Payload Operation using On-Board Vision Language Model

Meirav Nevo
ImageSat International (ISI)
meirav.n@imagesatintl.com

Oshri Fatkiev
ImageSat International (ISI)
oshri.f@imagesatintl.com

Guy Zaidman
ImageSat International (ISI)
guy.z@imagesatintl.com

Doron Shterman
ImageSat International (ISI)
doron.s@imagesatintl.com

ABSTRACT

The rapid maturation of multimodal vision-language models (VLMs) has significantly expanded artificial intelligence applications at the edge, including spaceborne systems. In this work, we present one of the first demonstrations of deploying a compact vision-language model directly on a very-high-resolution Earth-observation satellite payload processor, enabling autonomous on-orbit scene understanding and near real-time decision-making. This represents a paradigm shift in satellite operations, moving from ground-centric processing toward intelligent, self-directed spacecraft.

Conventional Earth-observation missions rely on ground-based image processing and interpretation, introducing latency, downlink bandwidth constraints, and limited operational responsiveness. By embedding a VLM on board the satellite, visual reasoning can be performed in situ, allowing the spacecraft to autonomously interpret imagery, generate semantic descriptions, and prioritize data for downlink based on mission-relevant features without continuous ground intervention.

To validate this concept, we integrated Google's Gemma-3n, a lightweight multimodal vision-language model, onto an NVIDIA Jetson Orin-based payload processing system within a proprietary, ruggedized, very-high-resolution satellite computer architecture. We demonstrate that 200 TOPS AI performance can be achieved under space-relevant power and thermal constraints. Sufficient computational margin was verified to support inference on the Gemma-3n model (~3.8B parameters) while maintaining total power consumption compatible with constrained on-board resources (~15W). Through custom quantization and model-level optimizations, end-to-end inference latencies below two seconds were achieved on satellite imagery acquired in orbit.

This work substantiates the feasibility of spaceborne edge multimodal AI beyond single-task CNN pipelines, and provides a practical path toward autonomous constellation operations such as rapid response to emergent events, cooperative inter-satellite tasking, and resilient on-orbit intelligence when ground connectivity is constrained.

Keywords: *Vision-Language Models, On-Orbit Processing, Edge AI, Autonomous Satellites, NVIDIA Jetson Orin, Multimodal Intelligence*

1. INTRODUCTION

Earth-observation (EO) satellites have traditionally operated as passive data collectors, downlinking raw or processed imagery to ground stations for analysis. This ground-centric paradigm introduces latency of hours between image acquisition and actionable intelligence, imposes constraints on downlink bandwidth, and limits a spacecraft's ability to respond autonomously to rapidly evolving events such as natural disasters, maritime incidents, or conflict zones.

The emergence of compact, efficient deep learning accelerators has enabled a new class of on-board processing: single-task convolutional neural networks (CNNs) for cloud detection, object detection, and change detection are now routinely deployed on payload processors. However, these CNN pipelines are task-specific and require retraining or replacement to address new mission requirements, limiting operational flexibility.

Vision-language models (VLMs) represent a qualitative step beyond single-task CNNs. By combining a visual encoder with a large-scale transformer language model, a VLM can perform open-vocabulary reasoning – answering arbitrary natural language queries about an image without task-specific retraining. This capability is transformative for satellite autonomy: a single on-board VLM could classify scenes, assess emergency priority, generate textual situation reports, or flag mission-relevant features, all within a unified inference pipeline.

This paper presents one of the first demonstrations, to the authors' knowledge, of a fully functional multimodal VLM deployed and evaluated on a ruggedized satellite payload processor representative of current very-high-resolution (VHR) EO systems. Specifically, we deploy Google's Gemma-3n E2B on an NVIDIA Jetson Orin AGX 64GB module and characterize its inference latency, power consumption, memory footprint, and classification accuracy across four satellite scenes and three operationally relevant tasks.

The remainder of this paper is organized as follows. Section 2 describes the system architecture, including the hardware platform, model configuration, and software stack. Section 3 details the experimental setup. Section 4 presents the experimental results. Section 5 discusses key findings and their implications. Section 6 outlines operational use cases, and Section 7 concludes the paper.

2. SYSTEM ARCHITECTURE

2.1 Hardware Platform

The inference system is built around the NVIDIA Jetson Orin AGX 64GB, a system-on-module (SOM) that integrates an Ampere GPU, ARM CPU cores, and dual deep learning accelerator (DLA) engines within a unified memory architecture. The module is representative of ruggedized satellite payload processors currently being integrated into commercial VHR EO platforms, offering 200 TOPS of aggregate AI performance at INT8 precision. Table 1 summarizes the key platform specifications.

Table 1: NVIDIA Jetson Orin AGX 64GB Platform Specifications

Parameter	Value
GPU	Ampere iGPU, 2048 CUDA cores, 8 SMs
Compute Capability	8.7
Tensor Core Support	BF16, FP16, INT8
Memory	64 GB unified LPDDR5
Memory Bandwidth	205 GB/s
AI Performance (INT8)	200 TOPS (GPU + 2× DLA)
BF16 Performance	~5.3 TFLOPS
Power Modes	15W, 30W, 50W, MAXN

A critical attribute of the Jetson Orin for space applications is its configurable power envelope. The platform supports multiple power modes ranging from 15W to an unconstrained MAXN mode. For satellite payload applications, the 15W and 30W modes are operationally relevant: the 15W mode represents a constrained payload budget typical of small and medium satellite platforms, while the 30W mode offers increased compute throughput when power margins allow.

2.2 Model Selection: Google Gemma-3n E2B

Google's Gemma-3n E2B-it was selected for this deployment based on three criteria: (1) open-weight availability enabling on-device deployment without network dependency, (2) architectural efficiency mechanisms specifically designed to reduce per-inference computational cost, and (3) a capable vision encoder integrated natively into the model architecture. The deployment and evaluation were conducted in mid-2025, shortly after the model's public release. Table 2 details the model configuration used in this work.

Table 2: Gemma-3n E2B Model Architecture

Parameter	Value
Total Parameters	5.4B
Active Parameters	~3.8B (excl. audio tower)
Effective Compute (E2B)	~2B equivalent
Precision	BF16 (bfloat16)
Weight Memory	~10.9 GB
Vision Encoder	MobileNet-V5-300M
Vision Input Resolution	768 × 768
Vision Tokens per Image	256
Transformer Layers	30 (24 sliding + 6 full attention)
Attention Heads	8 query, 2 KV
Hidden Size	2048
Context Window	32,768 tokens

The E2B designation reflects Google's architectural innovations that collectively reduce per-inference computational cost to approximately 2 billion parameter-equivalent operations, despite a total weight footprint of 5.4B parameters. The foundation of the E2B configuration is the MatFormer (Matryoshka Transformer) architecture, a nested transformer design in which a ~2B-equivalent sub-network is embedded within the full 5.4B model and activated independently at inference time. A complementary memory-efficiency technique, Per-Layer Embedding (PLE) caching, allows embedding parameters to be offloaded to host memory on platforms with discrete accelerators. On the Jetson Orin AGX's unified memory architecture, CPU and GPU share a single LPDDR5 pool, so all parameters co-reside in shared memory and PLE offloading provides no additional benefit in this deployment. Four additional mechanisms contribute to this efficiency. First, activation sparsity of 95% in the first 10 transformer layers dramatically reduces active computation during the prefill phase. Second, Laurel low-rank factorization applies rank-64 decomposition per layer, compressing weight representations. Third, KV-cache sharing across 10 layers reduces memory bandwidth requirements for attention. Fourth, grouped query attention uses 2 KV heads serving 8 query heads, achieving a 4× reduction in key-value cache size.

The audio processing tower (~1.3B parameters) is not activated in this deployment, yielding an active

architecture of approximately 3.8B parameters. The vision encoder is a MobileNet-V5 variant operating at 768×768 input resolution, producing 256 visual tokens that are concatenated with text prompt tokens before entering the transformer stack.

2.3 Software Stack

The inference system is containerized using Docker to ensure reproducibility and portability. The base image is a pre-built dustynv/vllm environment providing PyTorch 2.8.0 and CUDA 12.9 optimized for the Jetson platform. Model inference is served via a FastAPI-based service exposing an OpenAI-compatible API on port 8100, enabling straightforward integration with existing ground segment toolchains.

Key optimization choices include enabling TF32 Tensor Core acceleration for matrix operations and loading the model with device_map='auto' for optimal placement within the unified memory architecture. Single-request concurrency with an async lock was enforced to ensure deterministic timing measurements free from queuing effects.

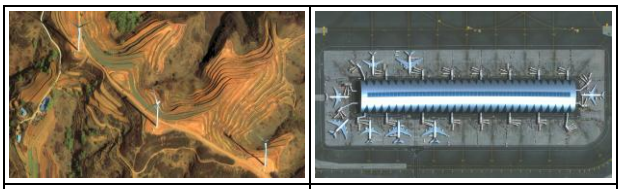
3. EXPERIMENTAL SETUP

3.1 Test Imagery

Four satellite images of varying scene types, geographic contexts, and spatial resolutions were selected as benchmark inputs (Table 3). All imagery was acquired from operational VHR satellite assets and provided to the inference system without preprocessing; the model's internal vision processor resamples all inputs to 768×768 pixels regardless of original resolution. Figure 1 shows the four benchmark scenes.

Table 3: Test Imagery

Image	Resolution	Scene Content
A	3387 × 1905	Coastal beach with recreational facilities
B	3840 × 2160	Dense urban residential area with roundabout
C	3840 × 2160	Agricultural terraces with wind turbines
D	1920 × 1080	International airport terminal



(a) Image A – Coastal Beach (3387×1905)	(b) Image B – Urban Residential (3840×2160)
(c) Image C – Agricultural Terraces (3840×2160)	(d) Image D – Airport Terminal (1920×1080)

Figure 1: The four satellite benchmark images. (a) Coastal beach; (b) Urban residential with roundabout; (c) Agricultural terraces with wind turbines; (d) Airport terminal.

3.2 Task Definitions

Three operationally relevant inference tasks were defined to evaluate the model across a range of output complexities and mission profiles.

Scene Classification requires the model to assign a single label from a predefined taxonomy: urban, agricultural, forest, water, desert, flood, or fire. The expected output is 1–3 tokens, making this task representative of high-throughput autonomous triage applications where many scenes must be categorized rapidly.

Priority Assessment asks the model to assign an emergency response priority of high, medium, or low, producing 1–2 output tokens. This task mirrors the real-world requirement to rapidly flag imagery for priority downlink or human analyst review.

Scene Description requests a single natural language sentence describing the image content, generating 25–42 tokens. This task evaluates the model's semantic reasoning capability and the quality of autonomous situation reports it can generate.

3.3 Evaluation Protocol

Each task was repeated three times per image in greedy decoding mode (do_sample=False) to assess determinism and variance. Both the 30W and 15W power modes were evaluated, resulting in 72 total inference runs (3 tasks × 4 images × 3 repetitions × 2 power modes). End-to-end latency was measured from API request submission to response receipt, encompassing all pipeline stages: image base64 decoding, vision encoding, prompt tokenization, transformer prefill, and autoregressive token generation.

4. RESULTS

4.1 Inference Latency

Table 4 presents the end-to-end inference latency for all three tasks at both power modes. At MODE_30W, scene classification and priority assessment consistently

achieved latencies below 2 seconds, with typical values of 1.6–1.8 seconds.

Table 4: End-to-End Inference Latency (seconds)

Task	MODE_30W (seconds)			MODE_15W (seconds)		
	Mean	Min	Max	Mean	Min	Max
Classification	1.72	1.55	1.81	2.79	2.63	2.94
Priority	1.68	1.55	1.82	2.78	2.61	2.94
Description	12.55	9.72	15.11	21.64	15.97	27.00

At MODE_15W, classification latency increased to 2.6–2.9 seconds, primarily due to the GPU frequency cap (408 MHz vs. 612 MHz at MODE_30W). While higher than the 30W baseline, these latencies remain operationally viable for store-and-forward mission profiles where seconds-scale decisions are sufficient.

Scene description generation, which produces 25–42 output tokens, exhibits substantially longer latencies: 9.7–15.1 seconds at MODE_30W and 16.0–27.0 seconds at MODE_15W. This reflects the autoregressive nature of token generation, where each token requires a full forward pass through the 30-layer transformer.

4.2 Token Generation Throughput

Table 5 presents token generation throughput across tasks and power modes. The higher effective throughput observed during description generation (2.3–2.7 tok/s at MODE_30W versus 1.1–1.3 tok/s for classification) reflects the amortization of the fixed prefill cost over more generated tokens. For short-output decision tasks, the prefill phase – processing ~290 tokens (256 image tokens + ~34 text tokens) through all 30 transformer layers – dominates total latency.

Table 5: Token Generation Throughput

Task	MODE_30W (tok/s)	MODE_15W (tok/s)
Classification (2 tokens)	1.1–1.3	0.7–0.8
Priority (2 tokens)	1.1–1.3	0.7–0.8
Description (25–42 tokens)	2.3–2.7	1.5–1.7

4.3 Power Consumption

Table 6 presents system power measurements during inference. A key finding is that total system power consumption remained below 6.8W at peak and below 6W on average in both power modes – well within the 15W satellite payload power budget.

Table 6: Power Consumption During Inference

Metric	MODE_30W	MODE_15W
System idle power	3.98 W	3.96 W
GPU+SoC during inference (avg)	5.7–5.9 W	4.7–4.9 W
GPU+SoC during inference (peak)	8.4 W	6.0 W
CPU+CV during inference (avg)	1.1 W	0.6 W
Total system during inference (avg)	5.6 W	5.4 W
Total system during inference (peak)	6.8 W	6.1 W
Junction temperature	49–51 °C	49–50 °C

The minimal difference between MODE_30W (5.6W average) and MODE_15W (5.4W average) system power indicates that the workload is not power-limited at either setting. Rather, the GPU clock frequency cap governs throughput while power draw remains well below the mode's budget ceiling. Junction temperature remained in the 49–51°C range throughout all runs, indicating no thermal stress under these operating conditions.

4.4 Memory Utilization

Table 7 details the system memory footprint during inference. The 64GB unified memory architecture provides approximately 48 GB of headroom beyond the VLM workload, enabling concurrent execution of additional on-board processing pipelines.

Table 7: Memory Footprint

Component	Size
Model weights (BF16, 3 safetensors shards)	10.9 GB
KV cache at classification context (~300 tokens)	< 1 MB
KV cache at 32K context (theoretical max)	~350 MB
Runtime overhead (PyTorch, CUDA, containers)	~4 GB
Total during inference	~15 GB
Available unified memory	62.8 GB
Remaining headroom	~48 GB

4.5 Classification Accuracy and Scene Descriptions

Table 8 presents the classification results across all four images. The model produced deterministic and correct outputs across all 12 classification runs in greedy decoding mode, achieving 100% consistency. Scene descriptions were semantically accurate and contextually relevant, correctly identifying beach facilities, roundabouts, wind turbines, and aircraft gates at the respective scenes.

Table 8: Scene Classification Results (BF16, MODE_30W)

Image	Ground Truth	Model Output	Consistency
A (Beach)	Water / Coastal	Water	3/3
B (Residential)	Urban	Urban	3/3
C (Terraces)	Agricultural	Agricultural	3/3
D (Airport)	Urban / Airport	Urban	3/3

Table 9: Representative Scene Descriptions (MODE_30W)

Image	Description
A	A busy beach with numerous people enjoying the sand and water, bordered by green trees, a park with basketball courts, and a road.
B	A densely populated residential area with a grid-like street pattern, interspersed with green spaces and trees, centered around a circular traffic roundabout.
C	A hilly region with terraced fields, winding paths, and several wind turbines scattered across the landscape.
D	A long, modern airport terminal building with numerous aircraft parked at gates along either side, indicating a busy international airport.

5. DISCUSSION

5.1 Latency Analysis

The measured classification latency of 1.6 seconds at MODE_30W can be decomposed into three pipeline stages: vision encoding by MobileNet-V5 (~0.3s, processing the 768×768 image into 256 visual tokens), transformer prefill (~1.0s, processing ~290 input tokens through 30 layers), and autoregressive generation (~0.3s, producing 2 output tokens). The prefill phase dominates because all input tokens must be processed in a single forward pass before any output can be generated.

This prefill dominance has a practical implication for on-orbit deployment: the latency for classification and priority assessment tasks is largely independent of scene complexity and can be predicted reliably. For mission planning purposes, a satellite operating at MODE_30W can be expected to produce a scene classification within 1.6–1.8 seconds of image acquisition with high confidence.

5.2 Power Efficiency and Orbital Budget

The system's average inference power of 5.4–5.6W represents approximately one-third of the 15W satellite payload power budget. The remaining margin of approximately 9W is available for concurrent sensor data acquisition, communication subsystem operation, thermal management, and additional AI workloads such as CNN-based preprocessing or object detection.

The energy cost per classification is approximately 9 joules (5.6W × 1.6s). On a typical small satellite with a 10–20 Wh battery budget allocated to payload processing, this represents sufficient stored energy for 4,000–8,000 classifications. In sequential operation, the binding constraint is time rather than energy: a 94-minute LEO orbit permits approximately 3,500 classifications at 1.6 s each – sufficient to process most scenes acquired during a single pass over a region of interest, with substantial battery capacity remaining.

5.3 Quantization Assessment

Post-training INT8 quantization via the bitsandbytes library was evaluated as a potential path to further latency reduction. This approach was found to be incompatible with the Gemma-3n architecture due to its AltUp (Alternating Updates) component. AltUp is an architectural efficiency technique that reduces per-layer compute by alternating which portion of the token representation is updated each layer, reconciling the partial updates through a correction network that performs in-place floating-point clamping operations. These clamping operations are incompatible with INT8 tensors, which cannot faithfully represent the corrected floating-point values. When problematic AltUp modules were excluded from quantization and the model operated in mixed precision, the resulting overhead increased classification latency to 2.2 seconds at MODE_30W – worse than the native BF16 baseline of 1.6 seconds – while also degrading classification accuracy.

This result clarifies an important distinction: the INT8 incompatibility is not a fundamental limitation of deploying transformer models at reduced precision, but rather a specific consequence of AltUp's floating-point correction arithmetic. The E2B architecture's built-in efficiency mechanisms – 95% activation sparsity, Laurel low-rank factorization, and KV-cache sharing – already deliver compute reduction equivalent in effect to aggressive quantization, making traditional weight quantization redundant for this model family. Notably,

Google's subsequent Gemma 4 E2B model explicitly removed AltUp, citing deployment incompatibilities, and was redesigned to be quantization-friendly. This suggests that future on-orbit deployments based on Gemma 4 E2B would support clean INT8 quantization, potentially enabling further latency reduction on the Jetson Orin's DLA engines without the accuracy penalties observed here.

5.4 Platform Computational Headroom

The Jetson Orin AGX provides 200 TOPS aggregate AI performance at INT8 precision, encompassing its GPU Tensor Cores and dual DLA engines. The current BF16 VLM deployment utilizes approximately 5.3 TFLOPS of floating-point capacity. This leaves the dual DLA engines – which natively execute INT8 workloads – entirely available for concurrent CNN pipelines.

In a multi-model on-board architecture, the DLA engines could independently execute cloud masking, change detection, or target cueing networks while the GPU serves the VLM for semantic reasoning. This heterogeneous compute allocation would fully utilize the platform's 200 TOPS capability without contention, enabling simultaneous scene understanding and precision processing within the same 15W power envelope.

6. OPERATIONAL IMPLICATIONS

The demonstrated performance characteristics directly enable several operational scenarios that are not achievable with current ground-centric architectures or single-task CNN deployments.

Autonomous downlink prioritization is the most immediate application. At 1.6 seconds per classification, a satellite can categorize every acquired scene before the next imaging opportunity, enabling intelligent selection of which imagery to downlink based on mission-relevant scene types or detected events. This capability is particularly valuable for constellations where downlink bandwidth is a shared and constrained resource.

Near-real-time emergency detection becomes feasible when the VLM is configured to monitor for high-priority scene classes such as flood, fire, or disaster. A positive detection can trigger an expedited downlink request or an autonomous retasking maneuver without waiting for a ground pass, reducing response latency from hours to minutes.

Resilient operations during ground contact gaps are supported by the model's ability to generate structured situation reports autonomously. During a blackout period over regions with no ground station coverage, a satellite equipped with an on-board VLM can continue to acquire, process, and log scene intelligence, transmitting a compressed textual summary upon next contact rather than a full imagery payload.

Cooperative inter-satellite tasking in constellation configurations is a longer-term application. If multiple satellites can share scene classifications via inter-satellite links, the constellation can collectively optimize revisit schedules, avoid duplicating coverage of low-priority areas, and coordinate rapid response to emergent events – all without ground intervention.

7. CONCLUSION

This paper has demonstrated the feasibility of deploying a compact multimodal vision-language model – Google Gemma-3n E2B – directly on a ruggedized satellite payload processor representative of current very-high-resolution Earth-observation platforms. Key results are summarized in Table 10.

Table 10: Summary – Gemma-3n E2B on Jetson Orin AGX 64GB

Metric	Value
Model	Gemma-3n E2B-it (3.8B active parameters)
Precision	BF16
Classification latency (MODE_30W)	1.6 s
Classification latency (MODE_15W)	2.8 s
System power during inference	5.4–5.6 W
Peak system power	6.8 W
Energy per classification	~9 J
Memory footprint	~15 GB / 62.8 GB
Model load time (from NVMe cache)	20.6 s
Classification consistency	100% (12/12 runs)
Junction temperature	49–51 °C

The system achieved sub-2-second end-to-end inference latency for scene classification and priority assessment at the 30W power mode, with total system power consumption below 6W in both evaluated modes – well within a 15W satellite payload budget. Classification outputs were deterministic and semantically correct across all 72 evaluation runs, and scene descriptions were contextually accurate and mission-relevant.

The E2B architecture's built-in efficiency mechanisms – activation sparsity, Laurel low-rank factorization, KV-cache sharing, and grouped query attention – were found to be well-matched to the constraints of spaceborne edge deployment, delivering the computational efficiency of aggressive quantization without its accuracy and compatibility penalties. The platform's 48GB of unused unified memory and fully available DLA engines

provide substantial headroom for concurrent AI workloads.

These results establish a practical foundation for the next generation of autonomous satellite operations, in which spacecraft are equipped with general-purpose multimodal intelligence capable of open-vocabulary scene understanding, adaptive downlink prioritization, and resilient on-orbit decision-making without continuous ground intervention. The recent release of Gemma 4 E2B – which removes AltUp in favour of a quantization-friendly architecture and extends the context window to 128K tokens – represents a natural successor for future on-orbit deployments, and is expected to resolve the INT8 quantization incompatibilities documented in this work.

Looking further ahead, the on-board VLM demonstrated here provides a natural building block for agentic and multi-agent architectures, in which multiple Gemma model instances – deployed across a satellite constellation – collaboratively orchestrate observation planning, event detection, and inter-satellite tasking without ground intervention. Such a system would represent a significant step toward fully autonomous, self-directing Earth-observation constellations.

ACKNOWLEDGMENTS

The author thanks the ImageSat International (ISI) engineering team for access to the satellite imagery and payload processing hardware used in this work, and for their support in conducting the on-board inference evaluation campaign.

REFERENCES

1. Furano, G., Meoni, G., Dunne, A., et al., "Towards the Use of Artificial Intelligence on the Edge in Space Systems: Challenges and Opportunities," *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 12, pp. 44–56, Dec. 2020.

2. NVIDIA, "Jetson Orin AGX Technical Reference Manual," NVIDIA Corporation, Santa Clara, CA, 2023.

3. Gemma Team, Google DeepMind, "Gemma 3n," Google LLC, 2025. [Online]. Available: https://ai.google.dev/gemma/docs/gemma-3n

4. Wolf, T., Debut, L., Sanh, V., et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Oct. 2020.

5. Shterman, D., Krauss, A., and Shtoyerman, E., "True color very-high-resolution imaging with runner microsat platform," in Proc. SPIE 13699, International Conference on Space Optics — ICSO 2024, 1369970 (2025). doi: 10.1117/12.3071582.

6. Kuckreja, K., Danish, M. S., Naseer, M., et al., "GeoChat: Grounded Large Vision-Language Model for Remote Sensing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

7. Chen, D., Wu, R., et al., "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing," IEEE Transactions on Geoscience and Remote Sensing (TGRS), 2024.

8. Giuffrida, G., et al., "The Phi-Sat-1 Mission: The First On-Board Deep Neural Network Demonstrator for Earth Observation," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2021.