Scaling Computer Vision at the Edge

From Prototype to Plant Floor

Eric Danziger

CEO + Co-Founder

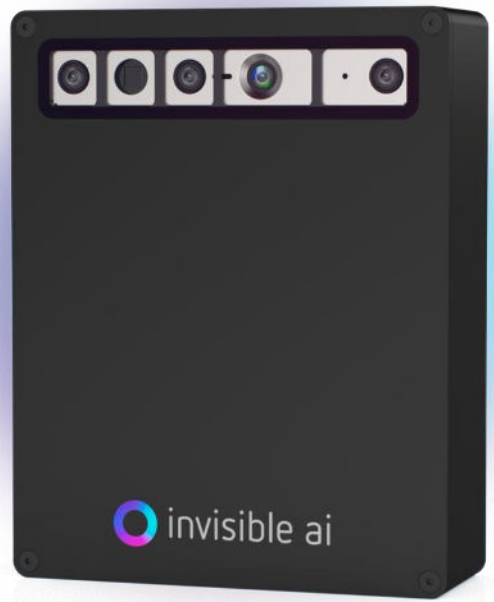
Invisible AI

invisible ai

Invisible AI: Computer Vision & AI to Accelerate LEAN Processes in Manufacturing

Invisible AI tracks what is happening on the factory floor and provides analysis for manual assembly optimization.

Our devices are primarily deployed at automotive OEM facilities globally.

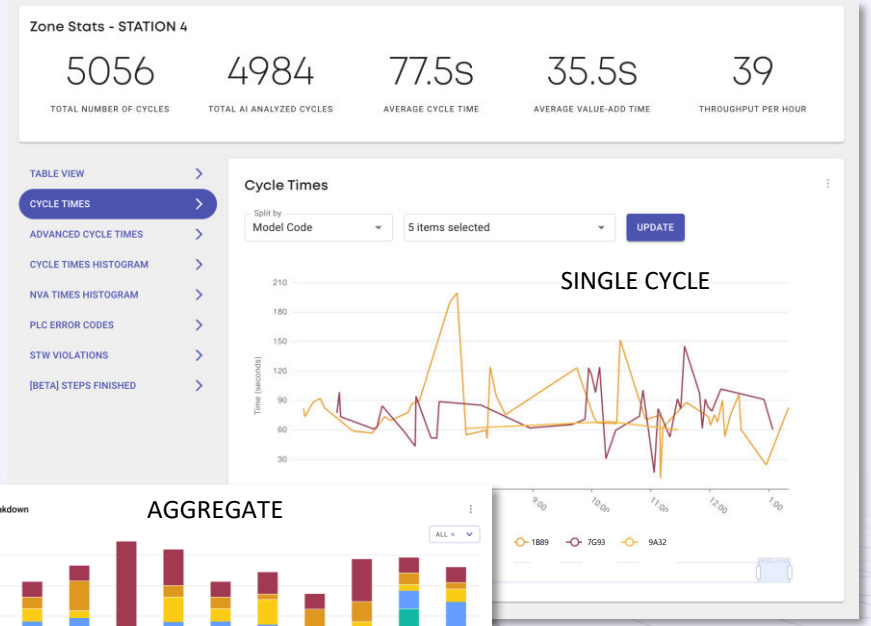


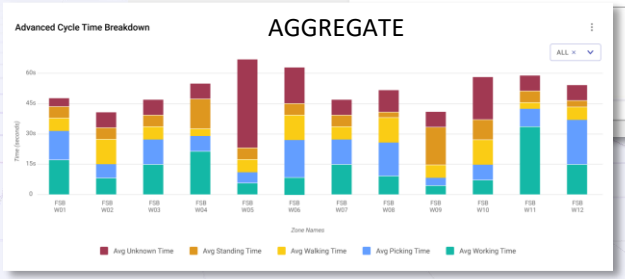
invisible ai


4.1m
Station Bounding Box

Automated time and motion studies on *every* cycle

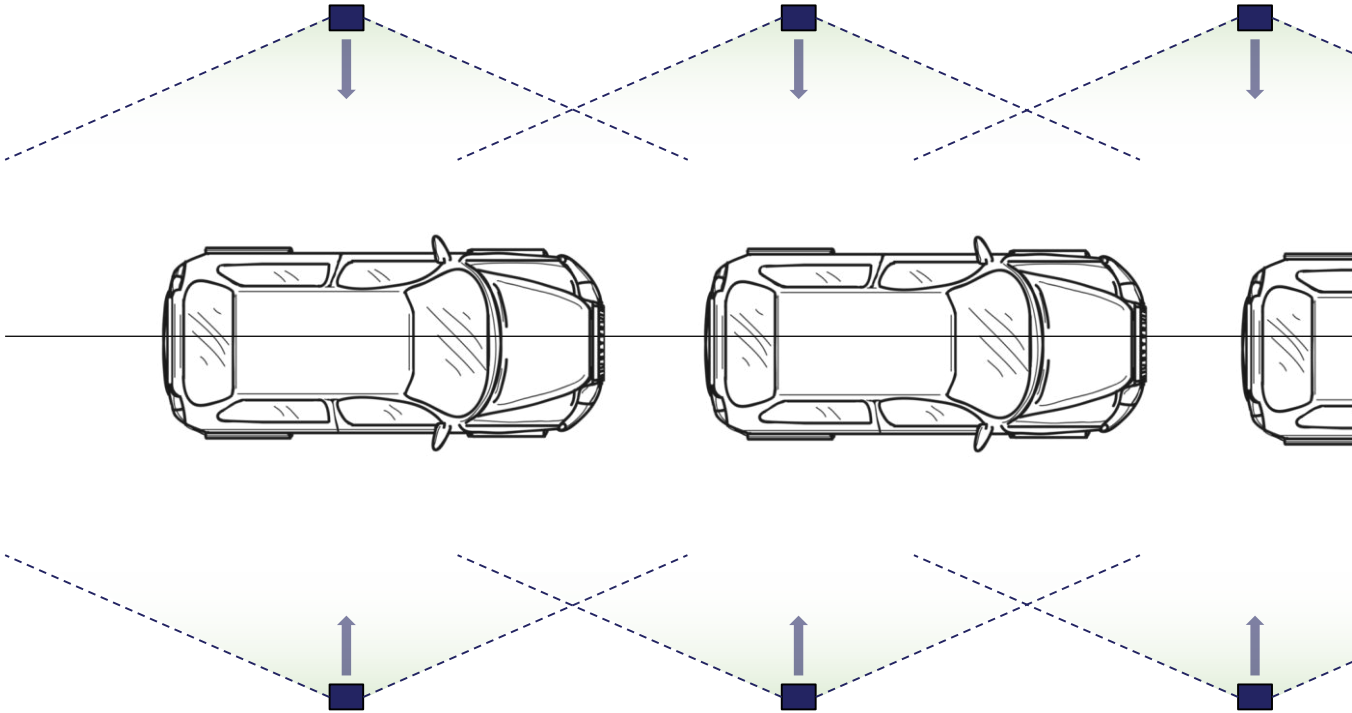
- Break down every cycle into work, walk, wait, pick, and unknown time.
- Identify stations that are under or overloaded with work for line-rebalancing.
- Eliminate clipboards and stop watches – removing operator bias.
- **Quickly** compare cycles pre- & post-kaizen activity, or across different shifts





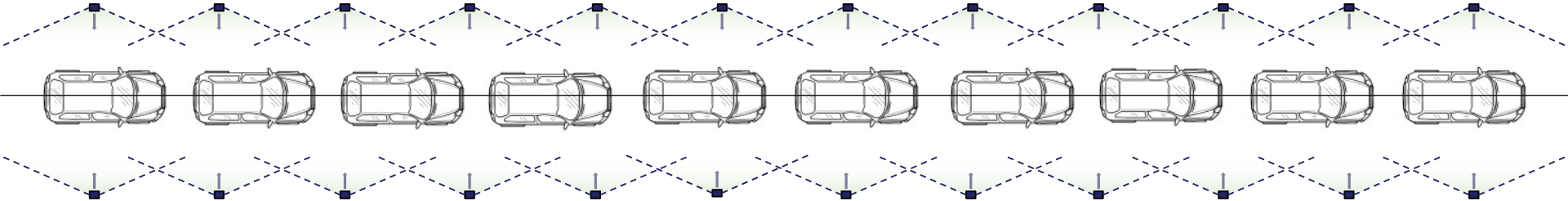
invisible ai

One Station

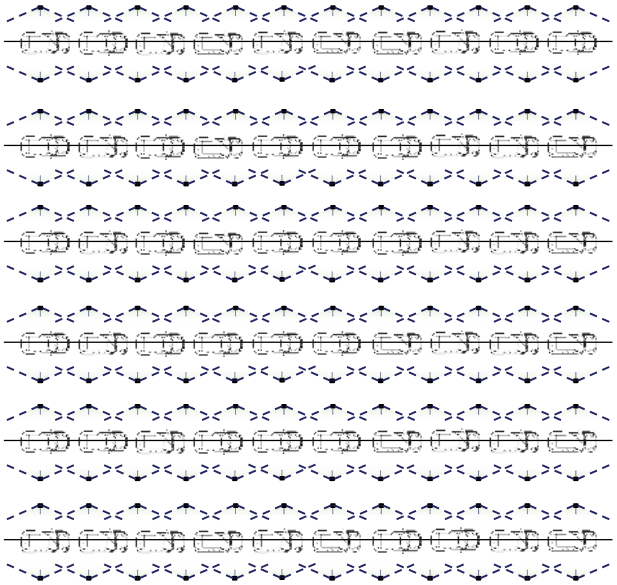


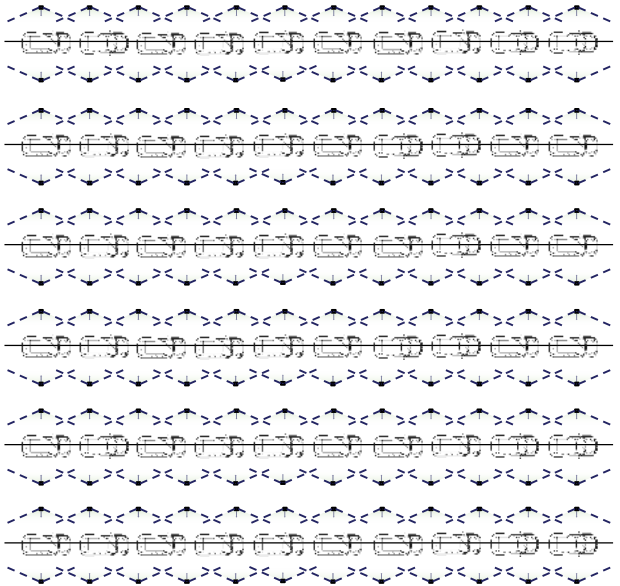
invisible ai

One Line

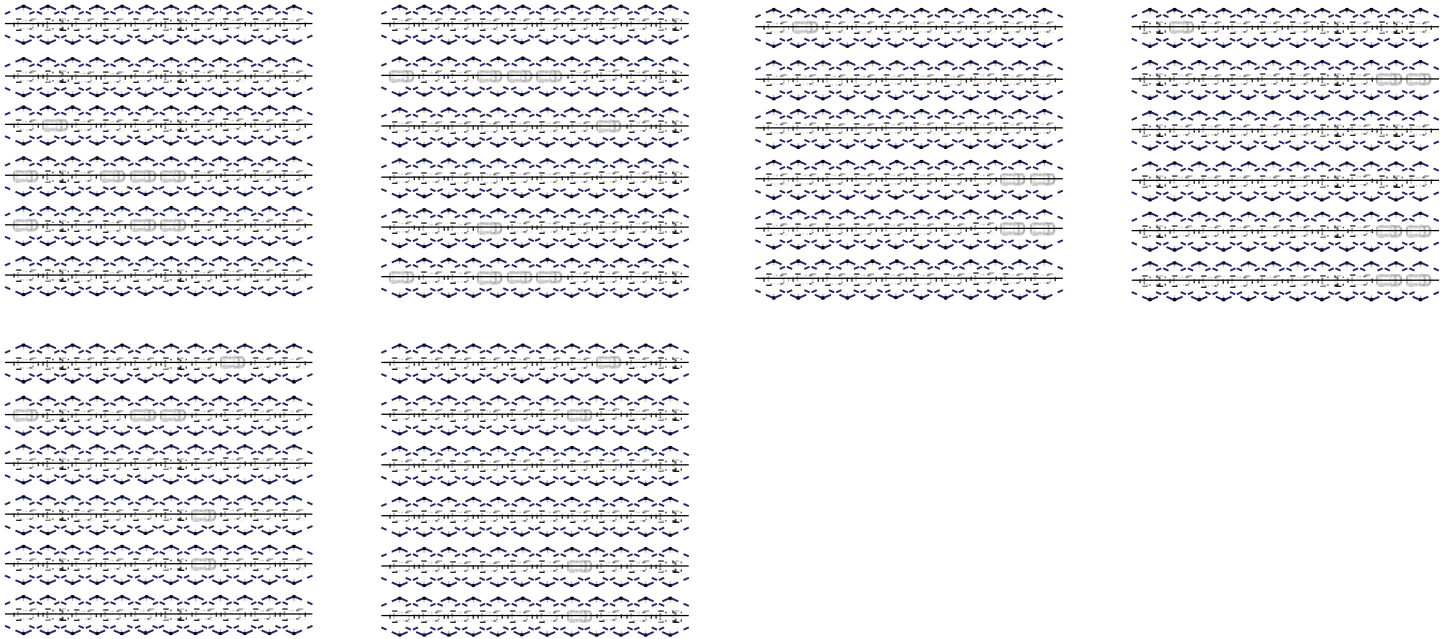


One Assembly Shop





One Facility



Mastering Scalable Vision Systems

An Invisible AI Case Study Through Three Dimensions

1.
Infrastructure

2.
Capability

3.
Deployment

Infrastructure

The Lab Demo – Just Show it Working Once

- Lab conditions
- Full bandwidth
- Constant engineering handholding

It was never going to survive the real world.

The Pilot

"Can you install this in a plant next month?"

Pilot Reality

- Workstations – 4
- Devices – 8
- Bandwidth needed to stream – 40 Mbps
- Bandwidth available – 10 Mbps

If we want to cover the plant, we need to be able to 400x the number of devices.

Infrastructure Lessons: One Device, All On-Board

- **Single-SKU edge node** — camera, depth, GPU, storage, networking in one rugged housing; order once, deploy anywhere, scale linearly
- **On-device inference & retention** — raw video never leaves the floor, only encrypted events/telemetry flow out → bandwidth drops 100× and links can't throttle production
- **Built-in privacy & compliance** — data stays on-prem, satisfying legal requirements
- **Uniform fleet lifecycle** — identical firmware, spares, and OTA updates cut maintenance cost and mean every feature ships to *every* node overnight
- **Engineering hurdle** — CV stack must squeeze into the node's TOPS, memory and power budget → aggressive quantization, pruning, and heterogeneous pipelines are mandatory

Capability

Pilot Reality Check: “Process Changed — Models Didn’t”

- **Pilot is now deployed –** data is labeled, models are trained, and we start getting data from each cycle.
- **Day 30:** plant updates torque sequence and moves a jig; cameras stay put.

The labels are now wrong.

- Edge hardware handled it perfectly — bandwidth, privacy, uptime **unchanged**
- The *model* was the bottleneck: needed fresh data, new labels, re-training → days or weeks of cycle-time blind-spots
- Operators asked, “How many new examples does it need??”

Three Levels of CV Capability

Level 1	Level 2	Level 3
Hot Dog or Not Hot Dog?	Interactive Vision Systems	General Purpose AI Vision
Pre-Training: Millions of Images	Pre-Training: The Entire Internet	Pre-Training: 
Model Size: 1-100+ MB	Model Size: 1-100+ GB	Model Size: Unknown
Task Training: 100s-10000s of Examples	Task Training: Only Needed for Non-Domain Use Cases	Task Training: None

Don’t Train the Model – Just Tell It What To Do
Our Shift to Prompt-Based Computer Vision

- Inspired by LLM prompting
- Decouples model from specific tasks
- Easier setup, faster reconfiguration

“Show me the station with the most ergo violations last shift”

“Show me the station with the highest walk distance”

“Show me the stations with the most variable cycle times”

What Prompting Unlocks

- **Setup in hours:** configure prompts, not data-label sprints
- **Adapt in minutes:** process change → revise prompt, redeploy instantly
- Customers can configure devices without engineering support
- Edge device time to **ROI is measured in minutes** not weeks

Deployment

Now Scale It 200x

Key Challenges:

- Engineers needed for every installation and adaptation
- Each on-site work change required physical reconfiguration
- Bandwidth limitations throttled system performance
- Training operators on complex interfaces slowed adoption
- Value delivery took weeks/months instead of hours/minutes

Each Unit Is A Self-Contained Node

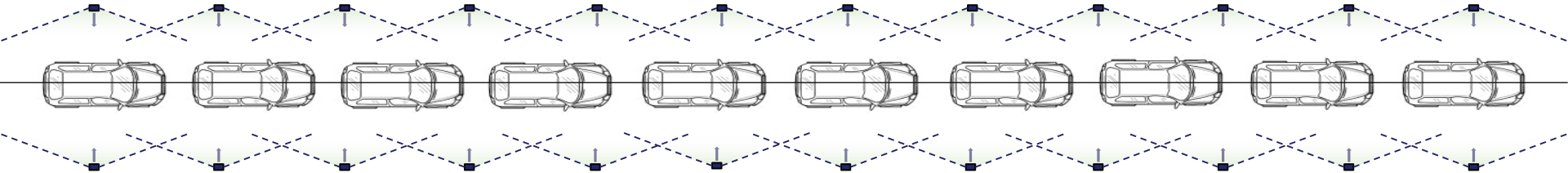
Atomic Devices, No Central Brain

- Compute + Storage onboard
- Identical units = plug and play
- Easy to swap, scale, maintain

invisible ai

Designing for Space, Not Tasks
The Scene, Not the Station

- 3D devices cover physical areas
- Processes within coverage = no device movement
- Just reconfigure the software



Auto-Adoption, Auto-Spatialization, Zero Setup

Deployment That Deploys Itself

- Regular spacing
- Self-registering devices
- Auto-extrinsics
- Fully unsupervised install
- 400 devices in less than a month

Conclusions

- “Quantity has a quality all its own”
  – Unnamed Automotive Engineer
- **Some problems only matter if they reach scale—so we design for 1,000+ nodes first, then work backwards**
- **Atomic edge nodes = zero racks, zero bottlenecks**
- **Space-centric design → tweak software, not hardware**

Questions?

invisible ai

Resources

Invisible AI:

- **Website:** https://invisible.ai
- **LinkedIn:** https://www.linkedin.com/company/invisibleai
- **Email:** sales@invisible.ai

Pat Gelsinger's Laws of the Edge:
https://www.youtube.com/watch?v=hOQ3uKcHXgo