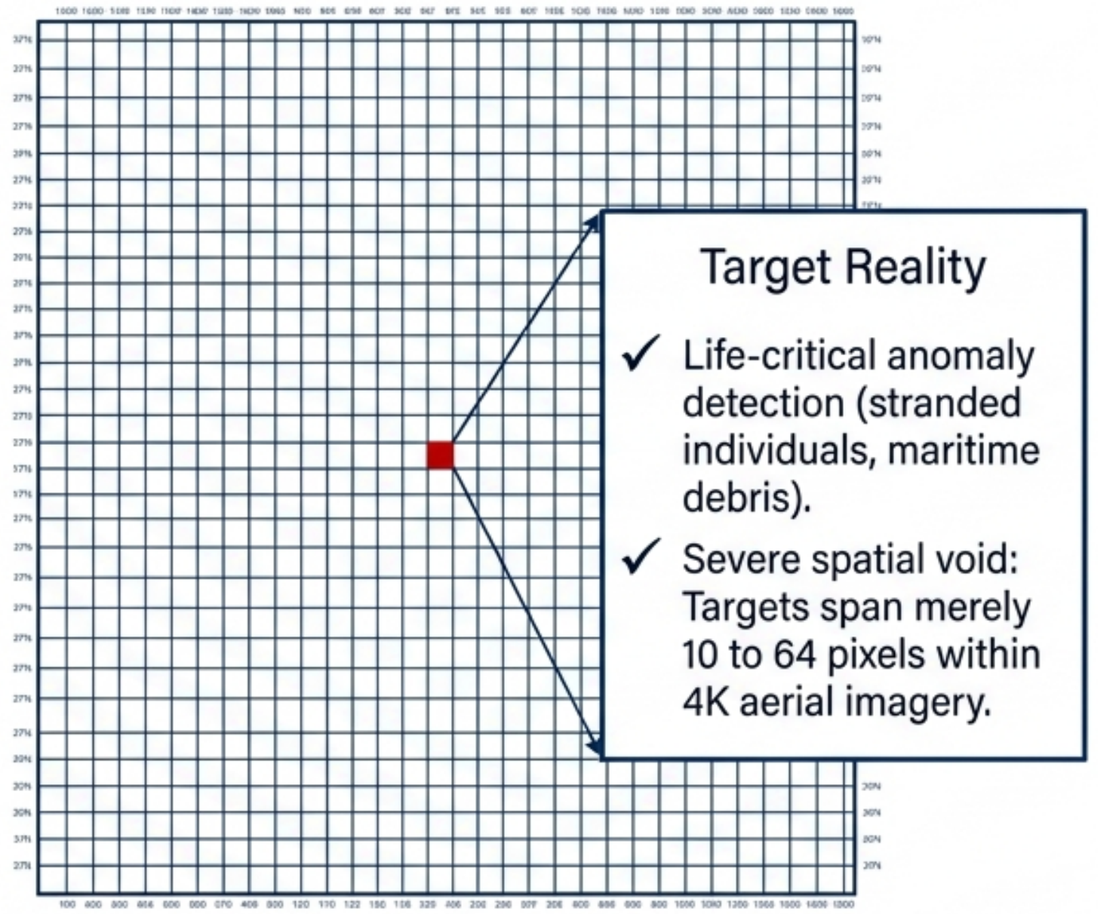
MITE-Net: SWaP-Optimized 4K Video Tiny Target Perception for Embodied Edge SAR

A Comprehensive Framework for Real-Time Onboard UAV Intelligence

The Mission Profile: Embodied Edge SAR

The Mission: High-Altitude Discovery



Target Reality

- ✓ Life-critical anomaly detection (stranded individuals, maritime debris).
- ✓ Severe spatial void: Targets span merely 10 to 64 pixels within 4K aerial imagery.

The Constraint: SWaP Boundaries

Size Weight Power

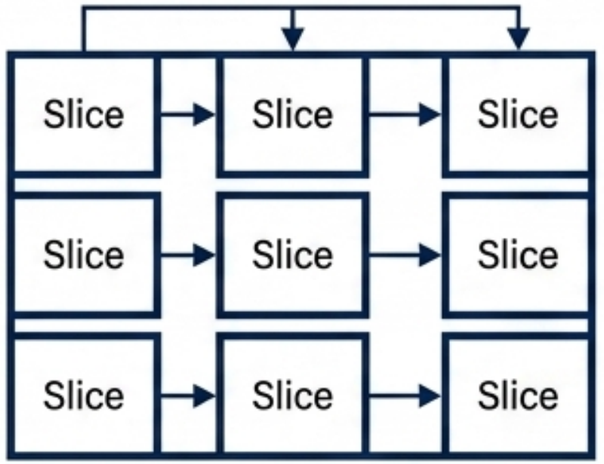
Deployment Platform: NVIDIA Jetson AGX Xavier

Strict Size, Weight, and Power limits govern all embodied edge intelligence. A UAV's operational flight endurance directly correlates with the onboard AI's energy efficiency. High power draw results in **mission failure**.

The Spatial-Computational Dilemma

Method A: The Slicing Approach

Slice-aided processing (e.g., SAHI)



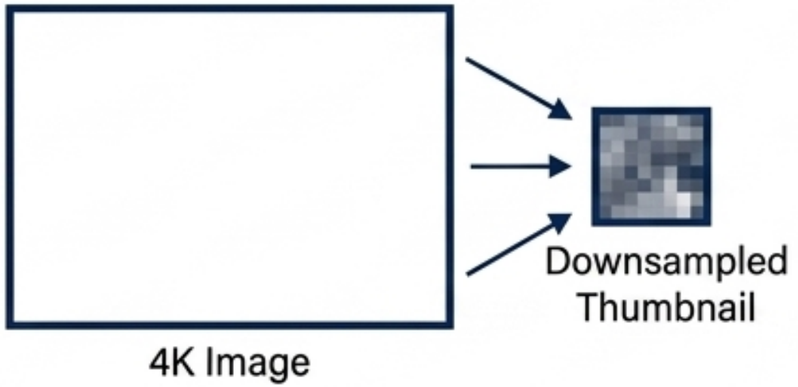
✅ **Result:** Preserves local high-frequency details (High Recall).

FATAL FLAW: Prohibitive computational overhead. Destroys real-time frames-per-second (FPS) and drains battery.

The Bottleneck

Method B: The Downsampling Approach

Direct image downsampling
(e.g., Standard YOLO pipelines)



✅ **Result:** Achieves rapid, real-time throughput (High FPS).

FATAL FLAW: Obliterates microscopic target features. Yields unacceptably low recall rates for tiny targets.

A Comprehensive Solution: The MITE-Net Framework

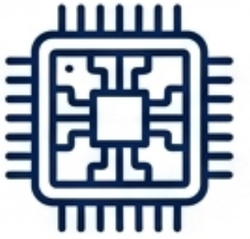
Pillar I: SWaP-Optimized Architecture

MITE-Net (Motion-Informed Tiny-target Edge Network). Decouples spatial and motion perception to break the feature-loss dilemma.

Pillar II: Standardized SAR-Tiny Datasets

Relabeled high-resolution 4K aerial datasets focusing strictly on sub-64 to 256-pixel targets.

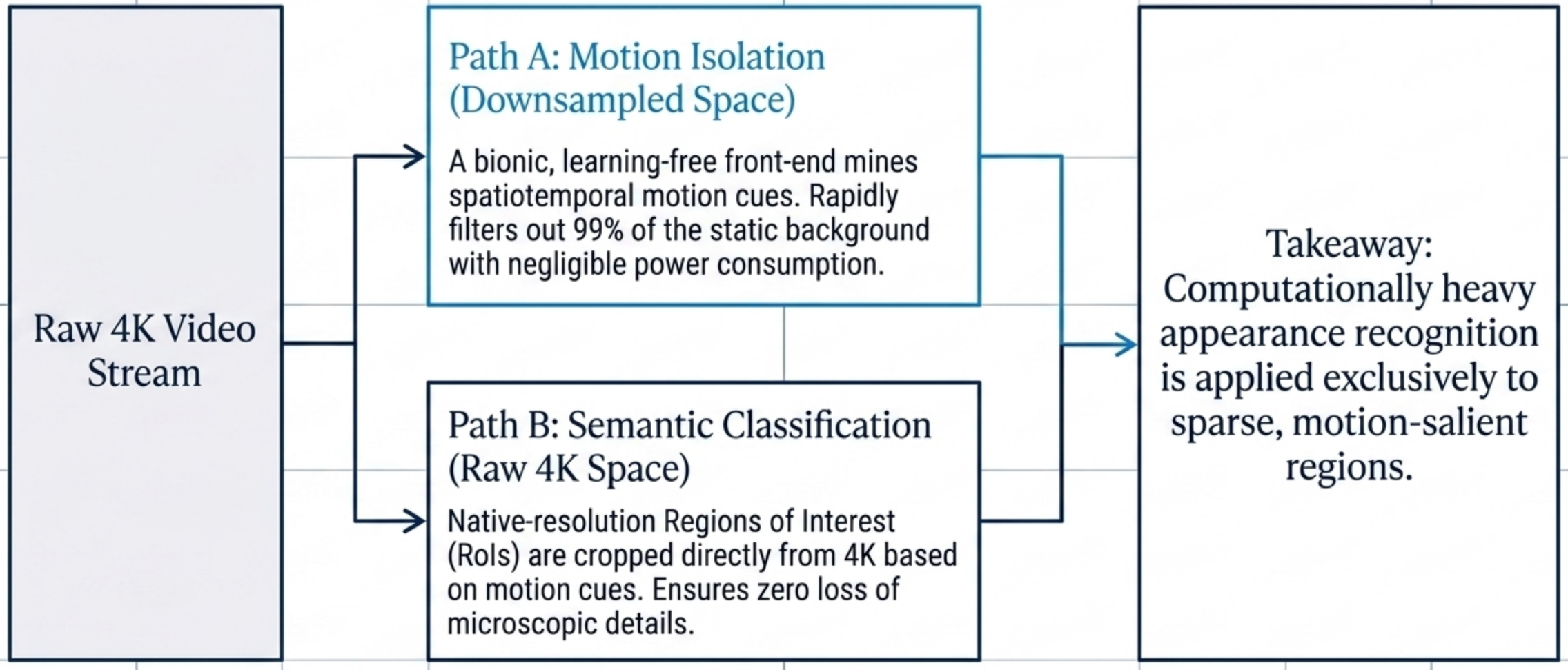
Pillar III: Edge Hardware Benchmarks



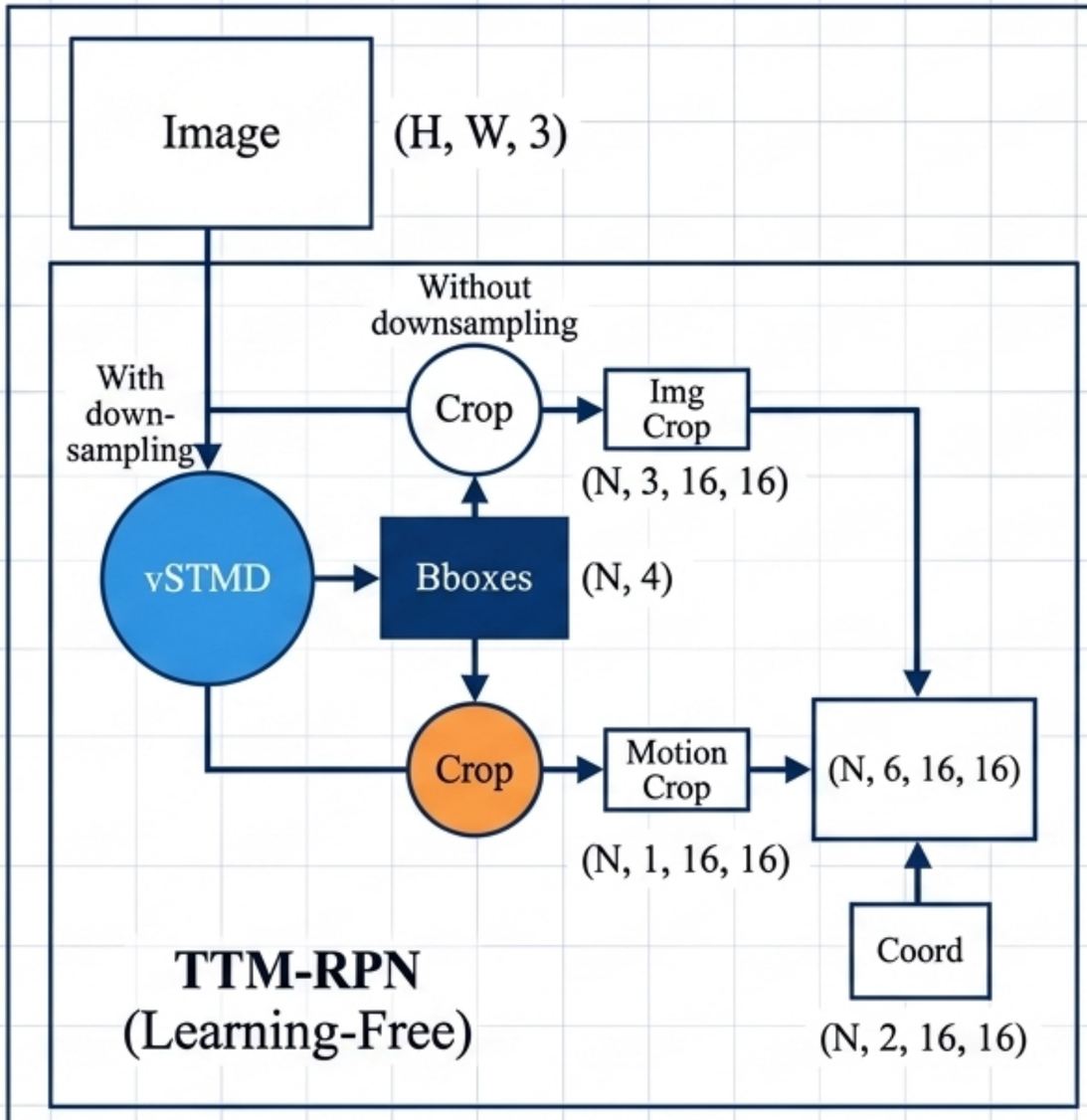
Rigorous hardware profiling on NVIDIA Jetson AGX Xavier to validate true real-time, energy-efficient deployment.

Real-Time Embodied Autonomy

Architectural Paradigm Shift: Decoupling Perception



Stage 1: Tiny Target Motion-Based RPN (TTM-RPN)



Key Bionic Mechanisms

1. Biological Origin:
Built upon the bio-inspired visual systems of dragonflies (vSTMD).

2. Filter Topology:
Employs a learning-free Spatio-Temporal filtering mechanism.

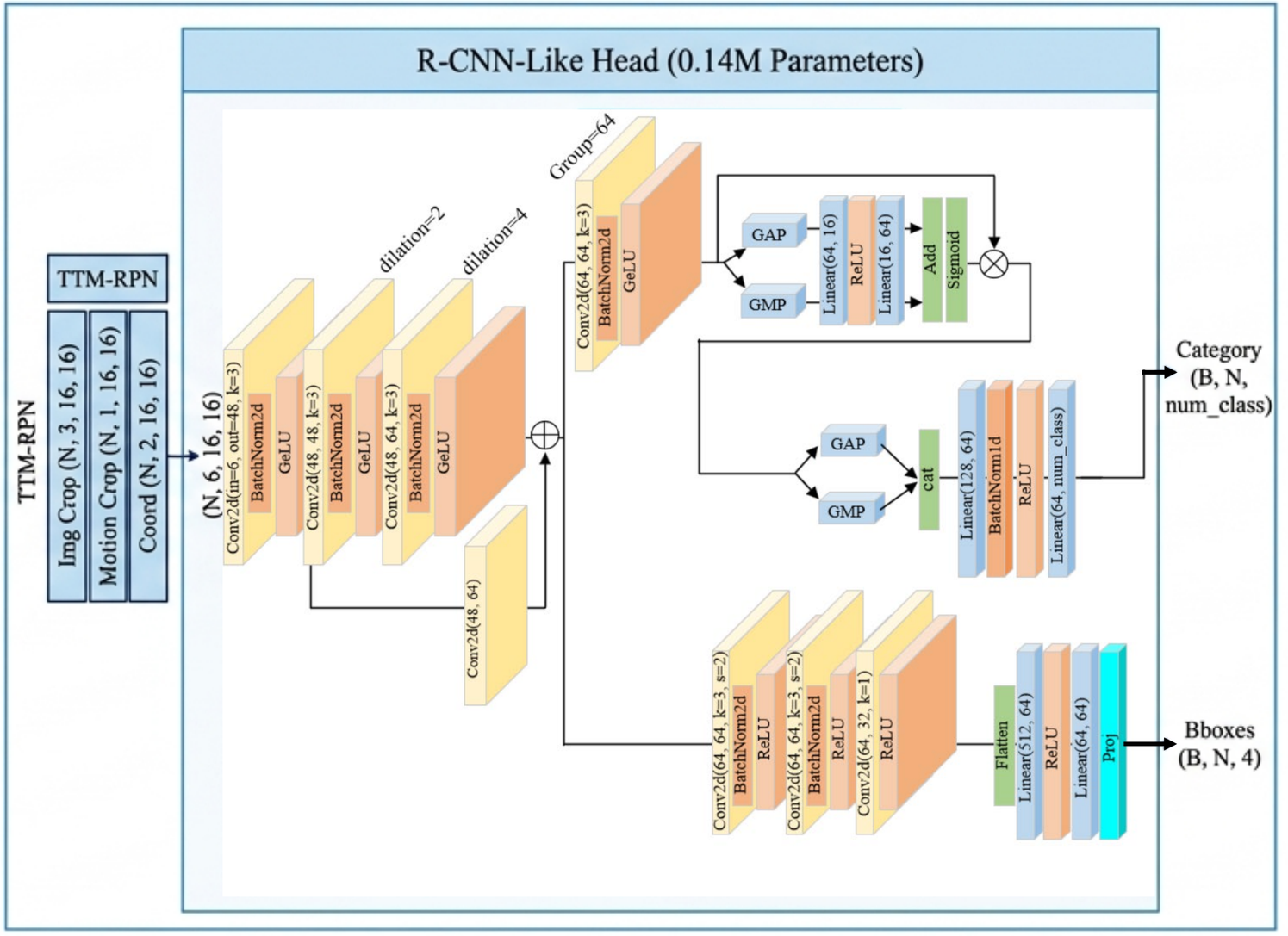
3. Signal Extraction:
Generates score maps and direction maps to extract isolated motion points.

4. Spatial Correction:
Applies directional forward shifts to align bounding box centers accurately with the moving target body.

Stage 2: Ultra-Lightweight Semantic Head

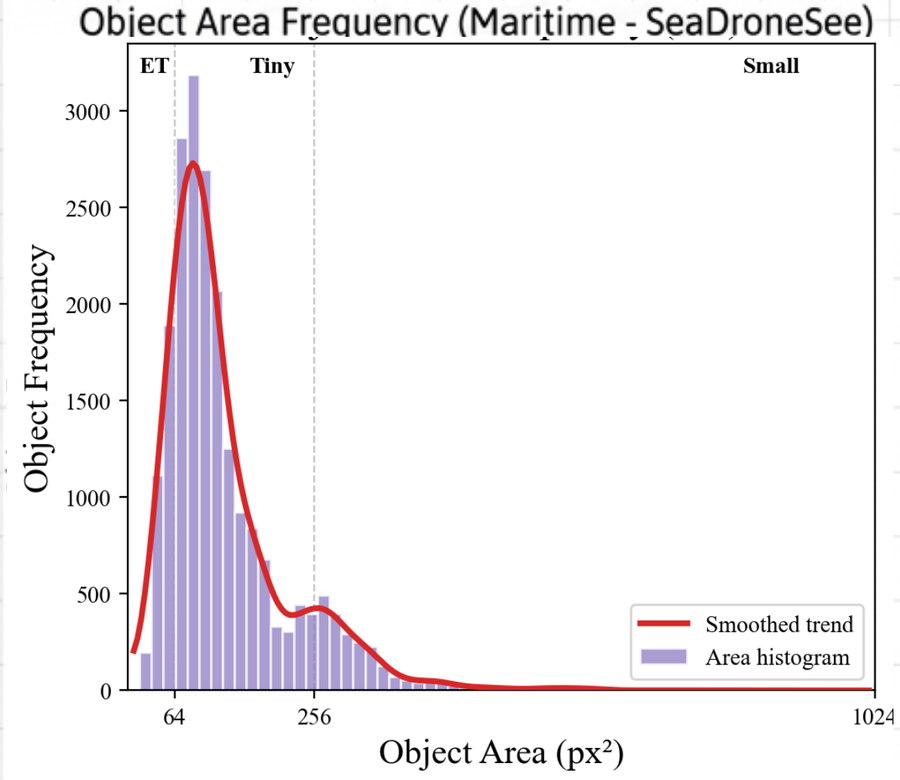
Design Specifications

- **Parameter Scale**
  Sub-0.14M parameters (less than 5% of standard lightweight YOLO models).
- **Multi-Modal Inputs**
  Fuses 16x16 raw image crops with explicitly separated ON/OFF motion polarity maps.
- **Spatial Awareness**
  Uses Coordinate Convolution (CoordConv) to anchor local features.
- **Receptive Expansion**
  Utilizes a Dilated Backbone (rates 2 and 4) to expand context without destructive pooling.
- **Optimization Strategy**
  Trained using Normalized Wasserstein Distance (NWD) explicitly suited for tiny target boundary regression.



Standardizing Evaluation: The SAR-Tiny Datasets

The Data Gap:
General UAV datasets either lack true 4K resolution or completely ignore extremely minute targets. We introduce curated sub-sets of **SeaDroneSee** and **UAVID** categorized **strictly** by microscopic pixel area.

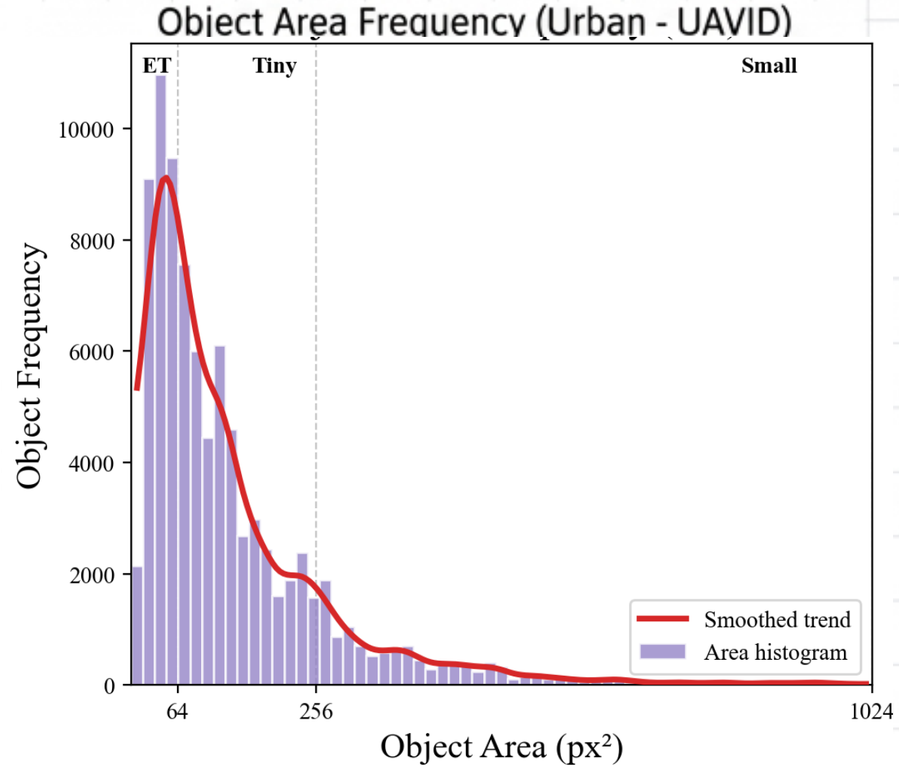


Scale Definitions

– **"Small"**:
256 < Area ≤ 1024 pixels.

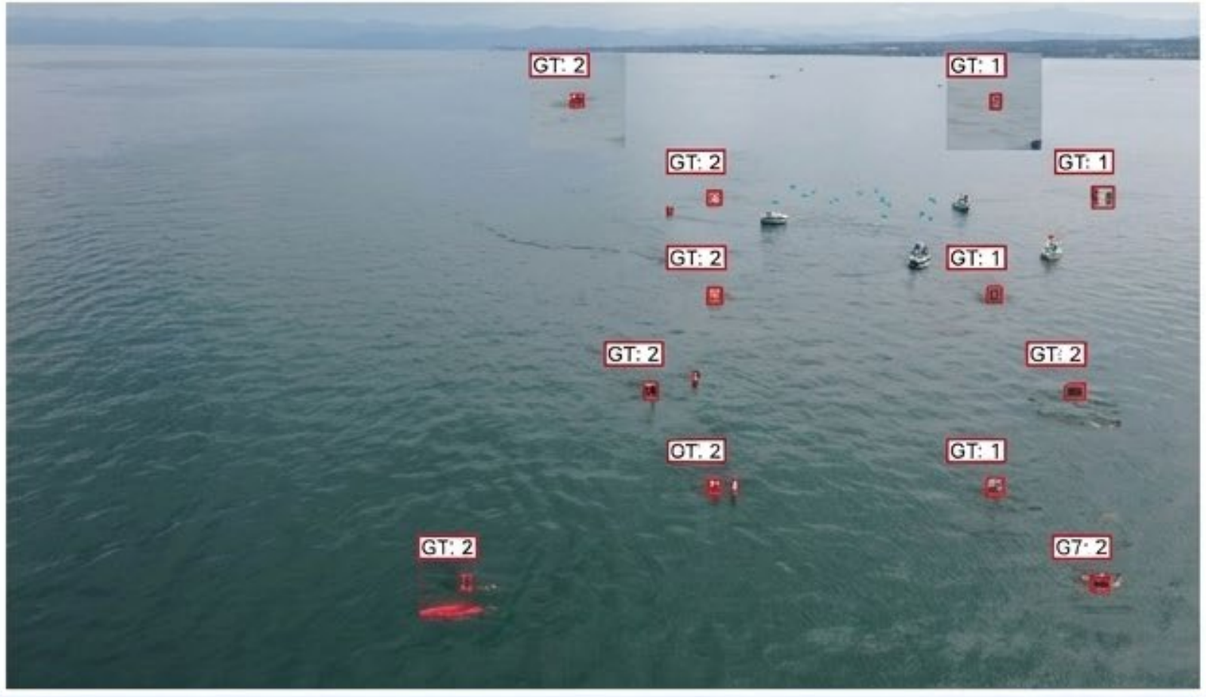
– **"Tiny"**:
64 < Area ≤ 256 pixels.

– **"Extremely Tiny (ET)"**:
Area ≤ 64 pixels.



Escalating Difficulty: Maritime Dynamics vs. Urban Clutter

SeaDroneSee-Tiny



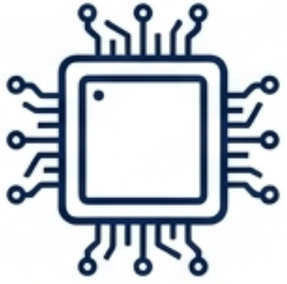
Environment	Dynamic maritime scenes (waves, sun glint).
Target Scale	Predominantly ‘Tiny’ (64-256 px). Swimmers and life jackets.
Stress Test	Sequence 1 mimics real-world crises with up to 12 scattered minuscule targets per frame.

UAVID-Tiny

Environment	Complex, highly cluttered urban landscapes.
Target Scale	Massive concentration strictly in the ‘Extremely Tiny’ domain (≤ 64 px).
Stress Test	Evaluates model capacity for early-stage target discovery against deceptive background noise and structures.

Edge Benchmarking Protocol

Hardware Spec Sheet



Deployment Hardware

Platform	NVIDIA Jetson AGX Xavier
Compilation	All models exported to highly optimized TensorRT engines
Precision	Float16 (FP16) Quantization

Dual-Domain Evaluation Criteria

SAR Mission Criteria

Search Success Rate (SSR)
% of ground truth trajectories locked in a sliding window.

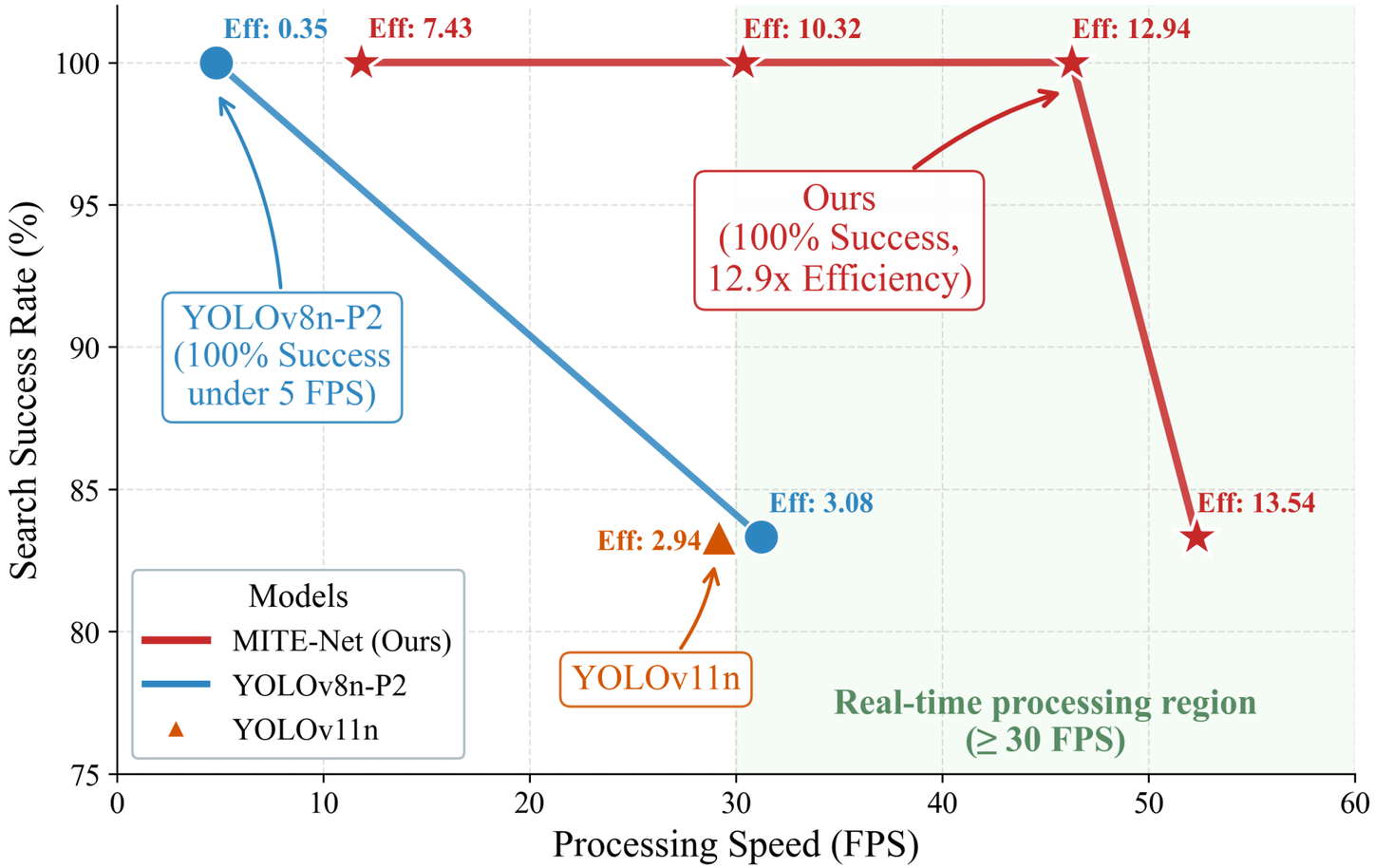
False Alarm Rate (FAR)
Secondary metric evaluating background noise rejection.

SWaP Constraints

Scale	Total parameters, G-FLOPs.
Viability	Inference Latency (ms), Total Power (W), Energy Efficiency (FPS/Watt).

Insight: In life-critical SAR, robust discovery (SSR) strictly supersedes precise bounding box overlap, while SWaP constraints dictate actual flight viability.

The Optimal Trade-off: Processing Speed vs. Success



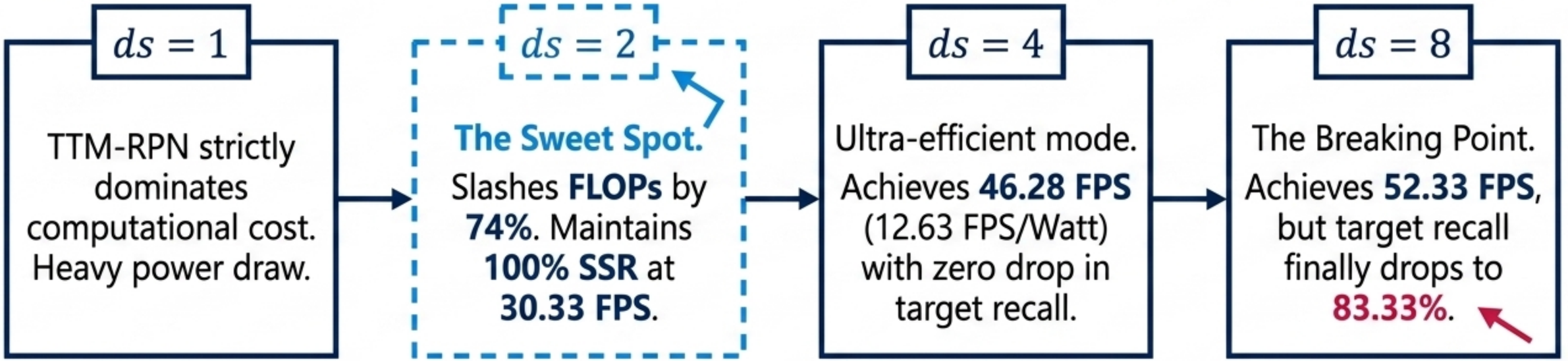
MITE-Net represents the sole architecture capable of occupying the optimal high-speed, high-recall quadrant under hardware constraints.

Comprehensive SWaP Dominance

Metric	YOLOv8n-P2	YOLOv11n	MITE-Net (Ours)
Total Parameters	3.01 M	2.59 M	**0.14 M**
Complexity (G-FLOPs)	41.80	32.85	**0.56**
Inference Power	10.15 W	9.91 W	**3.19 W**
Hardware Efficiency	3.08 FPS/Watt	2.94 FPS/Watt	**9.51 FPS/Watt**

Insight: MITE-Net operates on less than 5% of the parameter budget of the smallest YOLO baselines, while consuming just one-third of the power footprint.

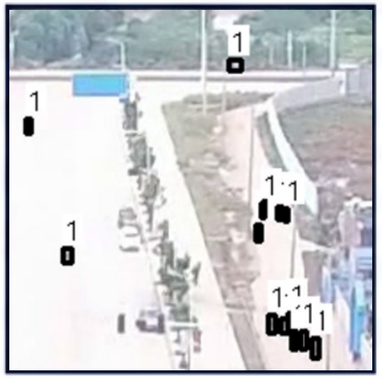
Scalability via Spatial Downsampling (ds)

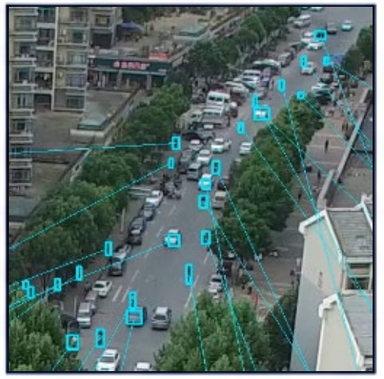


Architectural Lever:
As the downsampling factor (ds) increases, the primary computational burden dynamically shifts from the bionic front-end entirely to the semantic head, acting as a flexible hardware dial for edge deployment.

Structural Boundaries in Urban Hyper-Clutter

The Phenomenon





Performance Collapse: MITE-Net experiences severe degradation on the UAVID-Tiny dataset, with Search Success Rate plummeting from 100% to between 8.33% and 0.00%.

Diagnostic Analysis (Root Causes)

Cause 1: RPN Failure

- The non-learning TTM-RPN struggles to mathematically separate moving tiny targets from dense, static urban structural noise.
- Initial hit rate drops to a mere 0.1047.

Cause 2: Representational Collapse

- The sub-0.14M-parameter semantic head is sufficient for open maritime backgrounds but fundamentally lacks the mathematical capacity to resolve complex urban feature mapping.

Conclusion & Future Trajectory

Core Achievement

- **MITE-Net** successfully bridges the spatial-computational divide.
- Coupling bio-inspired motion extraction with an ultra-lightweight semantic head achieves true real-time, high-recall 4K autonomy.
- Operates under strict 3W edge hardware constraints.

The Trajectory Forward

- **Next Step:** Evolving the fixed-parameter bionic front-end into a fully learnable, end-to-end integrated architecture.
- **Goal:** Scaling spatial filtration to permit a slightly higher-capacity semantic head. This will conquer hyper-cluttered urban environments while strictly preserving the ultra-low SWaP profile.

Establishing a highly practical baseline for the next generation of energy-efficient onboard edge intelligence.