects, faces, or actions they show. While commercial, open-source, and academic labeled datasets remain

SYNTHETIC DATA TRANSLATION

important, they don't always meet the needs of specialized use cases with the requirement to train or tune models using tailored data inputs.

The labeling process is extremely time-consuming and expensive if done manually, especially for large datasets. As a result, organizations are approaching the point where it is no longer feasible to have hundreds of human workers label data. Automated data labeling uses techniques like pre-trained models, which leverage existing models to predict labels on new, unlabeled data; active learning, where a machine learning model asks for human input only for data points that are uncertain, reducing the number of labels needed; and semi-supervised learning, which combines a small amount of labeled data with a large amount of unlabeled data, allowing the system to automatically generate labels for the latter.

Through these techniques, automated data labeling significantly reduces the time, cost, and scalability challenges of manually labeling large datasets, with the result that small workforces can enrich their own datasets. It enables rapid annotation using pre-trained models, algorithms, or synthetic data generation. This automation cuts labor costs and scales efficiently, processing vast amounts of data without manual constraints. It also improves consistency and accuracy by reducing human error, ensuring uniform datasets for performant models.

## Software Gets More Intelligent

Computer vision applications combine integrated tools for managing and processing data with specialized algorithms for analyzing various media types. Given the diversity of media types and business applications, numerous combinations are often used to best address specific requirements.

### Algorithmic Improvements

The invention of convolutional neural networks (CNNs), together with implementation on graphics processing unit (GPU) hardware, represented a breakthrough in learning-based computer vision. CNNs provide a mathematical formalism for flexibly encoding spatially distributed information into the learned parameters of a model. Moreover, the fundamental operation of CNNs—that is, convolution—is shift-invariant, which allows the network to detect the same feature regardless of the location in the input. This property is crucial for tasks like object recognition, localization, and tracking, where it is important to classify patterns consistently regardless of their position in an image.

With the emergence of CNNs—such as widely used, highly performant models like You Only Look Once (YOLO)—formerly intractable computer vision problems like real-time and highly accurate image recognition and classification became manageable. Furthermore, these model architectures are well-suited for portable devices due to their lower resource requirements and are still advancing today. Nevertheless, CNNs have strict limitations. Without resizing or pooling, they require a fixed input image size. In addition, they have a relatively small receptive field, which is the specific area of an image that the vision application focuses on as it assesses what it is seeing.

In today's most advanced computer vision research, however, vision transformers (ViTs) have largely taken center stage, overshadowing CNNs in many leading-edge discussions—even though CNNs like YOLO remain widely used and highly effective. Fundamentally different than CNNs, ViTs divide the image into patches and convert them to feature representations called "tokens." Each feature is then combined with every other feature using a self-attention mechanism and passed to a feedforward network. Because the entire image is ingested by the model, ViTs do not suffer from a limited receptive field. As a result, they are extraordinary context learners, meaning that they excel at interpreting and using all of the information an image contains.

As an added benefit, the self-attention mechanism in ViTs is nearly identical to those in LLMs used for natural language processing, a realization that has paved the way for a surge in multimodal AI development. With ViT architectures, engineers can supplement training with context provided by user descriptions or ingest multiple wavebands like visible, infrared, and radar in a unified way of "seeing" that is far more than CNNs can accomplish on their own. A downside is that ViTs require tremendous amounts of data to train, even by CNN standards.

Self-supervised learning has also emerged as a powerful innovation. Traditionally, deep learning models required large, labeled datasets for training. Self-supervised learning allows models to learn representations of images without extensive manual labeling by leveraging unlabeled data. This has led to breakthroughs in image classification, enabling models to generalize better across different datasets.

### Vision Foundation Models

Vision foundation models (VFMs) are further revolutionizing the accessibility and adaptability of computer vision. Often using the ViT architecture, these large, pre-trained models are engineered to tackle a broad spectrum of visual tasks, including image classification, object detection, segmentation, and even the generation of descriptive captions for images. VFMs allow enterprises to readily adapt the


most powerful computer vision models to their unique requirements, often as easily as tapping into a cloud-based application programming interface.

Like LLMs, VFMs are trained on datasets comprising vast amounts of image data, enabling them to harness rich, nuanced representations of visual content. VFMs can interpret images and describe what is happening within them, moving toward reasoning capabilities and the handling of more complex cases like identifying suspicious behavior or understanding interactions in a scene. The generalization capabilities of VFMs extend beyond their initial training tasks to excel in various other visual applications. For example, these models have demonstrated strong zero-shot capabilities in segmentation, indicating their ability to adapt to new tasks.

A specific subset of VFMs, vision language models (VLMs) are transforming multimodal learning by integrating visual and text-based data to execute complex tasks. VLMs typically include an image encoder, a text encoder, and a fusion mechanism that combines these modalities to generate coherent outputs.

Combining deep learning models for both images and text has led to multimodal architectures like Contrastive Language-Image Pretraining (CLIP), which significantly improves zero-shot learning. This capability allows models to efficiently understand and relate images and text simultaneously, opening new possibilities in image generation. CLIP-enabled applications can create outputs based on concepts they haven't been explicitly trained on, offering versatility to support missions that require precise decision making in unpredictable, rapidly shifting operational environments.

Models like CLIP use large datasets to learn relationships between images and text, enabling image captioning, visual question-answering, and text-guided image generation. By blending visual and linguistic data, VLMs provide a basis for more intuitive AI systems that can seamlessly generate multimodal content.

In the area of optical character recognition (OCR), for example, VLMs simplify processes by leveraging zero-shot learning capabilities. Unlike traditional OCR systems that require extensive training on diverse datasets, VLMs generalize well to new, unseen data scenarios without exhaustive retraining. This agility supports applications such as automated document processing and enhanced accessibility features for digital platforms.

## Faster, More Efficient Hardware

The continued development of GPUs and tensor processing units (TPUs) has also democratized computer vision by making it possible to run deep learning models faster and more efficiently. These specialized processors allow for the parallel computation needed to handle many operations simultaneously, streamlining the process of training models on massive datasets. TPUs optimize the tensor operations that underlie neural network computations specifically. GPUs and TPUs are readily available as cloud-based resources, making it easier for organizations to get started using advanced hardware without a significant investment for on-premises/physical infrastructure. Advances in memory and storage technologies, such as high-bandwidth memory and non-volatile memory express storage, have further accelerated data access speeds and computational throughput.

Catalyzed by hardware and algorithmic breakthroughs alike, lightning-fast inference allows vision systems to process images and video streams in real time. This capability is critical for certain healthcare applications, such as when instant image analysis can inform surgical procedures as doctors perform them. Such improvements make computer vision systems more reliable and versatile in high-stakes environments including security and defense. However, despite the advancements in hardware, more powerful models like VLMs require massive amounts of compute and memory to train and for inference. Ultimately, the expansion of a wide range of capabilities—from data, software, and hardware to processing capacity, storage, and more—will continue to be essential to drive future progress in computer vision.

# Looking at the Future

Computer vision has undergone a remarkable transformation in recent years, propelling the technology beyond mere classification and identification and enabling more sophisticated analysis and prediction capabilities. However, capitalizing on this potential requires imagination and the ability to assess the status quo with fresh eyes. Here we explore the implications of these trends for four key scenarios—edge computing, multimodal AI, generative AI, and virtual reality (VR) or augmented reality (AR)—focusing on their potential impact on federal government operations and mission performance.

## The Edge Paradigm

Computer vision is increasingly moving to the edge, where sensors operate and data is processed locally on devices like smartphones and IoT devices. At the same time, a larger range of sensor hardware types—encompassing tools for gathering and processing LiDAR, radio frequency signals, synthetic aperture radar, and infrared data—is expanding the information streams available to deliver computer vision models. Overall, the edge paradigm offers significant benefits, such as reduced latency and enhanced privacy. However, it also presents several challenges, such as the need to address size, weight, and power (SWaP) and connectivity constraints. These challenges are significant for the warfighter in an increasingly digital battlespace, when the need to carry power-hungry equipment hinders mobility and when failsafe connectivity is needed for instantaneous decision making in denied, degraded, intermittent, and limited-bandwidth environments.

Hardware advances, model optimization, and more efficient software solutions are helping organizations overcome these challenges. For example, dedicated edge AI chips—such as the Google Edge TPU coprocessor, NVIDIA Jetson™ modules, the Apple Neural Engine chip, and Qualcomm's Snapdragon systems on chips—run deep learning models efficiently on edge devices with low power consumption, enabling real-time image processing on resource-limited devices. Intel is adding neural processing units to new central processing unit

designs to accelerate AI tasks. Field-programmable gate array and application-specific integrated circuit technologies, although not general-purpose, provide custom, energy-efficient computing solutions optimized for specific vision tasks.

Quantization reduces the precision of model weights as a means to reduce GPU memory demands while minimally affecting model accuracy. Other model compression techniques, such as pruning, which removes unnecessary parameters, and knowledge distillation, which is a way to train smaller models using larger ones, similarly reduce the computational and energy demands of vision models without compromising performance. Memory and storage can be optimized using pared-down CNN architectures and other techniques to reduce model size. Specifically, models like MobileNet and EfficientNet effectively balance accuracy and size in edge environments. Even architectures as large and computationally expensive as ViTs are making their way to edge devices. These find their way into edge-specific frameworks like TensorFlow Lite and PyTorch Mobile and allow developers to deploy lighter versions of deep learning models tailored for devices with limited memory.

### Additional enhancements for edge applications include:

#### SECURITY CONSIDERATIONS

Assuming edge devices may end up in the wrong hands warrants the need for model encryption, watermarking, and safeguards against reverse engineering techniques. On-device processing strengthens security and data privacy by reducing the need to send sensitive data to the cloud, minimizing breach risks. Encrypted communication and secure hardware, like trusted execution environments or hardware security modules, further protect data and models at the edge.

#### CONNECTIVITY

To address connectivity limitations, edge devices use edge-native architectures for real-time processing to reduce or avoid the need for cloud connectivity. Edge-cloud hybrid models handle critical tasks locally and less time-sensitive tasks in the cloud.

#### DEPLOYMENT AND MAINTENANCE

Platforms like Amazon Web Services IoT Greengrass and Google Cloud IoT Edge simplify edge deployment and maintenance by coordinating remote updates across intermittently connected devices, while containerization tools like Docker and Kubernetes enable consistent software-defined deployments across diverse edge devices.

## Multimodal AI Grows AI Adoption

Recent advancements in computer vision technology—particularly with ViTs—have paved the way for enhanced use of multimodal data. Multimodal AI combines diverse text, imagery, audio, and video data sources for better decision making. The model can integrate these multiple data types to generate descriptive outputs and can infer missing modalities from available data when necessary. The fusion of data types also facilitates a more natural human-model interaction and offers a transparent means of interpreting the model's decision-making process.

For example, one of the key benefits of recent computer vision advancements is the ability to recognize and describe actions in videos. By combining language, vision, and video data, multimodal AI can convert visual information into searchable text. This capability makes it easier to find specific visual information based on text-based queries, significantly enhancing the usability of video data. In surveillance operations, this technology can automatically identify and tag actions, making it simpler to locate and analyze critical events.

In addition, integrating LLMs with vision models has led to more natural and intuitive interactions with AI systems. This integration allows robots and AI systems to understand and respond to verbal instructions in real time. Advances in grounded language learning enable robots to comprehend visual scenes using language, allowing them to perform tasks based on verbal commands, such as moving objects to specific locations. This capability is particularly valuable in logistics and supply chain management within federal operations.

Multimodal AI systems that combine data from various sources provide a more comprehensive understanding of complex scenarios. The ability to integrate visual data with other operational inputs, such as real-time communication or environmental data, enhances situational awareness. In crisis management, multimodal AI can analyze visual, textual, and sensor data to provide real-time recommendations during national emergencies. This capability enables government officials to make informed decisions quickly, improving response times and overall effectiveness.

## Generative AI Takes Shape

Generative AI broadly refers to models and techniques that can create new content—such as images, text, audio, or video—by learning natural distributions from a given dataset. It encompasses several models and algorithms, such as GANs,

diffusion models, and variational autoencoders, which can generate realistic content based on user prompts and patterns learned from training data.

Generative AI can enhance computer vision performance in several ways. For example, in image synthesis, generative AI is often used to create synthetic images for tasks such as augmenting data or producing synthetic datasets to train computer vision models. GANs, for example, can generate realistic images that may be visually indistinguishable from real ones.

Some data augmentation techniques use generative AI to convert one type of image into another. Techniques such as CycleGAN can transform sketches into realistic photos or day images into night images to provide additional training data that simulates different real-world conditions. Engineers also use generative AI for super-resolution and image enhancement to improve the quality of images, such as converting low-resolution satellite imagery into high-resolution versions in order to enhance surveillance capabilities, pinpoint disaster response, or improve the accuracy of environmental analysis.

The difference between computer vision's ability to identify objects in images and generative AI's ability to create images lies in the tasks each approach is designed to perform, the underlying processes, and the type of output generated. In computer vision, the primary goal of object identification is to detect and recognize specific objects within an image by analyzing visual features such as shape, color, texture, and patterns. The output is typically structured data, such as bounding boxes around detected objects, labels categorizing each object, and confidence scores indicating the model's certainty. For example, the output from a computer vision model could be {"object": "car", "position": [x1, y1, x2, y2], "confidence": 0.95}.

In contrast, one goal of generative AI is to generate textual descriptions of an entire image, capturing the context, describing relationships between objects, and providing a narrative summary. This goes beyond just identifying objects, aiming to produce a natural language description of what is happening in the image. For instance, the output from a generative AI model might be "A red car is parked at the curb, and a man is walking his dog nearby." Alternatively, the model could process this output as a new text prompt and generate the corresponding image.

Generative AI models also have a deeper contextual understanding of images compared to traditional computer vision systems. Instead of merely recognizing objects in isolation, generative AI can interpret the relationships between objects and their context within the scene, allowing for more accurate and detailed search results, particularly for complex queries where understanding the entire image is necessary.

Traditional computer vision search typically relies on matching image features like color, texture, or shape to find similar images. However, the underlying representations employed in generative AI enable semantic search, meaning engineers can use the meaning behind an image or query and retrieve images based on concepts rather than just visual features.

Generative AI can not only search for existing images but also generate new ones on the fly based on a query or prompt. Aiding image search, this capability allows users to create images that may not exist in the dataset but can be synthesized to fit the search criteria. This capability gives generative AI a significant advantage over traditional computer vision search, which is limited to retrieving images or grouping similar images already present in the database.

## Bringing Virtual Reality to Life

Advances in computer vision and generative AI are playing a critical role in making the metaverse and other VR, AR, or mixed-reality (MR) applications more realistic, immersive, and useful. Enhanced realism and immersive environments are achieved through object and scene reconstruction, where computer vision enables highly detailed 3D reconstructions. Techniques like photogrammetry and 3D scanning capture and import real-world objects and spaces into virtual environments with extreme accuracy, allowing users to explore lifelike digital spaces in VR/AR applications. Additionally, computer vision helps simulate light and shadow by analyzing real-world lighting conditions, ensuring that objects cast shadows and reflect light realistically, thus enhancing the overall realism of virtual spaces.

Computer vision is critical for hand tracking and full-body motion capture, enabling natural interactions within the metaverse or VR/AR applications. Users can manipulate objects, move through environments, and interact with virtual characters using gestures and body movements. Advances in pose estimation and motion tracking facilitate these interactions without the need for additional hardware like controllers, making the experience more intuitive. Facial expression recognition allows avatars in the metaverse to reflect users' real-time emotions, making virtual interactions feel more personal and engaging, which is especially important in social applications.

Computer vision enhances real-time object recognition and interaction, which allows for the recognition of real-world objects and their interaction with virtual elements in MR applications. AR applications can overlay digital information or tools onto physical objects, enabling seamless interaction between real and virtual worlds. Semantic understanding is another key aspect, as computer vision can analyze the context of real-world scenes and objects to provide relevant and actionable virtual overlays.

Precise spatial mapping of physical environments allows VR/AR devices to understand the layout of the real world. This is particularly important for creating MR applications where virtual elements must be integrated into real-world spaces without disrupting the physical environment, allowing users to navigate seamlessly between physical and virtual spaces.

For AR applications, organizations can also use computer vision for indoor navigation, where GPS signals are weak, by building accurate maps of indoor spaces and guiding users through virtual overlays in settings like malls, airports, or hospitals. Finally, computer vision allows virtual environments in the metaverse to simulate real-world physics, making interactions with objects more believable. Whether it is objects falling, bouncing, or reacting to user input, physics-based simulations add an important layer of realism.

# Building a Computer Vision Technology Platform

A key focus of this paper is documenting computer vision's march forward in terms of its capabilities, utility, and value. And, as previously discussed, ensuring that enterprises can stay at the forefront of these innovations requires building the right competencies and processes. It also requires the right infrastructure and architecture.

A dedicated computer vision technology platform can help organizations build upon these competencies to deploy optimized solutions that address unique requirements at mission speed. Relying on an open, extensible, adaptive architecture also streamlines the integration of these future innovations.

Additional advantages of an integrated, logically structured platform include the following:

- Allows for the automation of processes such as data collection, labeling, and model training, improving overall productivity.
- Eases the deployment of models to the cloud or edge, enabling scalability across multiple applications and environments.
- Enables management of everything from data ingestion to inference and monitoring in a single workflow, reducing friction between different stages and allowing for faster model iteration cycles.
- Reduces the need for extensive manual labor and infrastructure investments via automated data pipelines and transfer learning models, leading to more cost-efficient development.
- Supports the addition of new computer vision methods in a modular way, from object detection to segmentation or image classification, and enables deployment in diverse environments (e.g., edge, cloud), improving flexibility.
- Incorporates tools for monitoring, retraining, and improving models over time, ensuring that performance remains optimal as data evolves and new challenges arise.

## Platform Components

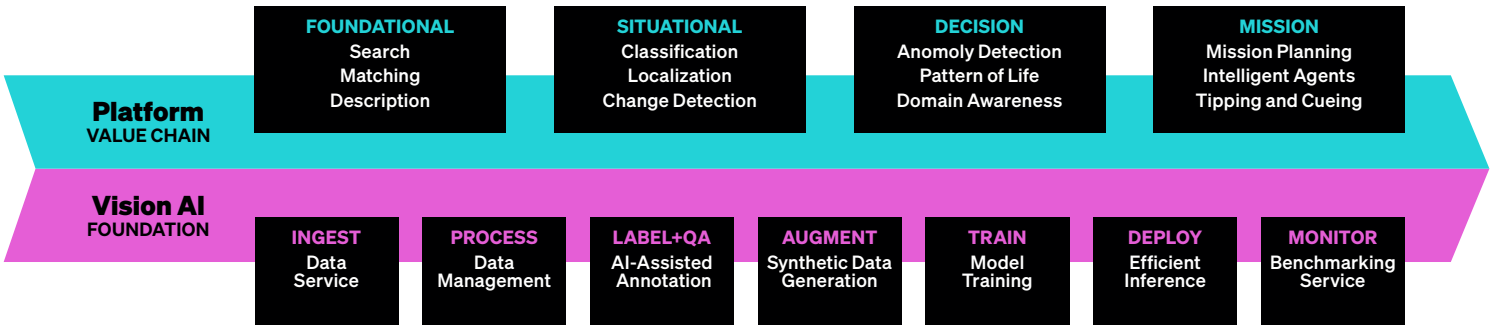
Key components of a computer vision technology platform include data collection, preprocessing, model training, inference, and deployment tools.

### DATA COLLECTION AND INGESTION

This component involves gathering images or videos from cameras, sensors, drones, or other devices, or generating synthetic data, and storing the data in cloud storage systems (e.g., AWS S3, Google Cloud Storage) or on-premises solutions. An extract, transform, load, or ETL, process can improve computer vision performance by standardizing and organizing file structure. Data labeling tools like Labelbox, Supervisely, or Computer Vision Annotation Tool assist in tagging objects for supervised learning models.

### DATA PREPROCESSING AND AUGMENTATION

Resizing, normalization, color correction, and noise reduction help ensure data consistency and readiness for model training. Preprocessing frameworks include OpenCV and Python Imaging Library. Data augmentation techniques, such as flipping, rotating, cropping, or adding noise, artificially increase dataset size and diversity, making models more robust to real-world variations. Libraries like Albumentations and Keras Preprocessing are commonly used for this step.



### MODEL DEVELOPMENT AND TRAINING

Model development and training are supported by deep learning frameworks like TensorFlow, PyTorch, and Keras, which provide the necessary tools to create architectures such as CNNs, ViTs, and GANs. Pre-trained models like ResNet, Inception, or YOLO can be fine-tuned for specific tasks, reducing the need for large amounts of labeled data and training time.

### SOFTWARE MODULARITY

Integrating modular, best-of-breed software components and AI technologies allows engineers to efficiently build, deploy, and operate computer vision models across various environments, such as the cloud and edge. Modular components simplify the process of adapting to new requirements, scaling to handle complex use cases, and integrating with other systems, reducing risk.

### HARDWARE

High-performance hardware, such as NVIDIA GPUs and Google TPUs, is essential for training deep learning models efficiently, especially when dealing with large datasets or real-time processing requirements. These hardware accelerators are optimized for parallel processing, which is critical for image data.

### MODEL INFERENCE AND DEPLOYMENT

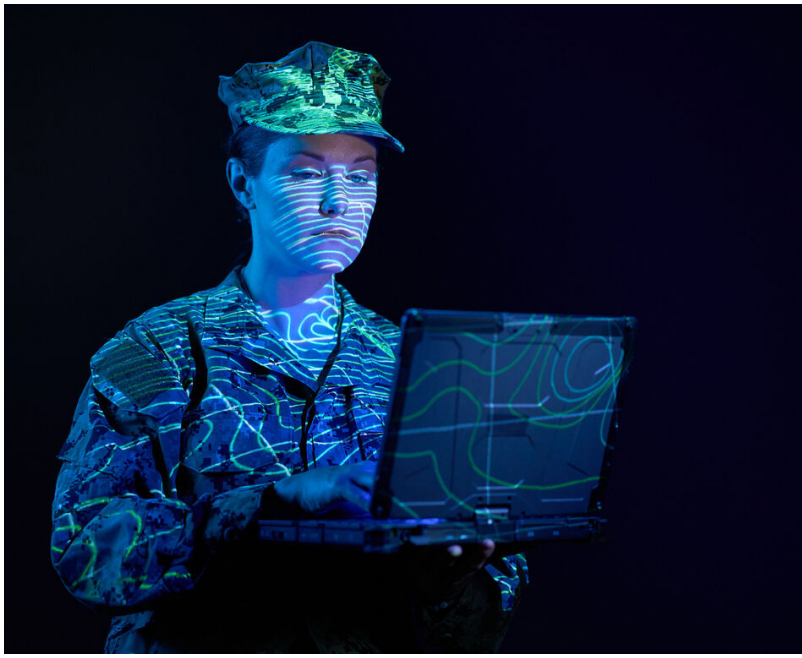
After training, models are deployed to perform inference on new data. Edge devices (e.g., mobile phones, IoT devices, drones) can use optimized models for real-time processing, while cloud solutions are typically used for more complex, resource-intensive tasks. Open neural network exchange models allow models to be portable across different platforms and devices. Tools like TensorFlow Serving, TorchServe, KServe, or AWS SageMaker are used to deploy models into production, supporting scaling to handle real-time requests and often integrating into larger systems through application programming interfaces.

### MONITORING AND FEEDBACK

After deployment, it's critical to monitor model performance in real-world scenarios. Tools like Prometheus and Grafana monitor metrics such as latency, accuracy, and performance drift. Continuous feedback loops allow for retraining models with new data to maintain accuracy over time.

### DATA MANAGEMENT AND GOVERNANCE

Just like code, datasets and models need to be versioned to track changes over time. Tools like data version control and Git are essential for establishing a logical governance flow that tracks changes in datasets and models over time. For privacy-sensitive data, tools ensure compliance with regulations like the General Data Protection Regulation and Health Insurance Portability and Accountability Act, protecting data during training and deployment. In addition, static code analysis tools such as SonarQube and container scanning tools such as Twistlock ensure that secure apps and containers are delivered in accordance with government and Department of Defense policies.



With a comprehensive computer vision technology platform, organizations can harness integrated software services to build, deploy, and operate advanced models in the cloud, secure environments, and the disconnected edge. Leveraging best-of-breed components for data ingestion, labeling, training, and optimization is the key to deploying faster at lower cost and risk while maintaining long-term extensibility. As organizations tap into these capabilities, they will strongly position themselves to capitalize on the shift from traditional computer vision to more sophisticated, context-aware, and interpretable models for increasingly complex missions.

## About the Authors


— **Michael Sellers, Ph.D.**, leads Booz Allen's Vision AI practice and consults extensively with the government on computer vision strategy.

— **Kevin Miller, Ph.D.**, is a senior lead scientist within Booz Allen's Vision AI practice, architecting advanced computer vision systems for strategic mission applications.

— **Sarah Shuhaibar** is a senior lead scientist within Booz Allen's Vision AI practice, leading the development of computer vision managed services and the integration of commercial technologies.

— **Rachel Lucas, Ph.D.**, is a senior lead scientist within Booz Allen's Vision AI practice, leading research and product development for radio frequency analysis and related electromagnetic spectrum solutions.

— **Drew Massey** is a senior lead engineer within Booz Allen's Vision AI practice, serving as the product lead for the company's Bighorn AI Kit™ and related edge and autonomous perception technologies.

— **Gawan Fiore** is a lead technologist within Booz Allen's Vision AI practice, focused on crafting advanced computer vision concepts into bespoke solutions to mission applications.

— **Morgan Githinji** is a lead technologist within Booz Allen's Vision AI practice, serving as the product lead for the company's Vision AI solutions.


## Contact


VisionAI@BAH.com

Booz Allen®

## About Booz Allen

Booz Allen is the advanced technology company delivering outcomes with speed for America's most critical defense, civil, and national security priorities. We build technology solutions using AI, cyber, and other cutting-edge technologies to advance and protect the nation and its citizens. By focusing on outcomes, we enable our people, clients, and their missions to succeed—accelerating the nation to realize our purpose: Empower People to Change the World®.

**BoozAllen.com**