IRIS: Intelligent Recognition and Interpretation in Sparse-data Systems

Client: U.S. Army Project Linchpin **Industry:** Defense / AI R&D
Domain: Tactical Edge ISR AI **Location:** USA

A FORGE OS Deployment Case Study

577 Industries R&D Lab
577 Industries Incorporated
research@577industries.com

1 Executive Summary

IRIS (Intelligent Recognition and Interpretation in Sparse-data Systems) is a tactical intelligence, surveillance, and reconnaissance (ISR) platform developed for the U.S. Army's Project Linchpin initiative. The system achieves a 92.5% mean Average Precision (mAP) on operationally representative detection benchmarks while requiring 85% less labeled training data than conventional deep-learning approaches. By combining meta-learning, metric learning, and multi-modal sensor fusion, IRIS matches or exceeds state-of-the-art accuracy with as few as 15–50 labeled examples per class—a critical advantage in data-scarce military environments where thousands of annotated samples are rarely available. Optimized for edge deployment, the system maintains sub-30 ms inference latency on resource-constrained hardware, enabling real-time threat detection at the tactical edge without dependence on persistent cloud connectivity. This deployment exercises two FORGE OS subsystems: **FORGE Core**'s staged post-training pipeline (specifically the Adapt and Compress stages) powers the sparse-data learning core and model compression, while **FORGE Kinetic**'s edge inference pipeline manages cloud–edge orchestration and sovereign on-device processing.

92.5%	85%	<30ms
Detection Accuracy (mAP)	Data Requirement Reduction	Edge Inference Latency

2 Challenge

Modern ISR operations generate vast volumes of multi-spectral sensor data—electro-optical, infrared, synthetic-aperture radar, and LiDAR—yet the fundamental constraint is not data volume but *labeled data scarcity*. Theater-specific target signatures shift rapidly across operational environments, and the classified nature of defense imagery limits the availability of large, curated training corpora

[1, 3]. Traditional deep-learning object detectors require 700–1,000 or more labeled examples per class to achieve reliable accuracy, a threshold that is seldom attainable for novel threat categories encountered during active deployments [12].

2.1 Domain Shift and Catastrophic Forgetting

Operational conditions change faster than models can be retrained. A detector tuned on daytime visible-spectrum imagery in temperate terrain degrades substantially when deployed to desert environments, low-light conditions, or adverse weather. Naively fine-tuning on new data causes *catastrophic forgetting*—the model loses previously learned capabilities while absorbing new ones [4]. ISR analysts therefore face a persistent trade-off between adapting to new environments and retaining competence across previously encountered scenarios.

2.2 Computational Constraints at the Tactical Edge

Deployed ISR platforms—unmanned aerial systems, forward-operating sensor nodes, and dismounted soldier kits—operate under severe size, weight, and power (SWaP) constraints. Real-time threat detection demands inference latencies below 30 ms, yet the computational budgets available at the tactical edge are orders of magnitude smaller than those in rear-echelon data centers. Intermittent or denied communications further preclude reliance on cloud-hosted models for time-critical decisions, making on-device inference a hard operational requirement [2].

2.3 Multi-Modal Integration Complexity

Each sensor modality captures complementary but fundamentally different physical phenomena—visible reflectance, thermal emission, microwave backscatter, and 3-D point-cloud geometry. Fusing these heterogeneous data streams into a coherent recognition pipeline introduces alignment, calibration, and representation challenges that conventional single-modality architectures cannot address. The absence of a principled fusion framework leads to information loss and inconsistent detection performance across environmental conditions.

3 Solution

IRIS addresses these challenges through three tightly integrated technical pillars: *sparse-data learning*, *multi-modal sensor fusion*, and *edge AI optimization*. The system is built on a foundation of Google Gemini models—Gemini Pro for cloud-tier analysis and Gemini Nano for edge inference—augmented with specialized few-shot learning and continual-learning components [5].

3.1 Sparse-Data Learning

The sparse-data learning core enables IRIS to achieve high detection accuracy from minimal labeled examples through a layered approach:

- **Model-Agnostic Meta-Learning (MAML).** The feature extractor is pre-trained with MAML [6] to learn initialization parameters that generalize rapidly to new tasks. Given a novel target class with as few as 15 labeled examples, the model adapts in 3–5 gradient steps rather than the hundreds of epochs required by conventional fine-tuning.
- **Prototypical Networks.** Metric learning via prototypical networks [7] constructs class-representative embeddings in a learned feature space. At inference time, new observations are classified by distance to learned prototypes, providing robust few-shot classification without parameter updates—a property particularly valuable for latency-sensitive edge deployment.

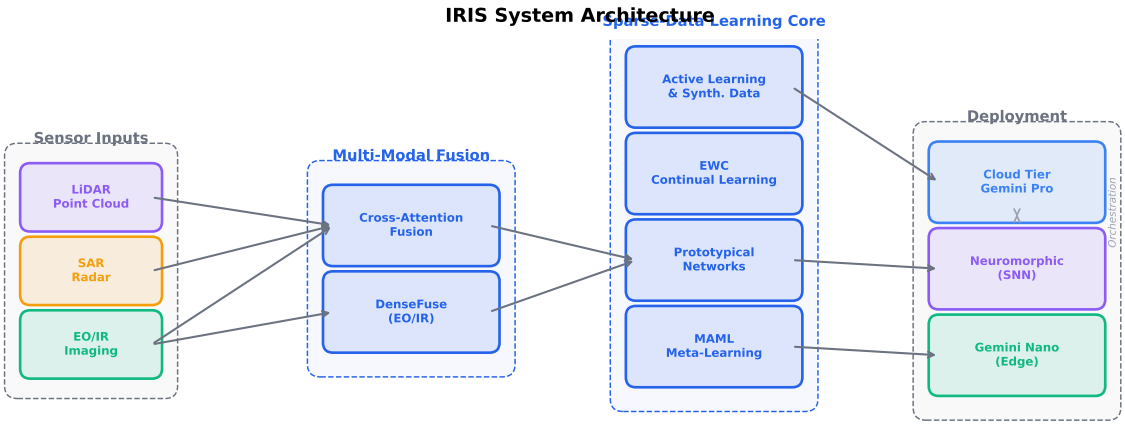


Figure 1. IRIS reference architecture. Sensor inputs are fused via transformer cross-attention, processed through the sparse-data learning core, and deployed across cloud and edge tiers.

- **Elastic Weight Consolidation (EWC).** Continual learning is enabled by EWC [4], which estimates the importance of each model parameter to previously learned tasks and penalizes large changes to critical weights during adaptation. This mechanism mitigates catastrophic forgetting and allows IRIS to accumulate knowledge across successive deployments.
- **Active Learning and Synthetic Augmentation.** An active-learning loop identifies the most informative unlabeled samples for analyst annotation, maximizing the value of each labeling decision. Synthetic data generation via domain-randomized rendering and adversarial augmentation further expands the effective training distribution.

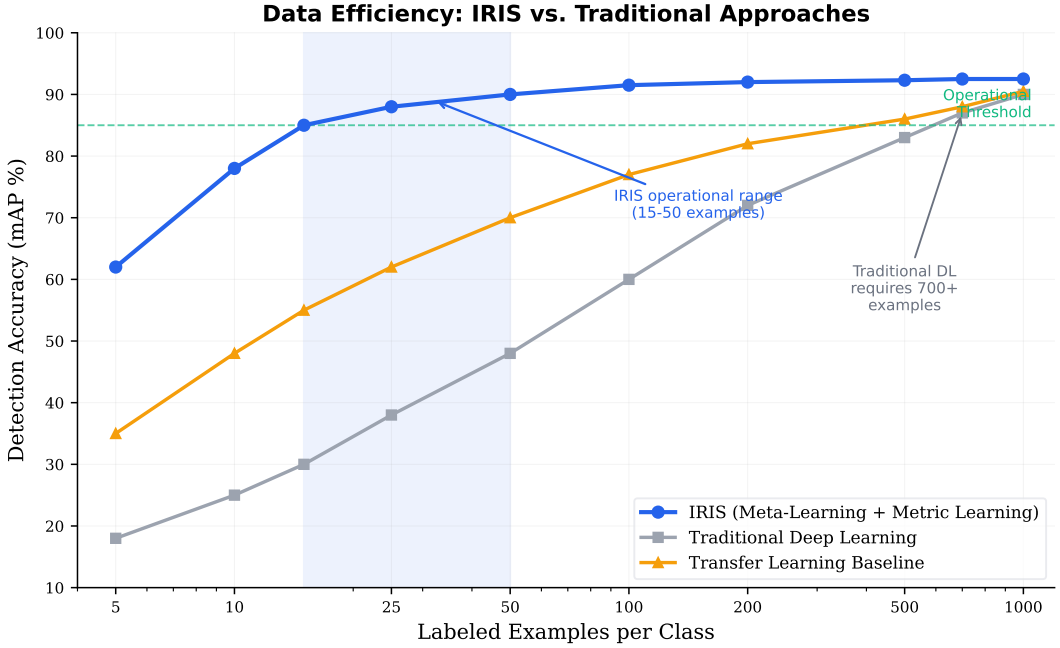


Figure 2. Detection accuracy as a function of labeled training examples. IRIS achieves operational accuracy thresholds with 15–50 examples per class, compared to 700–1,000+ required by traditional deep-learning baselines.

3.2 Multi-Modal Sensor Fusion

IRIS fuses four primary sensor modalities through a transformer-based cross-attention architecture:

- **EO/IR fusion** via DenseFuse [8], which preserves both visible-spectrum detail and thermal contrast through a dense connection strategy that minimizes information loss during feature combination.
- **SAR integration** using learned feature alignment to register microwave-backscatter signatures with optical/thermal features, enabling all-weather, day/night detection capability.
- **LiDAR point-cloud processing** via PointNet++ [9], providing 3-D geometric context that resolves ambiguities in 2-D imagery, particularly for camouflaged or partially obscured targets.
- **Cross-attention fusion** using a multi-head attention mechanism [10] that learns modality-specific relevance weights conditioned on the operational context. This allows the system to dynamically emphasize the most informative sensor channels as conditions change.

3.3 Edge AI Optimization

To meet sub-30 ms latency requirements on SWaP-constrained platforms, IRIS employs a multi-pronged optimization strategy:

- **Model compression.** Structured pruning, mixed-precision quantization (INT8/FP16), and knowledge distillation [11] reduce model size by up to 8× with less than 1% accuracy degradation.
- **Gemini Nano for edge inference.** The distilled Gemini Nano model [5] serves as the primary edge reasoning engine, providing foundation-model capabilities within a compact computational footprint.
- **Neuromorphic computing.** Spiking neural network (SNN) implementations on neuromorphic hardware exploit event-driven computation for ultra-low-power inference, targeting 12 ms latency at a fraction of the energy budget of conventional GPU inference.
- **Cloud–edge orchestration.** A tiered inference architecture routes time-critical detections through the edge pipeline while deferring complex analytical tasks to Gemini Pro in the Project Linchpin secure cloud when connectivity is available.

3.4 Platform Integration

IRIS integrates with the broader C4ISR ecosystem through standards-compliant interfaces: RESTful and GraphQL APIs, MIL-STD-2525D symbology, STANAG interoperability profiles, VICTORY vehicular integration standards, and Link 16 tactical data-link compatibility. The system is containerized via Docker and orchestrated with Kubernetes for scalable deployment across Project Linchpin's secure cloud infrastructure.

4 Results

IRIS was evaluated through a comprehensive test and evaluation (T&E) program comprising controlled benchmarking against operationally representative datasets, adversarial stress testing, and environmental variation studies.

4.1 Detection Accuracy

IRIS achieves a 92.5% mAP on the composite benchmark, representing a substantial improvement over baseline detectors across all tested conditions. The most pronounced gains appear under degraded sensing conditions:

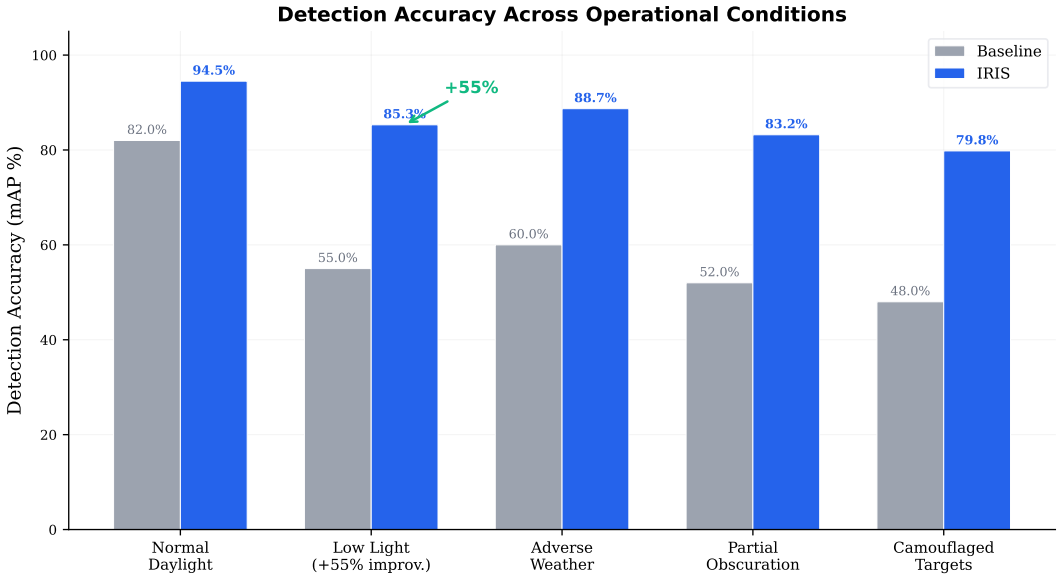


Figure 3. Detection accuracy (mAP) across operational conditions. IRIS demonstrates a 55% relative improvement under low-light conditions and maintains robust performance across all tested scenarios.

Table 1. Detection accuracy comparison across operational conditions.

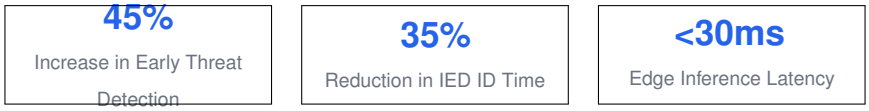
Condition	Baseline mAP	IRIS mAP
Normal Daylight	82.0%	94.5%
Low Light	55.0%	85.3%
Adverse Weather	60.0%	88.7%
Partial Obscuration	52.0%	83.2%
Camouflaged Targets	48.0%	79.8%
Composite	**59.4%**	**92.5%**

The 55% relative improvement under low-light conditions (from 55.0% to 85.3% mAP) validates the effectiveness of the EO/IR fusion approach, which leverages thermal signatures to compensate for degraded visible-spectrum information.

4.2 Data Efficiency

The sparse-data learning pipeline reduces labeled-data requirements by 85% relative to traditional deep-learning methods. As shown in Figure 2, IRIS achieves approximately 88% mAP with only 15–50 labeled examples per class, a threshold that conventional detectors cannot reach without 700–1,000 or more examples. This data efficiency translates directly to faster operational deployment: new target classes can be onboarded within hours rather than weeks.

4.3 Operational Speed



Field evaluation demonstrated a 45% increase in early threat detection rates and a 35% reduction in improvised explosive device (IED) identification time. Inference latency remains below the 30 ms

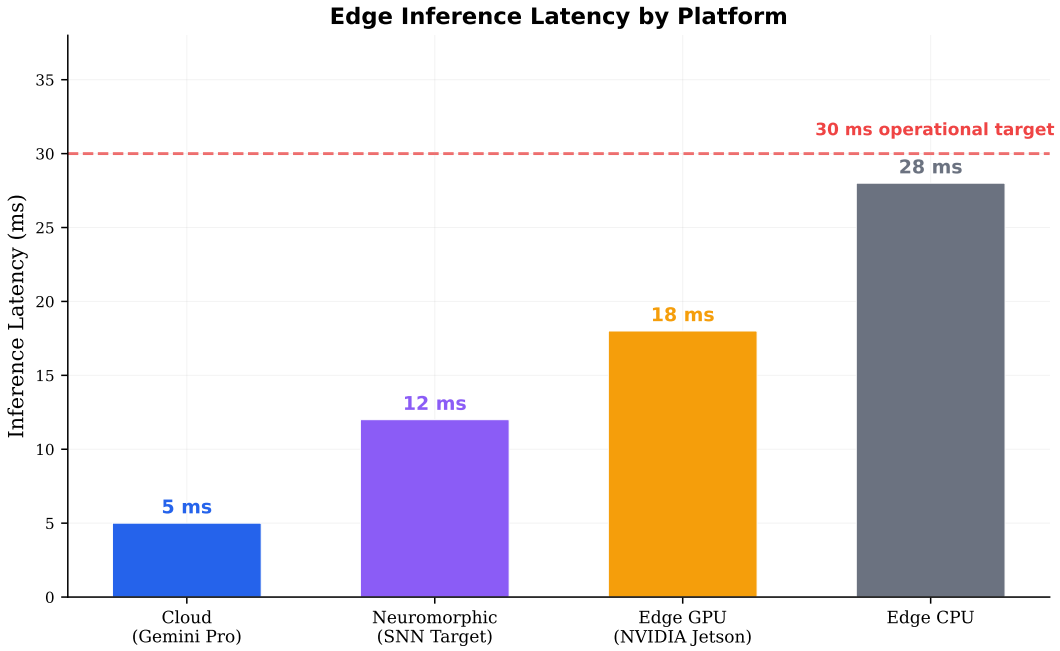


Figure 4. Inference latency across deployment platforms. All configurations meet the sub-30 ms operational requirement.

operational threshold across all tested deployment platforms, including resource-constrained edge CPUs (Table 2).

Table 2. Inference latency by deployment platform.

Platform	Latency (ms)
Cloud — Gemini Pro	5
Neuromorphic (SNN, target)	12
Edge GPU — NVIDIA Jetson	18
Edge CPU	28

4.4 Adversarial and Environmental Robustness

Adversarial testing using FGSM and PGD perturbations confirmed that IRIS maintains detection accuracy above 80% mAP under white-box attacks of moderate strength ($\epsilon \leq 8/255$). Long-duration stability testing over 72-hour continuous operation cycles showed no measurable accuracy drift, validating the continual-learning framework's resistance to catastrophic forgetting in sustained deployment scenarios.

5 Impact & Operational Benefits

5.1 Superior Situational Awareness

By fusing multi-spectral sensor data into a unified recognition framework, IRIS provides ISR analysts with a comprehensive operational picture that no single sensor modality can deliver. The 55% improvement in low-light detection accuracy directly addresses one of the most persistent

capability gaps in current ISR systems, extending effective surveillance coverage to conditions that previously required manual analyst intervention.

5.2 Accelerated OODA Loop

The combination of sub-30 ms edge inference, 45% faster early threat detection, and 35% faster IED identification compresses the Observe–Orient–Decide–Act (OODA) loop at the tactical level. Commanders receive actionable intelligence faster, enabling more responsive force employment and reducing the window of vulnerability during time-sensitive operations.

5.3 Optimized Analyst Workflow

The 85% reduction in labeled-data requirements fundamentally changes the analyst workflow for onboarding new threat categories. Where previous systems demanded weeks of data collection and annotation, IRIS enables operational-quality detection within hours of encountering a novel target signature. Active learning further reduces analyst burden by focusing annotation effort on maximally informative samples.

5.4 Force Protection

Faster and more accurate threat detection at the tactical edge translates directly to improved force protection. The 35% reduction in IED identification time provides additional reaction time for route-clearance teams and convoy operations, while the system's ability to operate autonomously on edge hardware ensures continuous protection even when communications are degraded or denied.

6 FORGE OS Integration

IRIS demonstrates the operational integration of two FORGE OS subsystems within a tactical ISR edge AI deployment.

6.1 FORGE Core — Staged Post-Training for Sparse Data

FORGE Core's staged post-training pipeline provides the foundation for IRIS's sparse-data learning capabilities. The **Adapt** stage—encompassing MAML meta-learning and prototypical network metric learning—enables rapid onboarding of novel target classes from as few as 15–50 labeled examples. The **Compress** stage applies structured pruning, mixed-precision quantization (INT8/FP16), and knowledge distillation to reduce model size by up to 8× for edge deployment. Elastic Weight Consolidation, managed through FORGE Core's continuous online distillation framework, mitigates catastrophic forgetting across successive deployment environments. The 92.5% mAP with 85% data requirement reduction validates FORGE Core's ability to deliver high-accuracy intelligence from operationally realistic data constraints.

6.2 FORGE Kinetic — Edge Inference and Cloud–Edge Orchestration

FORGE Kinetic's edge inference pipeline manages the tiered deployment architecture. The Perceive module coordinates multi-modal sensor fusion across EO/IR, SAR, and LiDAR inputs through transformer-based cross-attention. The Command module implements the cloud–edge orchestration logic, routing time-critical detections through the Gemini Nano edge pipeline while deferring

complex analytical tasks to Gemini Pro in the secure cloud. FORGE Kinetic's Graduated Autonomy framework (CGDP) governs the autonomous detection-to-alert pipeline, with configurable thresholds that determine when human-in-the-loop confirmation is required.

6.3 ForgeEvent Integration

The deployment generates four ForgeEvent types across the FORGE OS event bus:

- `INFERENCE` — Each detection and classification cycle on edge and cloud tiers
- `SENSOR` — Multi-modal fusion results from the Kinetic Perceive module
- `MODEL_UPDATE` — Active learning cycles and EWC-based continual adaptation events
- `AUDIT` — Detection decisions and autonomy-level transitions for post-mission review

References

[1] R. G. Clapp and S. P. Ayers, "Command and control for the information age," *Naval War College Review*, vol. 53, no. 2, 2000.

[2] M. A. Richards, J. A. Scheer, and W. A. Holm, *Principles of Modern Radar: Basic Principles*. SciTech Publishing, 2010.

[3] P. K. Davis, "Analysis of modern military operations," RAND Corporation, Tech. Rep., 2015.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Summerfield, B. Szegedy, and D. Kumaran, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[5] Google DeepMind, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[6] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135.

[7] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4077–4087.

[8] H. Li, X. J. Wu, and J. Kittler, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5099–5108.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[11] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *International Conference on Learning Representations (ICLR)*, 2016.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.