Satellite Detection and Characterization Using Alternate Transformer Architectures for Spectro-Spatial RF Signal Analysis

Sidney BESNARD
Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, France
Safran Data Systems, Colombelles, France

Jalal FADILI
Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, France

Frederic JURIE
Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, France

Steredenn DAUMONT
Safran Data Systems, Colombelles, France

***Abstract*— With the continuous increase in satellite launches and the growing complexity of RF signals, traditional methods struggle to fully exploit the rich data derived from passive frequency scanning. In this paper, we propose an innovative approach based on Transformer architectures to detect and characterize satellites. Our method leverages the intrinsic capabilities of Transformers to model both local and global dependencies in high-dimensional spectro-spatial images. By alternating Transformer layers along the frequency and spatial axes, our model extracts robust, invariant representations even in highly complex signal environments. The design incorporates attention mechanisms that facilitate the simultaneous separation and reconstruction of spectral and spatial footprints, proving especially effective in scenarios with densely overlapping signals and interference. Experimental results on both synthetic and real-world datasets demonstrate that our approach significantly enhances detection and characterization performance compared to traditional architectures, paving the way for more precise and resilient space situational awareness systems.**

***Index Terms*—Transformers, Attention Mechanism, Satellite Detection, RF Signal Characterization, Spectro-Spatial Analysis, Deep Learning, Space Situational Awareness (SSA)**

Manuscript received XXXXX 00, 0000; revised XXXXX 00, 0000; accepted XXXXX 00, 0000.

Sidney BESNARD is with Safran Data Systems, Colombelles, France (e-mail: sidney.besnard@safrangroup.com).

Jalal FADILI is with Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, France (e-mail: jalal.fadili@ensicaen.fr).

Frederic JURIE is with Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, France (e-mail: frederic.jurie@unicaen.fr).

Steredenn DAUMONT is with Safran Data Systems, Colombelles, France (e-mail: steredenn.daumont@safrangroup.com).

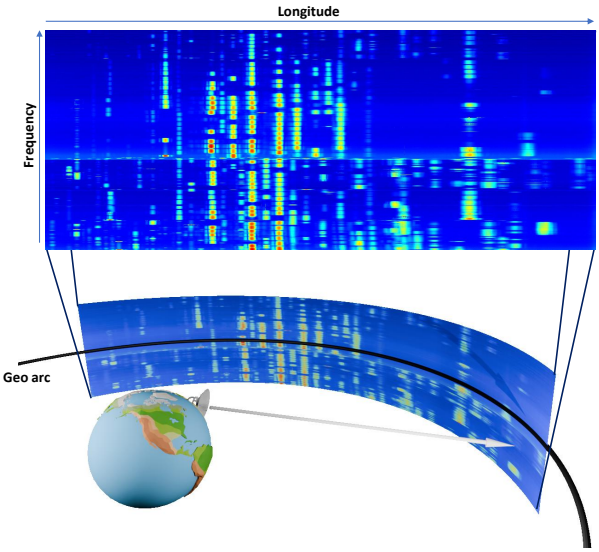

Fig. 1: Illustration of the Ku-band acquisition system, using an antenna located in France, with an OPS step of 0.2 degrees and covering a 140-degree range in longitude.

I. Introduction

Space Situational Awareness (SSA) [1] is crucial to ensure the uninterrupted functioning of vital space-based services, providing accurate information about the space environment. SSA encompasses space weather monitoring, near-Earth object observation, and space surveillance and tracking [2], [3], [4]. A network of diverse sensors (optical, RF, radar, etc.) is utilized to track satellites and estimate their activities and orbits, supporting collision avoidance, fragmentation analysis, reentry analysis, and recovery from service disruptions. However, the sheer volume of data and the complexity of the signals present significant obstacles to traditional manual processing. This paper explores these challenges, specifically the overwhelming data volume and intricate signal interpretation, and proposes a machine learning-based solution to achieve a deep, physics-consistent representation of the data to extract high-level features like satellite detection and orbit parameter estimation.

This paper addresses Space Situational Awareness (SSA) challenges by focusing on passive Radio Frequency (RF) solutions, specifically for tracking active objects. Passive RF offers significant advantages, enabling continuous tracking of active satellites regardless of weather, daylight conditions, cross-tagging complexities, or orbital limitations. The effectiveness of this approach has been demonstrated across various orbits, from Low Earth Orbit (LEO) to cis-lunar space [5].

Building upon these benefits, we have developed a novel passive RF system named ”Watchtower” [6] which is composed of a network of antennas distributed globally around the earth. Designed for the detection of any satellite telemetry and payload signals, thereby enabling

the determination of satellite positions. The "Watchtower" system employs RF sensors (receiver-only) to continuously monitor radio frequency activity from objects in orbit. Each sensor scans the orbits of interest, recording the entire signal spectrum to generate a Power Spectral Density (PSD) at each orbital position, which we refer to as the Object Phase Shift (OPS), as illustrated in Figure 2.

For this study, we focus on geostationary orbits, where the Object Phase Shift (OPS) holds significant physical meaning, directly corresponding to Earth's longitude. We then concatenate the Power Spectral Densities (PSDs) to generate a longitude frequency map. The "Watchtower" system employs small antennas with wide main lobes, allowing for rapid scanning of specific orbits without loss of signal or object exclusion, as illustrated in Figure 1. This approach enables the detection of satellite signals close to the monitored orbit. However, a wide main lobe prevents discrimination of closely spaced objects.

Although characterizing and detecting a satellite is straightforward when it is spatially distant from others, the task becomes complex when satellites are in proximity, causing signal mixtures. When satellite signals spatially and spectrally overlap, the challenge can be viewed as a blind source separation problem.

In addition, various artifacts and perturbations can significantly degrade data quality during acquisition, including:

- Terrestrial interference, such as GSM telecommunications. For example, the C band ($3.4 - 4.2$ GHz) is contaminated with 5G signal [7].
- Pointing error: An uncalibrated antenna (in terms of leveling or bracket adjustment) will cause distortion and highly affect the signal-to-noise ratio (SNR) of the received signal. (In this article, we assume that the antennas are well-calibrated).
- Unfocused source: Similarly to the previous point, an unfocused source will cause distortion and asymmetries in the radiation pattern of the antenna. (In this article, we assume that the source is well-focused).
- Noise: Multiple noise sources are present in the RF chain and disturb the signal. However, it highly depends on the devices that are used and how acquisitions are made. In our case, we assume that the noise floor is non-constant along the frequency and is empirically estimated.
- Antenna position: These parameters cause two effects: spatial distortion along the edge of the acquisition and the area of reception not necessarily aligned with the satellite's transponders [8], [9].

In this study, we explore the benefits of applying machine learning techniques to capture both spatial and spectral information from geostationary satellites within the framework of our "Watchtower" system. The integration of machine learning into Space Situational Awareness (SSA), including satellite tracking applications, has recently garnered considerable attention, with numerous studies proposing machine learning-based solutions in this domain [10], [11], [12], [13]. Many current systems now utilize these ML-powered techniques in scenarios closely related to raw data and sensor inputs [14], [15].

In this case, data acquisition has a strong physical basis, intrinsically linked to orbital dynamics and telecommunication protocols. This allows us to approximate a physical model, transforming data extraction into a non-linear inverse problem. Framing the problem as an inverse problem opens many research avenues, enabling the development of neural networks informed by physical principles, thus improving the robustness and resilience of the system.

With a robust physical model representing our raw acquisitions, we can use this information across various stages of our training framework. First, this model enables the creation of simulated acquisitions, offering fully labeled data, a significant advantage over real data, which are difficult or impossible to label due to limited information on satellite RF protocols. Secondly, through simulated data generation, a neural network can be trained on infrequent events [16], [17], [18], [19], [20], addressing data scarcity challenges. This lack of data is exacerbated by rare events such as RPO (Rendezvous and Proximity Operations) [21], satellite posting, and other unusual scenarios. Having this comprehensive dataset enables us to train a neural network to detect and characterize satellites present in received signals.

II. Related works

1. *Machine learning for SSA and RF applications*

In recent years, the integration of machine learning (ML) techniques in Space/Spectrum Situational Awareness (SSA) Radio Frequency (RF) applications has gained significant traction [22], [23], [24], [10], [11], [12], [13], [25]. ML approaches, particularly deep learning models, have demonstrated their efficacy in handling the complex and high-dimensional data characteristic of SSA and RF environments [22]. For example, neural networks (CNN and MLP) have been employed to identify and classify signals in spectrum sensing, achieving remarkable accuracy in cluttered RF spectra and waterfall [22], [26], [27], [28], [29], [30]. More modern neural network architecture has been used like in [31], [32], [33] which uses YOLO. In addition, reinforcement learning methods have been utilized to detect jamming [34], improving the robustness of the system. However, these methods focus exclusively on detection problems within relatively narrow frequency bands and small temporal scales, which poses a significant limitation in our case. This is because our goal is to share information across much broader frequency ranges and distant spatial scales.

Positioning Against Detection-Only Pipelines. Much of the existing RF-ML literature frames wideband spectrum sensing as an *object detection* task on time-frequency images (e.g., using YOLO-style detectors on waterfall plots). Although effective in localizing emissions, these methods typically output bounding boxes and class labels.

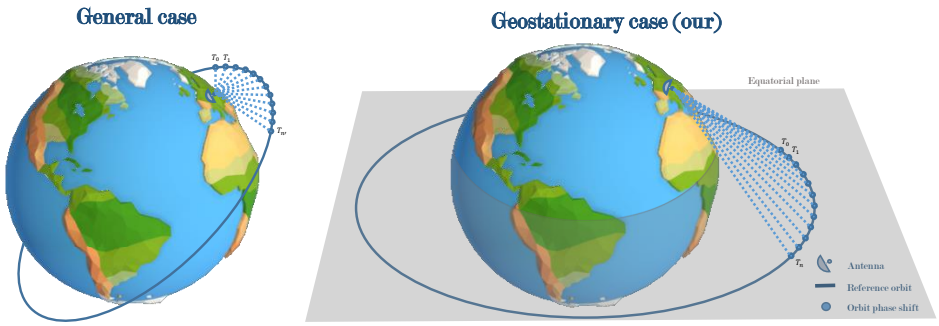

Fig. 2: Illustration of the system in the general case and in the geostationary case.

In contrast, our work addresses the more complex task of joint *detection and characterization*. Our objective is to produce distinct spectral and spatial (longitude) footprints for each satellite and perform blind source separation even under conditions of significant spectral and longitudinal overlap. This involves disentangling and reconstructing per-satellite spectro-spatial signatures from noisy and interference-laden data.

Other efforts focus on domain adaptation techniques to bridge the gap between simulated and real-world RF data, as real-world labeled datasets are often scarce and difficult to annotate. Transfer learning has shown promise in this context, enabling models pre-trained on synthetic datasets [25] to perform effectively in operational environments as in [29], [30] which proposes a generated dataset and achieves good generalization on real data performances. These advances underscore the critical role of ML in addressing the challenges of SSA and RF applications.

2. *Neural networks for large RF data*

Large-scale RF data presents unique challenges, including the need for efficient processing, storage, and analysis. Traditional signal processing techniques, while foundational, struggle to scale with the exponential growth of RF data. To address this, recent research has turned to machine learning (ML) models.

Many neural network architectures aggregate information locally, as numerous problems in the literature are spatially structured and linked. For example, convolutional neural networks (CNNs) are not ideal for very large RF datasets (in 1D temporal or 2D spectral forms), because these require global context to capture essential structure.

Broadly, four families of solutions have emerged. First, [35] enlarge the convolutional kernel to expand the receptive field; however, computational complexity grows quadratically with kernel size and is often controlled via sparsity. Moreover, for large images this approach remains problematic because the convolution window is still smaller than the full image. This limitation naturally motivates the second family of approaches: deeper CNNs that progressively reduce spatial resolution to enlarge context (e.g., ConvNeXt [36]). Yet they inherently struggle to model the long-range dependencies often required in RF analysis [37]; and as the input size grows, achieving sufficient context demands ever-deeper models, imposing significant compute and memory costs.

Third, attention-based models [38], [39] capture global relationships but are computationally expensive. For example, standard ViT approaches compute self-attention over all image tokens, causing compute and memory to explode for high resolutions due to the quadratic time and space complexity of the attention mechanism. To mitigate this, *factorized/axial attention* decomposes the 2D attention into two 1D attentions along the orthogonal axes, reducing complexity while preserving exact global interactions along each axis [40], [41], [42], thus lowering computation and memory costs. Other approaches, such as hierarchical/windowed ViTs (e.g., Swin [43]), scale to high resolutions via shifted local windows; however, cross-window context remains limited unless complemented by window shifting across stages, token merging, or explicit cross-window interaction modules.

Fourth, state-space models (SSMs) [44] and attention-free backbones like Vision Mamba [45] provide efficient long-range modeling in sequence-like data and are promising for RF signals that exhibit long temporal/spectral dependencies.

Finally, hybrid models combining CNNs with attention mechanisms have been proposed like in [46], [47], [48] to balance the strengths of both local aggregation and global contextualization. These approaches represent a promising direction, but still require further refinement to address scalability and efficiency challenges effectively.

Relation to Our Alternating-Transformer. Our architecture is a domain-specific implementation of axial attention, tailored for RF spectro-spatial grids as it was first proposed in [40], [41], [42]. We alternate axis-wise attention along the *frequency* and *longitude* dimensions, operating on tokens generated by a CNN backbone designed to preserve small carriers. This factorized approach reduces the computational complexity from $O((t_f t_l)^2)$ for standard 2D attention to $O(t_l t_f^2 + t_f t_l^2)$. Crucially, unlike windowed architectures like Swin Transformers or CNNs,

our method maintains exact global information exchange along each axis. This design aligns with the physical separability of the data and enhances the model's ability to associate spectrally distant carriers that originate from the same satellite.

3. Summary of Contributions

Our work distinguishes itself from prior art in the following ways:

1) *Physics-Aligned Separability:* We employ alternating axis-wise attention along the frequency and longitude axes. This architecture mirrors the physical structure of the acquisitions and preserves the exact global context along both dimensions.
2) *Structured Outputs for Characterization:* Instead of producing simple bounding boxes, our model predicts and reconstructs complete spectral and spatial footprints for each satellite, using a permutation-invariant matching loss for training.
3) *Scalability and Sim-to-Real Robustness:* Our framework integrates several key components to ensure robust generalization: a CNN tokenizer that preserves small carriers, a duplication-noise augmentation for scaling across frequency windows, a curriculum learning strategy for handling closely collocated satellites, and a physics-based data generator with domain randomization.

III. System presentation

This section details the operational principles of the "Watchtower" system and the construction of its data products. The system comprises an Earth-based antenna that sweeps across the geostationary arc within its field of view, determined by its latitude. In our configuration, the acquisition field of view is fixed, spanning 120° along the geostationary arc, from $+60°$ to $-60°$ relative to the antenna's longitude. The sweeps are performed in discrete steps, with a fixed step size of $\Delta\theta = 0.2°$ for X band. This discrete scanning methodology on the geostationary arc minimizes distortions at the edges of the acquisition, thus maintaining an undistorted acquisition along the longitude axis relative to a constant step in antenna azimuth.

At each longitude step, the antenna acquires signals within a fixed frequency band. While this study focuses on the X band, our methodology is generalizable to the C, Ku, and Ka bands. The system constructs a power spectral density (PSD) by averaging the signal over 1024 FFTs. This process, which requires a few seconds to minutes (specifically, we fixed $\Delta t = 30$ seconds for the X band), depends on the frequency band. With this $\Delta t = 30s$ between each longitude step, a full acquisition takes $T = \Delta t \times \frac{2L}{\Delta\theta} = 5$ hours do be done.

At each longitude step, the received power corresponds to the signals emitted by satellites within the scanned frequency band that fall within the antenna's visibility cone, defined by its radiation pattern. For our antenna, this visibility cone is approximately 2 degrees at -3 dB at 7.7 GHz, as shown in Figure 4. Consequently, signal variation along the longitude axis is influenced by both the antenna's radiation pattern and the angular deviation between the satellite's position and the antenna's pointing direction, as illustrated in Figure 5. This projection of the radiation pattern is further affected by the modulation of the emitted signals. Signal variations from a satellite over time, even when averaged over 1024 PSDs, can introduce discontinuities in the received signal. For instance, modulations like TDMA and FM-TDMA [49] can cause signal splitting across multiple longitude steps.

Throughout the remainder of this article, we assume that the antenna's radiation pattern is uncalibrated (i.e., unknown or only coarsely characterized) and varies with the antenna's geometry, aperture, and operating frequency. It may also drift due to operational factors such as pointing bias, thermal deformation, or mount leveling errors. To handle this uncertainty, we treat the radiation pattern as a noised variable and train our models on a large library of synthetically generated patterns to improve robustness and generalization. Under this assumption, the satellite–antenna angular deviation cannot be estimated precisely from the measurements alone.

A. Physical model

The physical model of our system is composed of several sub-systems, each representing a specific component. We have structured these systems into the following key components:

1. Satellite position

The initial component of our physical model involves simulating satellite movement and trajectory. We employ a Keplerian propagator [50], [51], assuming elliptical orbits. The precision of the propagator is not critical, as the system's overall accuracy is limited by the antenna's longitude step size. With a step size of $\Delta\theta = 0.2°$ and an Earth-based antenna, the error along the geostationary arc is approximately 147 km. This level of error and a propagation period of a few hours (see T) make a Keplerian propagator sufficient, introducing negligible errors. These orbits are modeled using five Keplerian elements: eccentricity (e), semi-major axis (a), inclination (i), longitude of the ascending node (Ω), and argument of periapsis (ω). See Section 1 for details on how these elements are determined for each satellite in the scene.

2. Antenna mount

The antenna is mounted on an altazimuth mount, which operates based on azimuth and elevation. This allows the antenna to move both horizontally and vertically, enabling precise positioning and alignment along the geostationary arc. This dual-axis control is essential for tracking such orbits and avoids issues like discontinuities found in equatorial mounts. Modeling this type of mount in our system enables us to use 3D radiation patterns from a simulated parabolic antenna.

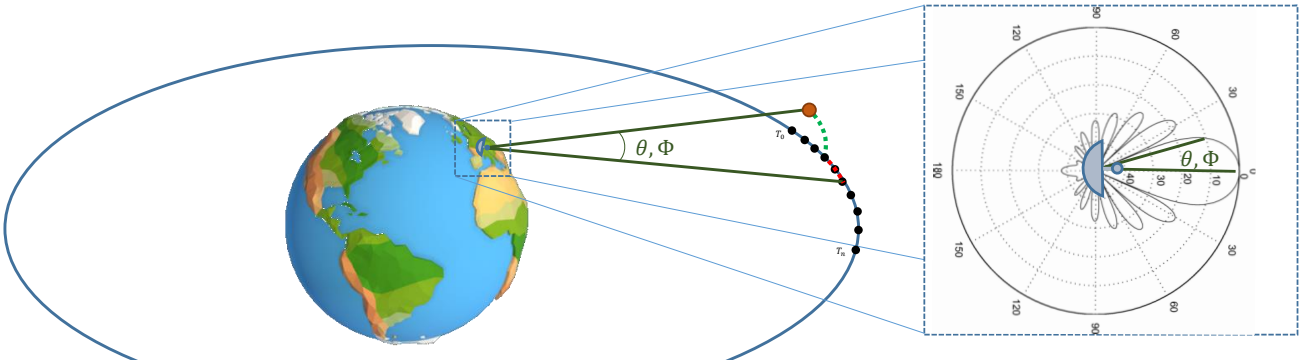

Fig. 3: Representation of the projection of a satellite for a fixed timestamp along the radiation pattern of the antenna for each azimuth and elevation angle.

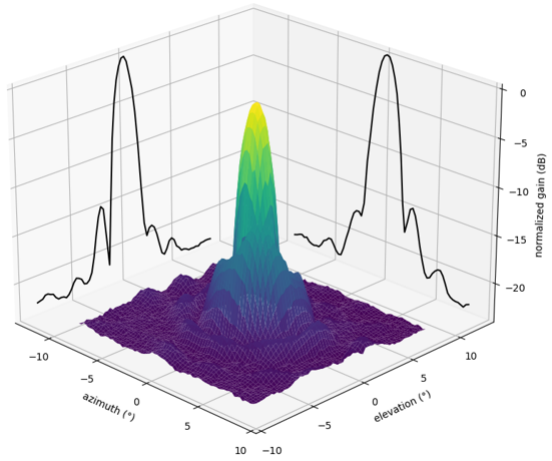

Fig. 4: Example of a measured radiation pattern in X-band, with a normalised gain of 23.34dB.

3. Antenna radiation pattern

To model the antenna's radiation pattern, we can use either empirical or analytical approaches. The first approach is to measure the antenna's radiation pattern in an anechoic chamber, and then use this pattern to generate our data. Although this method is simple and avoids complex modeling, we cannot consistently maintain a stable radiation pattern over time due to potential variations from RF source shifts or slight misalignment of the antenna. An alternative approach is to model the radiation pattern analytically [52], [53]. The electric field pattern $E_f(\theta)$ is given by

$$E_f(\theta) = \frac{2\lambda}{\pi D} \frac{J_1\left((\pi D/\lambda)\sin(\theta)\right)}{\sin(\theta)}, \tag{1}$$

where θ is the angle in radians from the antenna's symmetry axis, $D = 1.9\,\mathrm{m}$ is the antenna aperture, λ is the wavelength at the current frequency, and J_1 is the first-order Bessel function.

4. Reception model

The reception model integrates all previous components. Let S be the set of satellites; in our experiment, $\mathrm{card}(S) = 15$. Define $\Gamma_s(t, f)$ as the spectral activity of a satellite $s \in S$ at time t and frequency f. Let $\Psi_s(t)$ denote the spherical coordinates of the vector $\mathbf{Dir}(t)_{\mathcal{B}} - (\mathbf{Pos}_s(t)_{\mathcal{B}} - \mathbf{Pos}_{\text{antenna}}(t)_{\mathcal{B}})$ in the same reference frame as the antenna $\mathcal{B}$. Here, $\mathbf{Pos}_{\text{antenna}}(t)_{\mathcal{B}}$ is the antenna's position at time t (constant in ECEF), $\mathbf{Dir}(t)_{\mathcal{B}}$ is the antenna's direction at time t, and $\mathbf{Pos}_s(t)_{\mathcal{B}}$ is satellite s's position in ECEF coordinates at time t. Additionally, let ϵ represent additive noise, determined empirically based on real signal data. Is defined as :

$$R(f,l) = \frac{1}{1024} \sum_{i=1}^{1024} \sum_{s \in S} \int_{f-\frac{\Delta f}{2}}^{f+\frac{\Delta f}{2}} E_f \circ \Psi_s(T_i) \times \Gamma_s(T_i, \gamma) d\gamma + \epsilon, \tag{2}$$

with $T_i = t + i\frac{\Delta t}{1024}$.

B. Data generation and implementation details

1. Data generation

The use of a physical model allows for the generation of a synthetic dataset, enabling the testing and evaluation of our method with labeled data. However, training a neural network on synthetic data presents challenges, as highlighted in [54], [55], [20], and may not generalize to real-world data [55], [56]. These challenges stem from two primary issues.

The first reason is that training a neural network on synthetic data does not guarantee convergence over real data, as shown in [56], [57], [58]. The generated distribution and the real distribution are independent and may not overlap, as illustrated in Figure 6. Many sim-to-real data strategies, such as those in [57], [59], introduce reinforcement learning as a post-training step on real data. However, our network is optimized using the extracted physical model of the antenna and orbits, meaning the difference between generated and real data can be attributed to our model itself, or the lack of knowledge regarding unmodeled effects.

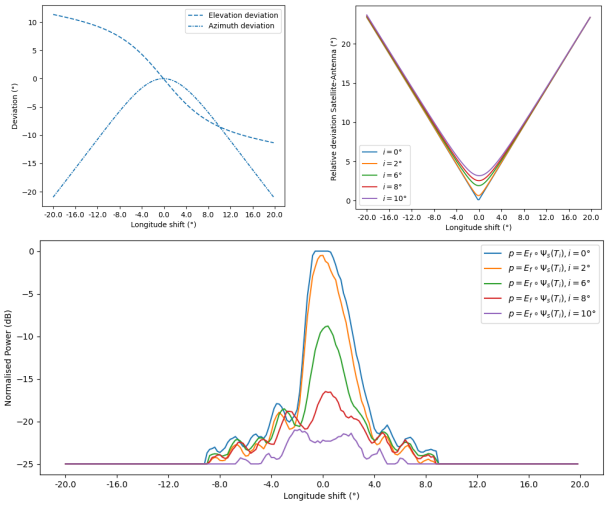

Fig. 5: **Simulation results illustrating satellite signal reception**. **Top left:** Azimuth and elevation deviation of a geostationary satellite ($i = 0°$) using an altazimuth antenna mount. **Top right:** Angular deviation between the antenna's pointing direction and the satellite's position. It should be noted that satellite inclination causes this angle to vary, even when the antenna is pointed towards the satellite's current longitude. **Bottom:** Received signal resulting from the interaction between the antenna's radiation pattern and the angular deviation to the satellite.

	Property	Variable	Probability law
	Num satellite per acquisitions	n	$\sim \mathcal{U}(\llbracket 0, 15 \rrbracket)$
Orbit	Inclination	i_k	$\sim \mathcal{U}([-5°, 5°])$
Orbit	Eccentricity	e_k	$\sim \mathcal{U}([0, 1[)$
Orbit	Semi-major axis	a_k	$\sim \mathcal{U}([0, 2\pi])$
Orbit	Longitude of the ascending node	Ω_k	$\sim \mathcal{U}([0, 2\pi])$
Orbit	Argument of periapsis	ω_k	$\sim \mathcal{U}([0, 2\pi])$
Spectrum	Num carrier	c_k	$\sim \mathcal{U}(\llbracket 0, 15 \rrbracket)$
Spectrum	Carrier types	t_k	$\sim \mathcal{U}(\{$ Classic, TDMA, TM , ...$\})^{c_k}$
Spectrum	Carrier power	$p_{k,l}$	$\sim \mathcal{U}([0, 1])$
Spectrum	Carrier bandwidth	$bw_{k,l}$	$\sim \mathcal{U}(]0, 30e6])$

TABLE I: Probability laws for each parameter that define an acquisition with the goal of building a sufficiently large dataset distribution that encompasses the real data distribution.

	Classic	RPO	Deployment	Colocalised	Random Drift
Probability	0.7	0.05	0.05	0.1	0.1

TABLE II: Probability for each events corresponding to a satellite class, empirically fixed.

The second reason is that synthetic data may not capture the full range of variations in real-world data, resulting in a neural network that performs well on synthetic data but poorly on real data. To mitigate this, our synthetic data are designed to encompass as many realistic variations as possible, and the network performance is validated against real data where available.

However, such an approach also has advantages, as our data generator can also create scenarios that are rare or underrepresented in real datasets. For example, RPO [60], [61] and deployment events are relatively rare in real-world data [60]. Thus, our generator enables us to produce more samples of these cases, balancing these events, and avoiding biases in training.

Sampling from these distributions yields a dataset that spans a broad range of scenarios and closely mirrors real acquisitions in terms of satellite orbit, carrier occupancy, and the induced structure of the projected radiation pattern. This, in turn, helps the neural network generalize to real data. To explicitly stress robustness, we sample orbital elements that produce noticeable distortions in the projected radiation pattern (see Fig. 5); relying solely on strictly geostationary satellites would not expose the model to such effects.

We further constrain the orbital elements so that most objects remain near the geostationary arc, consistent with the limited integration time and the practical inability to observe LEO/MEO traffic within a sweep. Concretely, we narrow the ranges of inclination, semi-major axis, and eccentricity for computational efficiency and to ensure that, at some acquisition step, each satellite enters the antenna's visibility cone. These constraints also motivate bounds on the argument of periapsis ω_k, $k \in \llbracket 1, n \rrbracket$.

For each satellite, we optionally apply an event (spatial or spectral) with probabilities given in Table II. We model four events: RPO, satellite deployment, collocated satellites (multiple satellites within $1°$ of longitude), and random drift. These processes tend to produce small longitudinal separations between satellites. Given the discretization step $\Delta\theta = 0.2°$, we must account for the system's source-separation limit, which we model using the Rayleigh criterion (Fig. 7).

Finally, we inject measurement noise ϵ to promote robustness. In our setting, noise is heterogeneous and arises from terrestrial emitters, thermal noise, and RF front-end impairments. We therefore adopt a frequency-dependent (non-constant) noise floor estimated empirically, rather than assuming a stationary white process.

2. *Implementation details*

The generator is purpose-built for machine learning: it must produce large training corpora for neural network training (see Sec. C). We therefore designed it to be GPU-efficient and highly parallel. The implementation uses

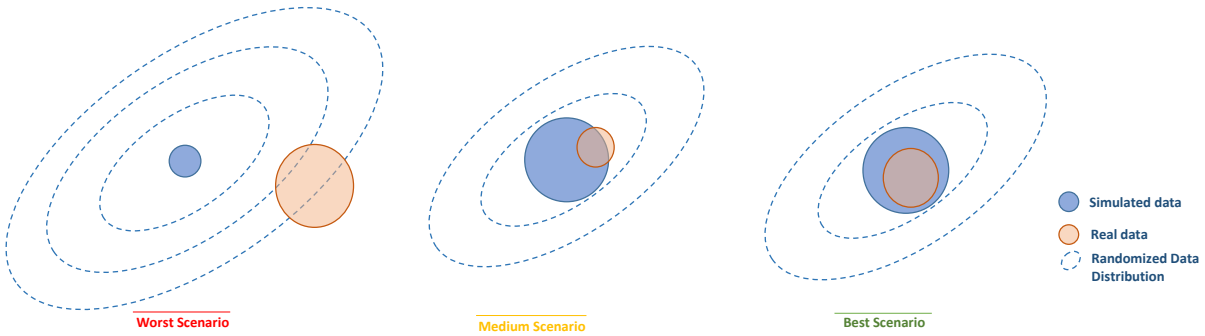

Fig. 6: Inspired by [57], this figure illustrates the concept of domain randomization, emphasizing the importance of aligning simulated and real data distributions. Ideally, the real data distribution should be entirely encompassed by the simulated data distribution; the optimal scenario occurs when both distributions are identical.

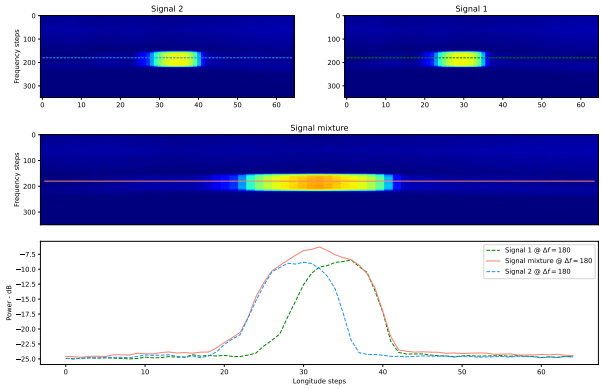

Fig. 7: Example of overlapping carriers from two different satellites that sum and mix into a single signal under the angular resolution defined by the Rayleigh limit.

	VRAM (Mo)	vram/sample (Mo)	Time (ms)	s/sample (ms)
300x200 ($b = 20$)	140	7	660	33
3000x200 ($b = 20$)	1000	50	6800	340

TABLE III: Performance to generate data samples for a batch of 20 images.

CUDA kernels and PyTorch. As reported in Table III, these optimizations allow us to generate a full dataset (500k examples) in a few hours on NVIDIA V100 GPUs.

Satellite trajectories are propagated with a Keplerian model, which is computationally inexpensive and yields limited drift over the acquisition horizon T compared with higher-fidelity propagators.

The main performance bottleneck is the synthesis of telemetry carriers over a wide bandwidth, which demands fine frequency resolution. In practice, we render these carriers on a frequency grid that is $20\times$ finer than the acquisition's discretization step, and this stage dominates the computational cost.

IV. Proposed architecture and training

The purpose of using machine learning here is to achieve tasks that standard expert algorithms cannot. Specifically, we aim to infer satellite information from each acquisition. This requires extensive knowledge of acquisition and satellite signal structure. Essentially, extracting a satellite from an acquisition involves detecting, extracting, and associating all signals representing carriers present for each satellite, which can be viewed as a blind source separation problem. These high-level features can be broken down into lower-level detection and characterization tasks.

The straightforward approach would be to extract a separate acquisition for each detected satellite, similar to standard detection methods (e.g., Unet [62], Swin Unter [43]). However, this approach demands excessive memory and computational power due to large image sizes, especially in the frequency domain. Instead, our approach extracts information that requires similar knowledge and features. Thus, we extract the peak coordinates in both longitude and frequency for each satellite. Let $R_s(f, l)$ be the signal for a unique satellite based on (2):

$$R_s(f,l) = \frac{1}{1024}\sum_{i=1}^{1024}\int_{f-\frac{\Delta f}{2}}^{f+\frac{\Delta f}{2}} E_f \circ \Psi_s(T_i) \times \Gamma_s(T_i,\gamma)d\gamma. \tag{3}$$

We want to infer:

$$f_s = \max_f R_s(f,l) \text{ and } l_s = \max_l R_s(f,l), s \in S \tag{4}$$

A. Constraints on the architecture

Overall, we want to define a parametric model that takes a (longitude, frequency) image I (see Fig.1) and outputs a list of (spatial, spectral) signatures, one pair of vectors per satellite, as defined Equation 4.

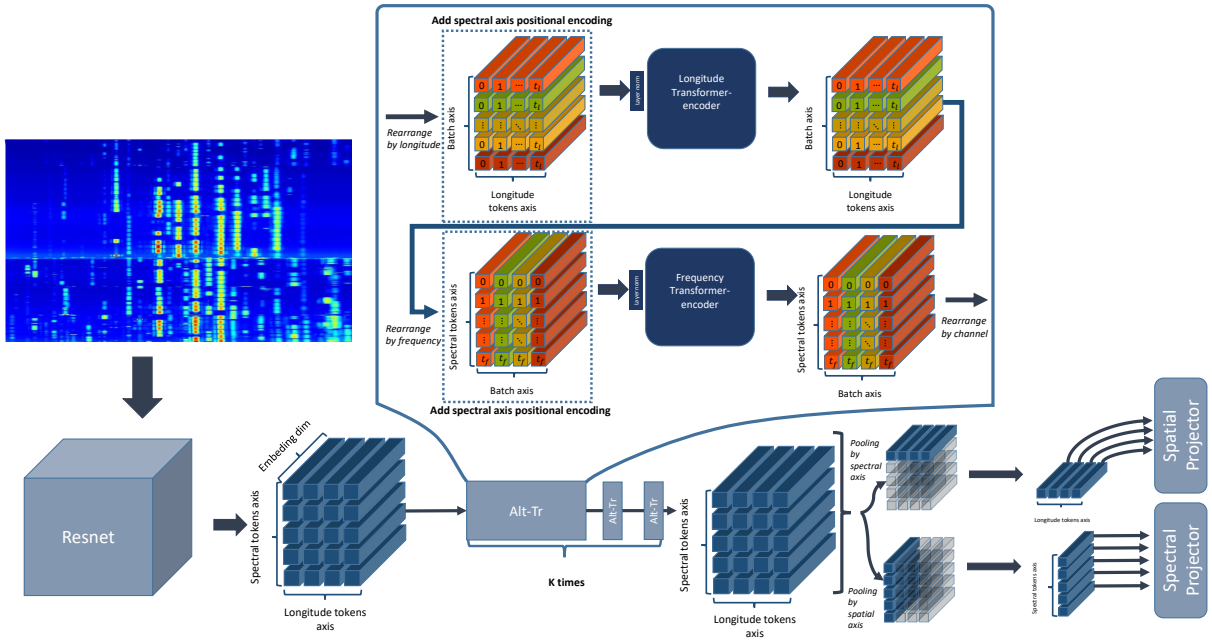

Fig. 8: Overview of the proposed model architecture. The input is a large, single-channel acquisition. Patch embeddings are computed using a ResNet-18 architecture [63]. These embeddings are processed sequentially by alternating transformer layers applied along the frequency and longitude axes. Finally, the patches from each axis are projected into the initial domain for each detected satellite. The projector uses linear layers.

This model architecture must respect some key properties, due to the nature and size of the data to be processed. First, input images are very large, covering a wide frequency band from 8 to 12 GHz in the X band. For instance, a discretization step of $\Delta f = 30kHz$ results in acquisitions with thousands of frequency bins (approximately 16,000 for our X-band acquisitions, ranging from 7.250 to 7.750 GHz; and up to 60,000 in C and Ku band). As discussed in Section III, processing such large images is difficult [64] due to spatial and temporal limitations on GPUs. Furthermore, Transformer encoders scale poorly with long sequences because self-attention forms the dense score matrix $QK^\top$, with Q the queries and K the keys. Therefore, a vanilla Vision Transformer (VIT) [39] is unsuitable due to its memory consumption and poor scaling. However, our system does not require real-time inference, allowing for a broader range of solutions.

Another constraint is that our architecture should be flexible in terms of frequency windows. Our goal is to develop an architecture that can generalize over multiple bands and frequency windows, which can range from 1 to 1000 times the size of the training data. To achieve this generalization, we utilize two properties of our acquisition.

Frequency invariance: This property allows us to make multiple approximations. As shown in the physical model A, only the radiation pattern depends on the frequency (see (3)). The radiation pattern is broader at lower frequencies and more directive at higher frequencies. Carriers can also be assumed to be uniformly distributed across the frequency band. With this property, we can train on smaller acquisitions and still generalize across the full acquisition range.

Separability: This property implies that our frequency/longitude acquisition has orthogonal and independent axis correlations. A carrier's bandwidth and frequency do not depend on the longitude. This also allows us to treat the longitude and frequency axes independently, similar to structured spectrograms [42]).

These properties enable an architecture that is more memory-efficient than standard Vision Transformers (VIT) for large images and that can generalize independently of the frequency window size. This generalization depends on two intrinsic properties. First, the architecture needs to extract local information from acquisitions, such as information about single carriers (e.g., type, frequency, bandwidth). Second, the transformers layers must extract information over a wide window and share information between distant tokens (e.g., using attention mechanisms); for example, the neural network need to associate carriers that are emitted from the same satellites.

B. Definition of the architecture

The proposed architecture naturally arises from the properties presented earlier. The overall architecture is shown in Figure 8. It includes 3 different blocks implementing 3 different elements of the model. First, a CNN is applied to the input image. This block uses a ResNet-18 [63] with max pooling as residual connections to preserve small objects and carriers. Thus, the CNN serves

as a tokenizer for the next transformer layer. Second, we employ alternating transformer blocks, each composed of two successive transformer encoder blocks along the chosen axis. This addresses the separability properties mentioned earlier. Third, a decoder head infers the spatial and spectral footprints.

These three blocks are described in the following subsections.

1. *Image encoder*

Having a convolution based image encoder has several advantages. The received signal is well-structured, and a CNN as a tokenizer and local feature extractor makes sense because CNNs can easily converge to filters that extract the underlying radiation pattern shape and higher-level features. However, the CNN's depth must be appropriately set. The convolution windows depend on the filter size and CNN depth, edge effects can appear, potentially making the alternate transformer block unnecessary if the encoded tokens see every other token (i.e., when the CNN window is larger than the image). This hinders generalization across various frequency windows. A key advantage is that a CNN with well-fixed depth can spatially encode tokens, allowing us to forgo positional encoding (i.e., no positional encoding (NOPE) [65]), as relative positions can be encoded between tokens given the low constraints on the frequency axis and image structure. However, PE was added to ensure the encoding of each token's position.

2. *Alternating Transformer Architecture*

Based on the previous considerations, we introduce the alternate layer for transformer (Alt), an architecture composed of two successive transformer encoder blocks that aggregate information along two orthogonal spatial axes of the image. In particular, this architecture is equivalent to a series of transformer layers with axis-specific masking.

Let $X \in \mathbb{R}^{n_f \times n_l}$ denote the input acquisition, where n_f is the number of frequency bins and n_l the number of spatial bins (e.g., longitude). After tokenization via a convolutional encoder, the data is transformed into a tensor $T \in \mathbb{R}^{c \times t_f \times t_l}$, where c denotes the number of channels, and t_f and t_l are the reduced spatial dimensions along the frequency and longitude axes, respectively.

To encode positional information, we apply additive positional encoding along both axes. The resulting tensor $T^{pe} \in \mathbb{R}^{c \times t_f \times t_l}$ is given by:

$$T^{pe}_{:,i,:} = T_{:,i,:} + PE(c, t_l), \quad T^{pe}_{:,:,j} = T_{:,:,j} + PE(c, t_f),$$

for all $1 \leq i \leq t_l$, $1 \leq j \leq t_f$, using the standard sinusoidal encoding from [39].

We propose an alternating transformer structure in which self-attention is successively applied along orthogonal axes. Let $Tr^t_{\text{lon}}/Tr^t_{\text{freq}}$ a transformer layer for the longitude/frequency axis at stage $1 \leq t \leq d_{lon}/d_{freq}$. First, we process the longitude axis. Given $K^t \in \mathbb{R}^{c \times t_f \times t_l}$ the current stage of the network, we reshape it into a batch of t_f sequences of length t_l:

$$K^{(t)}_{\text{lon}} = \left[K^{(t)}_{:,1,:},\ K^{(t)}_{:,2,:},\ \ldots,\ K^{(t)}_{:,t_f,:}\right].$$

Each sequence is independently processed by a stack of d transformer encoder blocks:

$$K^{(t)}_{\text{lon}} = Tr^t_{\text{lon}}(K^{(t-1)}_{\text{lon}}), \quad K^{(0)}_{\text{lon}} = T^{pe}, \quad 1 \leq t \leq d_{lon}.$$

We then process the frequency axis using a symmetric operation. From the same tensor $K^t \in \mathbb{R}^{c \times t_f \times t_l}$, we construct:

$$K^{(t)}_{\text{freq}} = \left[K^{(t)}_{:,:,1},\ K^{(t)}_{:,:,2},\ \ldots,\ K^{(t)}_{:,:,t_l}\right],$$

which represents a batch of t_l sequences of length t_f, passed through a second stack of d' transformer layers:

$$K^{(k)}_{\text{freq}} = Tr^k_{\text{freq}}(K^{(k-1)}_{\text{freq}}), \quad K^{(0)}_{\text{freq}} = K^{(d_{lon})}_{lon}, \quad 1 \leq k \leq d_{freq}.$$

The alternating application of attention along both axes allows the model to efficiently capture 2D dependencies. Since the data structure is separable, the order of axis processing is not critical: the first transformer stack can approximate the identity if needed, enabling implicit axis reordering.

For the complete architecture, we stack multiple layers of alternating transformers, which aggregate and propagate information across all tokens in both the longitude and frequency dimensions.

Theoretical complexity gains

This approach leads to a significant performance gain. A standard Transformer layer has a theoretical minimum complexity of $O(t_l^2 t_f^2 c)$, as shown in [66]. By processing reduced sequence lengths and treating one axis as a batch, we eliminate the quadratic term for that axis. By alternating transformer layers (horizontal and vertical axes), we reduce the overall complexity to $O(t_l t_f^2 c + t_l^2 t_f c)$, which is cubic rather than quartic as in a vanilla attention layer.

3. *Decoder head*

The decoder head is a projection from the latent space of the spatial and frequency information, providing a reconstruction of the longitude footprint and associated spectrum for each satellite present in the image (see Figure 8). It is important to note that the outputs are unordered and do not follow any predefined sequence. To train the network, we match these predictions with the targets using the Hungarian algorithm. This strategy is not optimal and can lead to issues of responsibility [67] and discontinuities in the network [68]. Although various methods have been proposed to address these issues by using permutation-invariant outputs, as in [69], they are not used in this work, as such issues were not observed during training.

C. Dataset and loss function

Synthetic Dataset. The model is trained on a large-scale synthetic dataset generated by the physics-informed simulator detailed in Section A. The data set consists of

500k training samples, 10k validation samples, and 10k test samples. Each sample is a 2D acquisition of size 3072×200 pixels (frequency × longitude). The composition of each scene, including the number of satellites, carrier properties, and orbital elements, is governed by the probability distributions in Tables I and II. For supervised training, we store the ground-truth spatial and spectral footprints for each satellite. **Real-World Dataset.** For evaluation, we use real-world X- and Ku-band acquisitions from the Watchtower system. Raw measurements are processed into non-overlapping 3072×200 crops. Ground-truth labels for this dataset are generated through manual annotation and are therefore considered weak labels.

1. Data augmentation

Firstly introduced in [70]. To achieve better generalization over the size of the frequency window, we tested and implemented a data augmentation strategy that artificially duplicates tokens along the frequency sequences (see Figure 8). This strategy takes advantage of the fact that tokens along the frequency axis (i.e. K_{freq}) are processed independently by each transformer operator, except for the attention mechanism.

Thus, the modified attention mechanism for the frequency transformer is as follows.

$$\text{Attention}_\alpha(Q,K,V) = \text{Softmax}_j\left(\frac{QK^T}{\sqrt{d_n}} + \log(\alpha)\right)V, \tag{5}$$

where Q, K, and V are the query, key, and value matrices, respectively, and $\alpha \in \mathbb{N}\setminus\{0\}^n$ represents the duplication factor for each token. In this formulation, the duplication factor α is treated as a weight that adjusts each token in the sum of the matrix product between the attention matrix and the values matrix. By integrating this factor into the attention matrix weights, we consider the vector α as a noise element that disrupts the softmax operation by altering the contribution of each token in the sum of token weights. For our experiment, we imposed a constraint on the α tensor such that $\max_{\alpha\in\mathbb{N}\setminus\{0\}^n}\sum_k \alpha_k \leq 200$. This implies that $\log(\alpha)$ is in a range of $[0, \log 200) \approx 5.3]$ and can be seen as adding noise in the model, that we deactivate during the inference time.

2. Curriculum learning

To properly address the issue of matching noise and errors at the beginning of training caused by the Hungary algorithm and leading to learning instabilities, we employ a curriculum learning strategy [71]. The primary matching difficulties arise with co-located satellites, i.e., those that are spatially close in terms of longitude. The main challenge lies in separating satellites that differ by only one longitudinal bin, where their predicted spatial footprints are very similar, leading to incorrect associations.

To mitigate this, we control the difficulty of the problem by adjusting the minimal longitudinal distance between satellites (see Figure 9). We begin with simpler training scenarios, where satellites are widely spaced, and gradually introduce more challenging cases by reducing the distance between them. This approach stabilizes the matching process in the early stages of training and ultimately leads to more robust network learning.

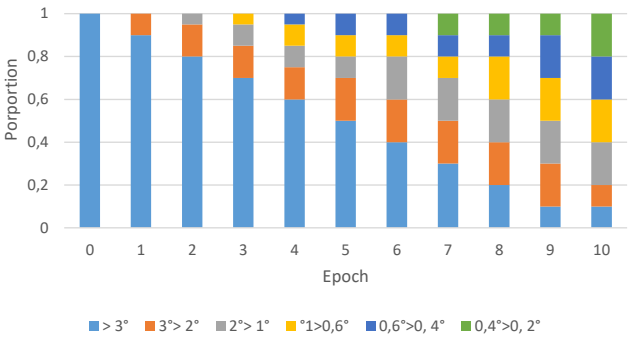

Fig. 9: Evolution of the difficulty for each set at each epoch. The idea is to have an increasing difficulty, which in our case is related to the proximity between satellites.

3. Loss function

The choice of loss function is critical for achieving good model performance and generalization. A key characteristic of our model is that it outputs unordered longitude footprints and associated spectra for each satellite in the image, similar to object detection algorithms. To effectively train the model, a method is needed to match these outputs with the ground truth. We accomplish this by using the Hungarian algorithm, which finds the optimal permutation and matches between the predicted longitude/spectral footprints and their corresponding labels.

Let Π the space of all n-length permutations, with n the number of predicted signals. Let $\hat{l}_i^\theta$ and $\hat{f}_i^\theta$ the spatial and spectral footprints predicted by the model for the i-th satellite with the neural network parameters θ, and l_j and f_j the ground truth signature. By applying the Hungarian algorithm, the goal is to find the permutation which solves:

$$\hat{\pi} = \arg\min_{\pi\in\Pi}\sum_{i=1}^{n}(1-\lambda_{match})\mathcal{L}(\hat{l}_i^\theta, l_{\pi(i)}) + -\lambda_{match}\mathcal{L}(\hat{f}_i^\theta, f_{\pi(i)}). \tag{6}$$

Then the loss function, with respect to the neural network parameter θ, is defined as :

$$E(\theta) = \sum_{i=1}^{n}(1-\lambda_{train})\mathcal{L}(\hat{l}_i^\theta, l_{\hat{\pi}(i)}) + \lambda_{train}\mathcal{L}(\hat{f}_i^\theta, f_{\hat{\pi}(i)}), \tag{7}$$

where $\mathcal{L}$ is the Mean Square Error (MSE) between the signatures. During training, the gradient of $E(\theta)$ is computed using automatic differentiation (autodiff).

The hyperparameter $\lambda_{match/train} \in [0,1]$ governs the trade-off between spatial and spectral information in the comparison. Notably, the value of λ_{match} in (6) (matching criterion) and λ_{train}in (7) (training loss function) can differ. Through experimentation, we found that the Hungarian matching error should prioritize the spatial

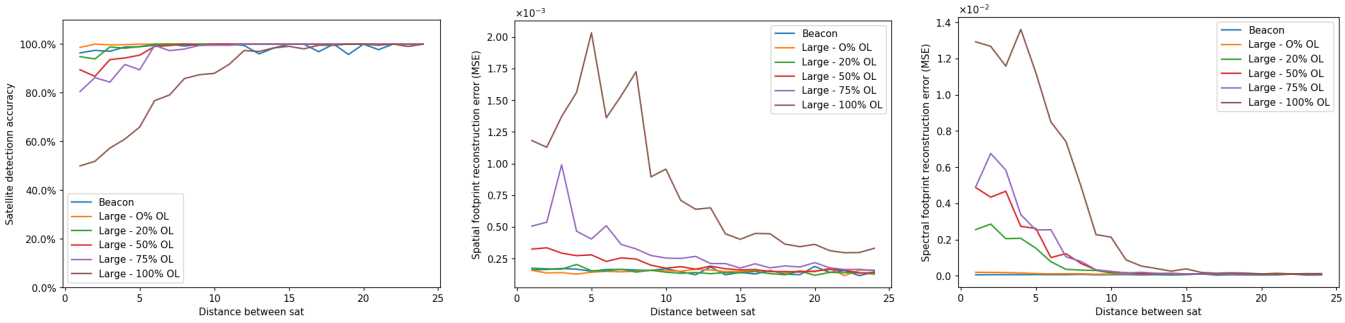

Fig. 10: Performance of the proposed architecture in multiple critical scenarios: we tested the model with varying distances between two satellites in terms of longitude steps. For each distance, various carrier configurations were evaluated, including small carriers (such as beacon carriers) and overlapping carriers. Overlap percentages were also considered. We assessed the model's detection accuracy, as well as spectral and spatial footprint reconstruction errors

position (longitude) of the deformed radiation pattern over the frequency footprint, leading to an optimal value of $\lambda_{match} = \frac{1}{6}$. However, empirical results indicate that a position-based loss function is suboptimal for training, where a balanced trade-off with $\lambda_{train} = \frac{1}{2}$ yields better performance.

V. Experimental Setup

A. Tasks and Metrics

We evaluate five aspects:

- **Satellite detection:** accuracy/precision/recall after Hungarian assignment. A prediction is considered a successful match to a ground-truth target if two conditions are met: (i) the spectral-peak error is within a tolerance τ, and (ii) the predicted spectral footprint has contains at least a proportion γ_f within the ground-truth spectral support, which is dilated by a margin of $\pm\tau$. First we define the set of elements that are in the margin of the GT such as

$$\mathcal{S}_i^\tau = \{\hat{f}_{i,j}^\theta \in [f_{i,j}^\theta - \tau; f_{i,j}^\theta + \tau] | 1 \leq j \leq n_f\}$$

  Finally, a satellite is characterized as a positive detection if :

$$\frac{card(\mathcal{S}_i^\tau)}{n_f} \geq \gamma_f$$

  where $\hat{f}_i^\theta$ is the predicted spectrum of a satellite i.
- **Longitude error:** MAE and std (in bins and degrees) of the spatial footprint argmax.
- **Spectral reconstruction:** MSE and std between predicted and target spectra per matched satellite.
- **Separation rate under proximity/overlap:** success rate when two targets are within $\Delta l \in \{1, 2, 3, 5\}$ bins and spectral overlap $\in \{0\%, 25\%, 50\%, 75\%, 100\%\}$.
- **Robustness to SNR:** detection accuracy stratified by carrier power above the local noise floor (bins: {0–3, 3–6, 6–9, >9} dB).

B. Baselines and Ablations

We compare Alt-Tr to: (i) a strong **CNN** tokenizer-only model with global pooling, (ii) **Alt-Tr (no-CNN)** tokens from raw patches. We ablate the proposed components: axis alternation (replace by a single 2D attention), positional encoding (remove PE), curriculum learning (disable), duplication noise α (disable).

C. Implementation Details

Our models were trained using the AdamW optimizer with a cosine learning rate schedule and a one-epoch warm-up, implemented in bfloat16 for efficiency. The CNN tokenizer is a ResNet-18 with max-pooling residual connections. Our Alternating-Transformer applies self-attention successively along longitude and frequency (Sec. 2) with 2D sinusoidal positional encodings. We use a simple curriculum on the minimum longitudinal gap and a duplication augmentation for scaling; predictions are matched with a Hungarian assignment. Exact hyperparameters (optimizer β's, weight decay, LR bounds/warm-up, batch size/epochs, dataset sizes, architectural widths/depths, loss weights, and tolerances) are summarized in Table IV.

VI. Results

In this section, we present the results of our proposed method and training framework. We aim to assess the model's performance across several tasks, including satellite detection, longitude pattern reconstruction, spectral pattern reconstruction, and separation performance in collocated satellite scenarios.

Component	Hyperparameter	Value / Description	Ref.
Optimization	Optimizer	AdamW (β_1=0.9, β_2=0.999), weight decay 0.05	Sec. C
	LR schedule	Cosine decay: $1\times10^{-4} \rightarrow 1\times10^{-6}$, warm-up 1 epoch	Sec. C
	Batch size / Epochs	**24 / 50**	Sec. C
	Precision	bfloat16	
Data	Train/Val/Test	500k / 10k / 10k (size 3072×200)	Sec. C
	Curriculum (min gap)	15→1 bins over first 10 epochs (linear)	Sec. 9
	Duplication α	$\sum_k \alpha_k \leq 200$; disabled at inference	Sec. C
	Noise, shifts	Empirical noise floor; random freq/lon flip	Sec. 1
Model	CNN tokenizer	ResNet-18 with max-pooling residual connections	Sec. 8
	Embedding dim	D=384 ; Heads H=6 ; MLP ratio 4.0 ; Dropout 0.0	
	Alt-Tr depth	Longitude stack d=3 ; Frequency stack d'=3 ; #Alt blocks 2	Sec. 8
	Positional enc.	Sin-longitude + Sin-frequency	
Loss/Matching	Matching λ	$\lambda_{\text{match}} = \mathbf{1/6}$ (balanced)	Eq. (6)
	Training λ	$\lambda_{\text{train}} = \mathbf{1/2}$ (balanced)	Eq. (7)
Hardware	A100 80GB × 4; PyTorch 2.5.0, CUDA 12.2		

TABLE IV: Training and model hyperparameters.

A. Results on Generated Data

1. Reconstruction error

First, we analyze the reconstruction errors for the spatial and spectral footprints in the test set (Table VII). While absolute values are not immediately intuitive, they enable fair comparisons across methods. Compared to a naive CNN, our Alt-Tr (ours) achieves substantially lower errors (≈5.3× lower spatial MSE and ≈8.3× lower spectral MSE). A vanilla Vision Transformer is impractical at our resolution due to quadratic memory scaling (out-of-memory).

Ablation study. We performed an ablation study to isolate the contribution of each key component of our framework. The results underscore the importance of our training strategy: removing curriculum learning caused the most significant performance degradation, confirming its critical role in stabilizing the permutation-invariant matching process. The data augmentation and positional encoding (PE) were also shown to be beneficial, as their removal consistently increased reconstruction errors on both axes. Finally, replacing the CNN tokenizer with raw patches also harmed performance, highlighting the tokenizer's effectiveness in extracting robust local features and preserving the structure of small carriers before they are processed by the attention layers.

2. Detection accuracy

Figure 12 presents the model's accuracy, precision, and recall as a function of signal coverage within a given tolerance interval, τ. The results demonstrate the model's robustness across different operating points. Under stringent requirements ($\tau = 0.25\,\text{dB}$), the performance degrades gracefully as coverage increases, which is the expected behavior for a tight error margin. For more typical tolerance levels ($\tau = 1.25\,\text{dB}$), the model achieves a strong balance of precision and recall, maintaining high performance even at high coverage rates. For more lenient applications ($\tau \geq 2.5\,\text{dB}$), the detector remains highly accurate across nearly all coverage levels, confirming its reliability in diverse operational scenarios.

By using synthetic data, we can explore a multitude of scenarios regarding satellite positions, events, and spectral footprints. Therefore, we applied our metrics to several datasets, each representing various scenarios of interest. We selected three significant case studies: satellites with small-bandwidth carriers, satellites with large non-overlapping carriers, and collocated satellites with overlapping carriers. These cases assess the model's ability to reconstruct small spectral details and separate carriers in scenarios with collocated satellites that have spectral overlap.

As shown in Table V, small carrier bandwidth has a minimal impact on satellite detection accuracy. However, the slight reduction in performance (approximately 1%)

results from the network needing to generalize over small signals that can easily be masked by acquisition and model noise. Acquisition noise is especially detrimental because low-intensity carriers lead to a reduced signal-to-noise ratio (SNR), making detection more challenging. As shown in Figure 11, detection performance is significantly affected for non-overlapping low-power carriers as the carrier power nears the noise level. In such cases, accuracy drops drastically until the signal becomes undetectable. However, for carrier powers above 5 dB, performance remains very good, with detection accuracy consistently exceeding 95%.

Dealing with larger carriers is easier due to their larger footprint and distinct structures in the acquisition. Thus, the best performance is achieved when non-overlapping large signals are present, as seen in Table V.

Figure 10 shows the model's response with varying satellite proximity and different levels of carrier overlap. When satellites are collocated with high overlap, distinguishing them becomes difficult. The merging of individual signals into a single composite signal indicates the satellites' proximity falls below the Rayleigh resolution limit (see Figure 7). This behavior is evident in the results for high overlap ratios (100% to 75%) in Figure 10. Such scenarios significantly impact spatial reconstruction error, increasing the overall error rates. Another effect is the presence of equivalence classes, unsolvable when the same carrier is shared for distinct satellites. In such cases, it is impossible to determine if the signals are from distinct satellites or a single inclined satellite, causing the appearance of two main lobes (see Figure 5).

Other noteworthy results concern telemetry carriers, which often have a narrow frequency bandwidth and appear as small signals. Telemetry carriers can occupy a single frequency step in the data, especially beacon carriers (pure sinusoidal signals). These carriers are important because they can easily be lost within the model's noise if not addressed by the training loss. However, as shown in Figure 10, our method can detect such carriers effectively, with an accuracy of about 97% for satellites collocated at a single longitudinal step ($\Delta l = 0.2°$).

3. *Computational costs*

Table VI compares inference and training times for each architecture. Our proposed Alt-Tr model achieves a strong balance between speed and performance, with high inference efficiency (186 samples/sec) and moderate training cost. In contrast, the vanilla Vision Transformer is significantly slower due to its memory-heavy attention mechanism. Removing the CNN from Alt-Tr increases training time, highlighting the CNN's importance for efficient tokenization. While the CNN-only model is slightly faster in inference, it lacks the accuracy and robustness of Alt-Tr. These results confirm that Alt-Tr is well-suited for large-scale RF analysis, offering both scalability and efficiency.

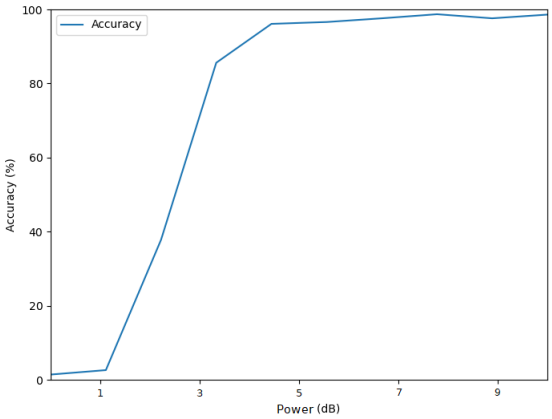

Fig. 11: Response of the model to low-power carriers. We evaluated the detection performance when large carriers have low power (between 0 dB and 9 dB above the noise level). The results suggest that our model is not significantly affected by low-power carriers when they are visible.

	Accuracy	Error in longitude (Δl)
Only small carriers (bw $<$ 200kHz)	98.9%	0.82 ± 0.66
Large carriers (overlap 0%)	99.9%	0.80 ± 0.61
Large carriers (overlap 25%)	99.5%	0.83 ± 0.65
Large carriers (overlap 50%)	98.5%	0.84 ± 0.71
Large carriers (overlap 75%)	97.4%	0.86 ± 0.72
Large carriers (overlap 100%)	88.5%	0.93 ± 0.81

TABLE V: Results of the proposed model across various synthetic satellite detection scenarios. The accuracy pertains to the correctly detected satellites, while the longitude error represents the difference between the predicted and ground truth positions of the spatial footprint's maximum (i.e $|\text{argmax}(\hat{l}_i) - \text{argmax}(l_{\pi(i)})|$. For a test set of 2000 samples.

B. Results on real data

Obtaining real-world results is difficult due to a lack of large labeled datasets. Therefore, we trained our model on synthetic data. Figures 13, 14, and 15 show qualitative results on real data, focusing on spatial reconstruction, satellite separation, and detection. The model shows good qualitative performance in satellite detection and spatial reconstruction, even with collocated satellites. The method handles collocated satellites well, especially in

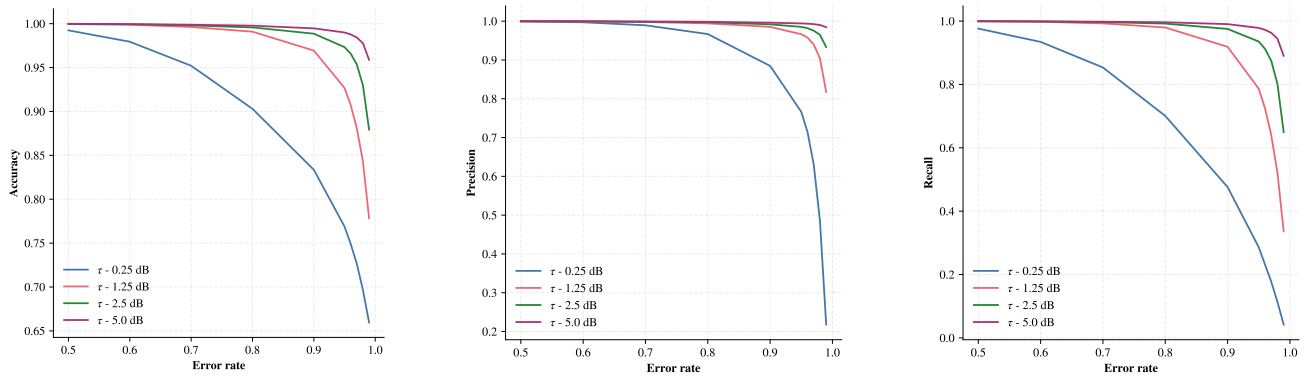

Fig. 12: Accuracy, precision and recall curves in case of multiple error rate i.e γ_f. The error rate define the minimal proportion of signal that need to fall in the tolerance window to be classified as a positive detection. The τ denotes the maximum tolerated shift between the reconstructed footprint and the ground truth.

	Inference (eval)	Inference (train) + backpropagation
Alt-Tr (ours)	186±5.1	68±3.2
Alt-Tr without CNN	320±6.8	123±5.8
CNN	216±7.1	81±3.5
Vanilla transformer [39]	2.8±0.1	1±0.1

TABLE VI: Number of inference per second for all evaluated architectures with image size 3072×200. All experiments were conducted on a NVIDIA A100 GPU with bfloat16 precision.

	Spatial pattern error (MSE)	Spectral pattern error (MSE)
Alt-Tr (our)	$\mathbf{6.9e{-}4}$ ± *1.3e-4*	$\mathbf{2.55e{-}3}$ ± *2.1e-3*
Alt-Tr without data aug	$7.6e{-}4$ ± *1.2e-4*	$2.84e{-}3$ ± *2.2e-3*
Alt-Tr without curriculum learning	$11.5e{-}4$ ± *8.6e-4*	$6.72e{-}3$ ± *15.3e-3*
Alt-Tr without PE	$7.5e{-}4$ ± *2.3e-4*	$3.46e{-}3$ ± *2.3e-3*
Alt-Tr without CNN	$9.6e{-}4$ ± *2.7e-4*	$3.71e{-}3$ ± *2.6e-3*
CNN	$36.6e{-}4$ ± *7.3e-4*	$21.26e{-}3$ ± *15.6e-3*
Vanilla transformer [39]	OOM	OOM

TABLE VII: Results on the test set distribution for the reconstruction error of the spectral and spatial satellite footprint. We compare our proposed method to a standard CNN that uses max-pooling to aggregate tokens at the end. Our approach, Alt-Tr, achieves superior performance. The comparison with the vanilla Transformer is challenging due to its quadratic scaling factor in memory consumption, which complicates training without the possibility of using masked attention to reduce requirements, except for the proposed approach. OOM = Out of Memory.

the Ku band (Figures 13 and 14), where satellites are closer compared to the X band.

Figure 16 shows spectral reconstruction and satellite separation. The model achieves promising results for carrier separation and reconstruction, even with overlapping signals. However, reconstruction is limited with overlapping carriers when one signal is much stronger, drowning out the weaker signal. Some artifacts also appear when satellites are close together; signals may be reconstructed with low amplitudes for neighboring satellites.

More detailed quantitative results are presented in Table VIII, which evaluates model performance for the Ku and X bands. The model performs best in the X band, which is simpler because of the larger average distance between satellites, as compared to the Ku band, where satellites are often collocated. These quantitative and qualitative results, consistent with those on synthetic data, show that detection accuracy is strongly influenced by the maximum received power of a satellite signal. The model's detection accuracy decreases for weaker signals, indicating reduced sensitivity under such conditions. These weaker signals can be attributed to cross-polarization effects, which are particularly significant in the Ku band, where vertical polarization (VP) is frequently used but not well-suited to mitigating cross-polarization. Because we do not have access to each satellite's polarization plane, the antenna may end up pointing toward satellites whose polarization planes differ from its own, thereby exacerbating cross-polarization issues.

VII. Conclusion and Future Work

Conclusion. We have introduced the Alternating-Transformer (Alt-Tr), a physics-aligned architecture for joint satellite detection and characterization from passive RF spectro-spatial acquisitions. By factorizing self-attention along the frequency and longitude axes, our model scales efficiently to wideband data while retaining exact global context along each dimension. A CNN tokenizer preserves narrow carriers, while a curriculum on satellite proximity and a duplication-based scaling aug-

	Spatial pattern error (MSE)	Spectral pattern error (MSE)	Satellite detection accuracy
X band	**0.0002**	**0.0011**	**98%**
Ku band	**0.0007**	**0.0052**	**95%**

TABLE VIII: Results for the X-band and Ku-band in terms of Spectral and Spatial pattern reconstruction error. And in terms of satellite detection accuracy for 5 acquisitions each (splitted in 15 test images of 3072x200 each).

mentation stabilize training and improve generalization. The model's ability to reconstruct per-satellite spectral and spatial footprints provides operationally richer output than conventional bounding-box detectors.

Limitations and Operational Envelope. Our single-site, single-polarization approach faces fundamental identifiability limits when satellites are spaced below the Rayleigh criterion and have complete spectral overlap. In these scenarios, signals merge into equivalence classes that cannot be disambiguated without additional information (e.g., multi-site observations). Model performance also degrades for weak carriers with low signal-to-noise ratios. Empirically, reliable operation is achieved for carriers $\gtrsim 5\,\mathrm{dB}$ above the noise floor and for satellites separated by at least one longitude bin ($0.2°$). Other practical challenges include potential sim-to-real gaps from uncalibrated antenna patterns, reliance on weak labels for real data, and the absence of explicit uncertainty quantification. Finally, our framework is currently designed for the quasi-stationary geometry of geostationary orbits; extending it to Non-Geostationary Satellite Orbits (NGSO) presents significant challenges due to rapid satellite motion, Doppler effects, and temporal warping of the spatial signatures.

Future Work. Our primary future research directions include:

1) *Adaptation to NGSO (LEO/MEO) Orbits.* Addressing NGSO tracking requires fundamentally new approaches. The Doppler shift, negligible in the GEO context, becomes a dominant feature. Furthermore, the transient nature of LEO/MEO passes results in distorted and incomplete spatial footprints, demanding models that can characterize satellites from partial observations.
2) *Temporal Association and Tracking.* We plan to extend our model to incorporate multi-frame temporal context. This could be achieved via temporal self-attention to form tracklets from sequential frames or by adding a higher-level expert layer that performs data association to track satellites across multiple revisits.
3) *Multimodal Data Fusion.* Fusing data from heterogeneous sources such as multi-band or multi-polarization RF measurements, data from different ground sites, or optical observations is a promising avenue for resolving ambiguities. Transformer cross-attention is a natural mechanism for integrating such multimodal data to improve performance in challenging, low-SNR, or highly congested scenarios.

In parallel, we will continue efforts on system calibration and principled uncertainty quantification to further enhance result quality and interpretability. Collectively, these developments aim to deliver automated, resilient, and scalable SSA solutions that remain effective beyond GEO and under realistic operational constraints.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014862 made by GENCI.

Frederic Jurie was supported by the ANR under award number ANR-19-CHIA-0017.

REFERENCES

[1] Council of the European Union European Parliament. Regulation (eu) 2021/696 of the european parliament and of the council. *Official Journal of the European Union*, L 170:53–55, 2021.

[2] Niladri Das and Raktim Bhattacharya. Privacy and utility aware data sharing for space situational awareness from ensemble and unscented kalman filtering perspective. *IEEE Transactions on Aerospace and Electronic Systems*, 57(2):1162–1176, 2021.

[3] Sevda Sahin and Tolga Girici. Resource allocation in networked joint radar and communications. *IEEE Transactions on Aerospace and Electronic Systems*, pages 1–11, 2024.

[4] Akram Al-Hourani. In-orbit space situational awareness using doppler frequency shift. *IEEE Transactions on Aerospace and Electronic Systems*, 60(5):7542–7547, 2024.

[5] C.D. Johnson. *Handbook for New Actors in Space*. Secure World Foundation, 2017.

[6] Steredenn Daumont, Yann Picard, and Baptiste Guillot. Device, method and program for recording the radiofrequency activity of artificial satellites, Jan 2024. Safran Data Systems SAS, Active. Anticipated expiration: 2042-02-02.

[7] 3GPP. 3gpp specification series: 38 series. http://www.3gpp.org/DynaReport/38-series.htm, 2024. Accessed: 2024-09-02.

[8] satbeams. Satellite footprints - satbeams. https://www.satbeams.com/footprints, 2024. Accessed: 2024-09-02.

[9] Intelsat. Intelsat fleet maps.

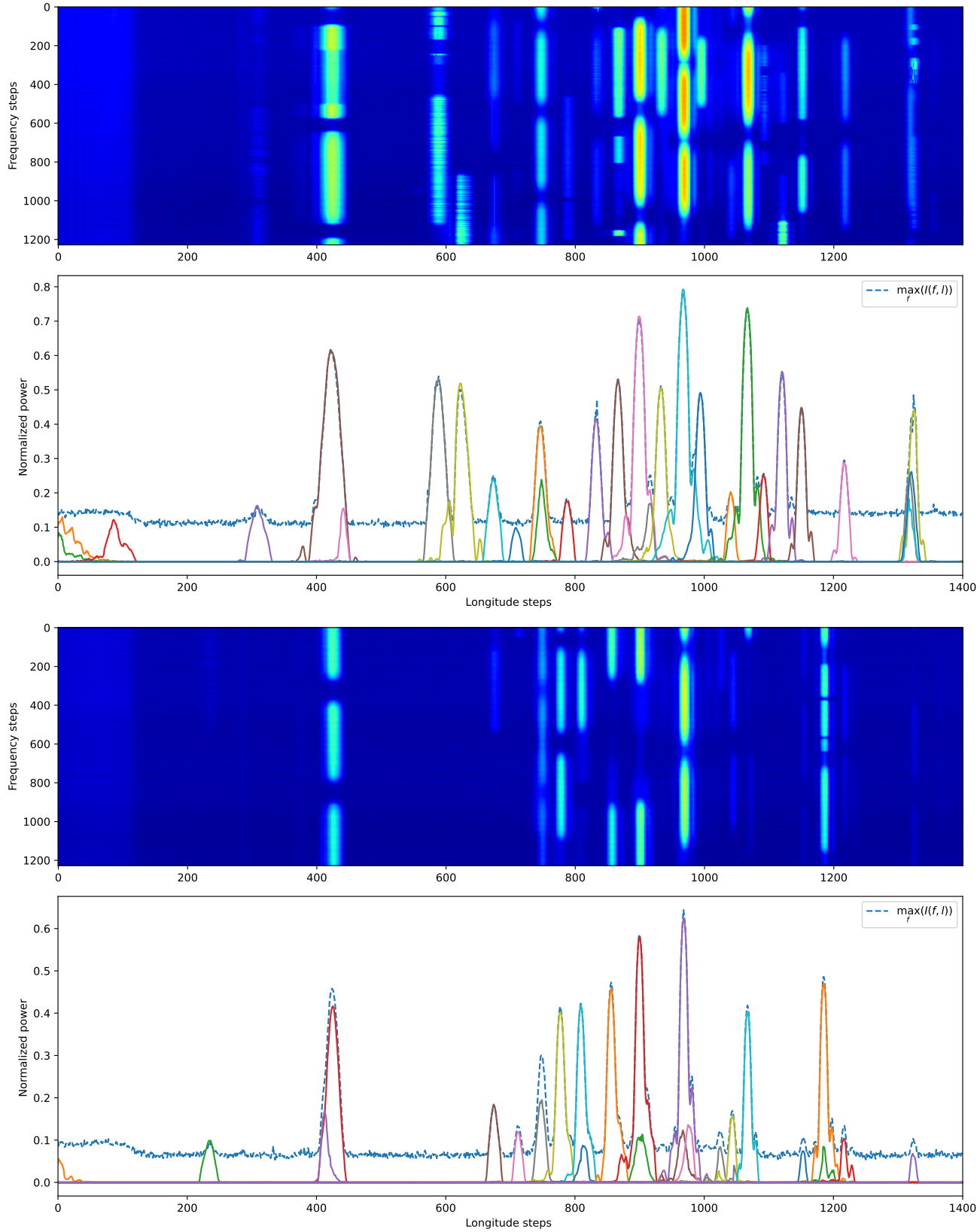

Fig. 13: Spatial footprint reconstruction in the Ku band (Page 1)

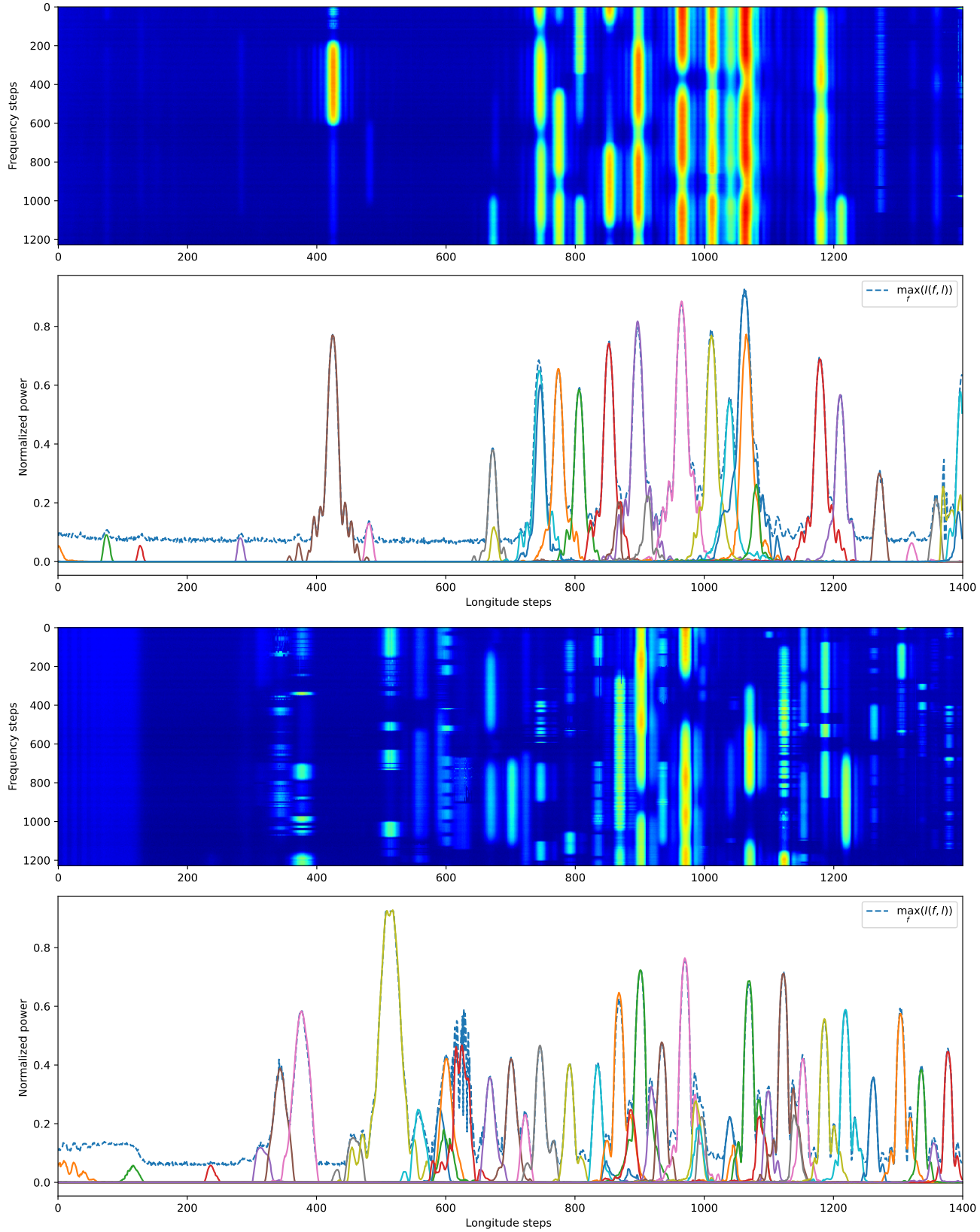

Fig. 14: Spatial footprint reconstruction in the Ku band (Page 2)

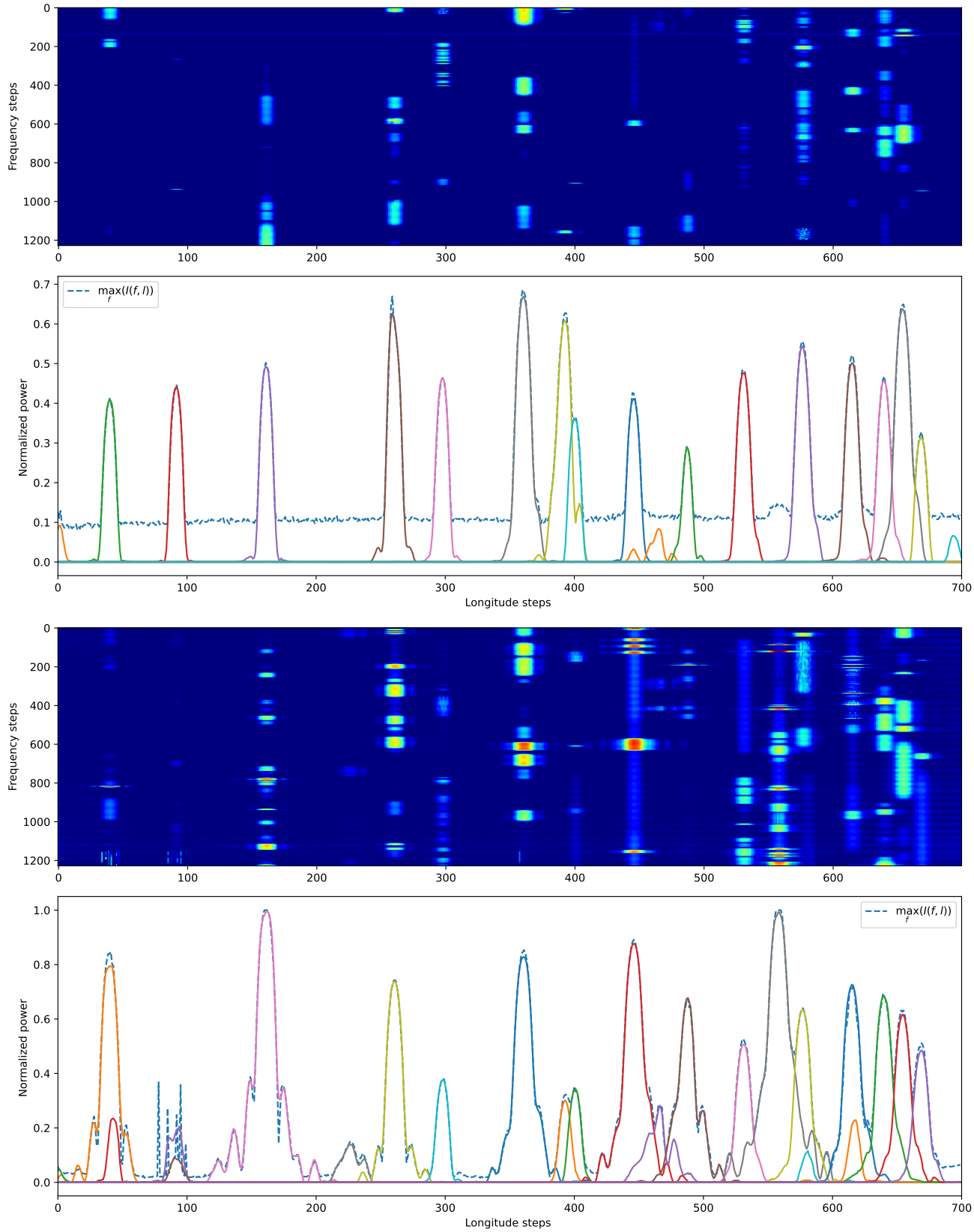

Fig. 15: X band reconstruction of spatial footprint

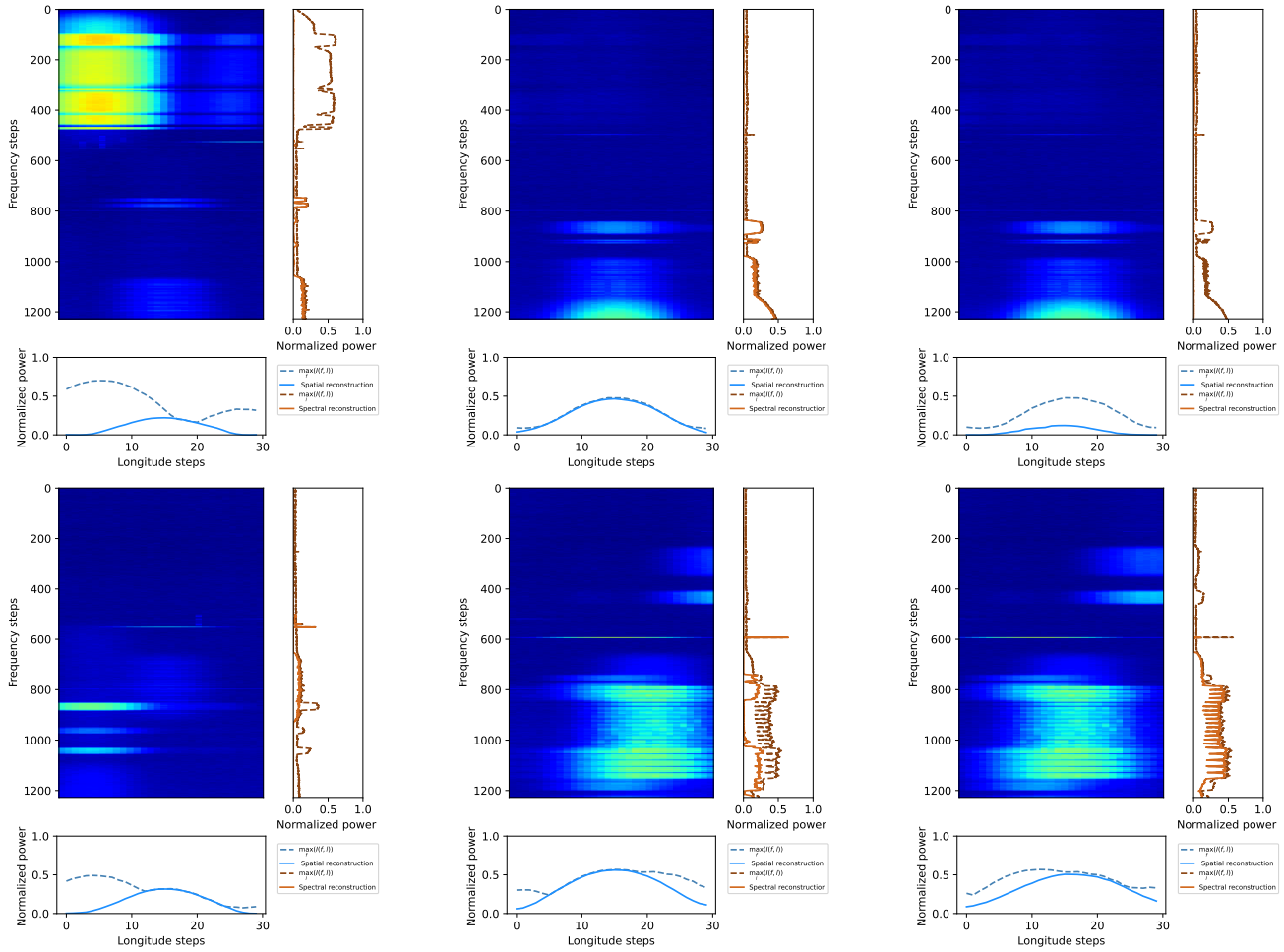

Fig. 16: Examples of reconstruction of the satellites spectral and spatial footprint in various scenario from the Ku-band (VP).

https://www.intelsat.com/fleetmaps/, 2024. Accessed: 2024-09-02.

[10] Zhengnan Xie, Alice Saebom Kwak, Enfa George, Laura W. Dozal, Hoang Van, Moriba Jah, Roberto Furfaro, and Peter Jansen. Extracting space situational awareness events from news text. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6077–6082, Marseille, France, June 2022. European Language Resources Association.

[11] Trier Mortlock and Zaher M. Kassas. Assessing machine learning for leo satellite orbit determination in simultaneous tracking and navigation. In *2021 IEEE Aerospace Conference (50100)*, pages 1–8, 2021.

[12] Dan Shen, Jingyang Lu, Genshe Chen, Erik Blasch, Carolyn Sheaff, Mark Pugh, and Khanh Pham. Methods of machine learning for space object pattern classification. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, pages 565–572, 2019.

[13] Dylan Grigg, Steven J. Tingay, Marcin Sokolowski, and Randall B. Wayth. DUG insight: A software package for big-data analysis and visualisation, and its demonstration for passive radar space situational awareness using radio telescopes. *Astron. Comput.*, 40:100619, 2022.

[14] Mohammad Monirujjaman Khan, Sazzad Hossain, Puezia Mozumdar, Shamima Akter, and Ratil H Ashique. A review on machine learning and deep learning for various antenna design applications. *Heliyon*, 8(4), 2022.

[15] Yongguang Mo, Jianjun Huang, and Gongbin Qian. Deep learning approach to uav detection and classification by using compressively sensed rf signal. *Sensors*, 22(8), 2022.

[16] Jesse Claiborne, Tivon Brown, Paul Rodriguez, and Sambit Bhattacharya. Enhancing rare object detection in ai: Leveraging synthetic data for improved model training. In *SoutheastCon 2024*, pages 56–60, 2024.

[17] Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, and Lin Ma. Instagen: Enhancing object detection by training on synthetic dataset. *CoRR*, abs/2402.05937, 2024.

[18] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. Synthetic data for video surveillance applications of computer vision: A review. *International Journal of Computer Vision*, pages 1–37, 05 2024.

[19] Sergey I. Nikolenko. Synthetic data for deep learning, 2019.

[20] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification.

Machine Learning, 108(8-9):1329–1351, September 2019.

[21] Kaitlyn Johnson.
Rendezvous and proximity operations.
https://www.jstor.org/stable/resrep26047.7.pdf, 2020.
Accessed: 2024-09-02.

[22] Chris Yakopcic, Tarek M Taha, Sanjeevi Sirisha Karri, Guru Subramanyam, Aaron D Smith, and Janette C Briones.
Design and analysis of convolutional neural network for rf signal modulation classification for in-orbit deployment.
In *2021 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)*, pages 1–6. IEEE, 2021.

[23] Jonathan Tran, Prateek Puri, Jordan Logue, Anthony Jacques, Li Ang Zhang, Krista Langeland, George Nacouzi, and Gary J. Briggs.
Artificial intelligence and machine learning for space domain awareness: The development of two artificial intelligence case studies.
Research Report RR-A2318-2, RAND Corporation, September 2024.
Includes research summary and one-page overview.

[24] Michael Lim, Payam Mousavi, Jelena Sirovljevic, and Huiwen You.
Onboard artificial intelligence for space situational awareness with low-power gpus.
In *Advanced Maui Optical and Space Surveillance Technologies Conference*, 2020.

[25] Lauren J. Wong, William H. Clark, Bryse Flowers, R. Michael Buehrer, William C. Headley, and Alan J. Michaels.
An rfml ecosystem: Considerations for the application of deep learning to spectrum situational awareness.
Institute of Electrical and Electronics Engineers (IEEE), 2:2243–2264, 2021.

[26] Ruolin Zhou, Fugang Liu, and Christopher W. Gravelle.
Deep learning for modulation recognition: A survey with a demonstration.
IEEE Access, 8:67366–67376, 2020.

[27] Merima Kulin, Tarik Kazaz, Ingrid Moerman, and Eli De Poorter.
End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications.
IEEE access, 6:18484–18501, 2018.

[28] Yuan Zeng, Meng Zhang, Fei Han, Yi Gong, and Jin Zhang.
Spectrum analysis and convolutional neural network for automatic modulation recognition.
IEEE Wireless Communications Letters, 8(3):929–932, 2019.

[29] Tim O'Shea, Tamohgna Roy, and T. Charles Clancy.
Learning robust general radio signal detection using computer vision methods.
In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 829–832, 2017.

[30] Qihang Peng, Andrew Gilman, Nuno Vasconcelos, Pamela C. Cosman, and Laurence B. Milstein.
Robust deep sensing through transfer learning in cognitive radio.
IEEE Wireless Communications Letters, 9(1):38–41, 2020.

[31] Hai N. Nguyen, Marinos Vomvas, Triet D. Vo-Huu, and Guevara Noubir.
WRIST: wideband, real-time, spectro-temporal RF identification system using deep learning.
IEEE Trans. Mob. Comput., 23(2):1550–1567, 2024.

[32] Jakob Wicht, Ulf Wetzker, and Vineeta Jain.
Spectrogram data set for deep-learning-based rf frame detection.
Data, 7(12), 2022.

[33] Adela Vagollari, Viktoria Schram, Wayan Wicke, Martin Hirschbeck, and Wolfgang Gerstacker.
Joint detection and classification of rf signals using deep learning.
In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pages 1–7, 2021.

[34] Xin Liu, Yuhua Xu, Luliang Jia, Qihui Wu, and Alagan Anpalagan.
Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach.
IEEE Communications Letters, 22(5):998–1001, 2018.

[35] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Zhangyang Wang.
More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity.
In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie.
A convnet for the 2020s.
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022.

[37] Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy.
Convolutional radio modulation recognition networks.
In Chrisina Jayne and Lazaros S. Iliadis, editors, *Engineering Applications of Neural Networks - 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings*, volume 629 of *Communications in Computer and Information Science*, pages 213–226. Springer, 2016.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.
An image is worth 16x16 words: Transformers for image recognition at scale.
In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[40] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans.
Axial attention in multidimensional transformers.
CoRR, abs/1912.12180, 2019.

[41] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen.
Axial-deeplab: Stand-alone axial-attention for panoptic segmentation.
In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 108–126. Springer, 2020.

[42] Nicolae-Catalin Ristea, Radu Tudor Ionescu, and Fahad Shahbaz Khan.
Septr: Separable transformer for audio spectrogram processing.
In Hanseok Ko and John H. L. Hansen, editors, *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4103–4107. ISCA, 2022.

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.
Swin transformer: Hierarchical vision transformer using shifted windows.

In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[44] Albert Gu, Karan Goel, and Christopher Ré.
Efficiently modeling long sequences with structured state spaces.
In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[45] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang.
Vision mamba: Efficient visual representation learning with bidirectional state space model.
In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[46] Shuai Hu, Feng Gao, Xiaowei Zhou, Junyu Dong, and Qian Du.
Hybrid convolutional and attention network for hyperspectral image denoising.
IEEE Geosci. Remote. Sens. Lett., 21:1–5, 2024.

[47] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens.
Attention augmented convolutional networks.
In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3285–3294. IEEE, 2019.

[48] Pengfei Gu, Yejia Zhang, Chaoli Wang, and Danny Z. Chen.
Convformer: Combining CNN and transformer for medical image segmentation.
In *20th IEEE International Symposium on Biomedical Imaging, ISBI 2023, Cartagena, Colombia, April 18-21, 2023*, pages 1–5. IEEE, 2023.

[49] G. Miao, J. Zander, K.W. Sung, and S.B. Slimane.
Fundamentals of Mobile Data Networks.
Fundamentals of Mobile Data Networks. Cambridge University Press, 2016.

[50] Gérard Maral, Michel Bousquet, and Zhili Sun.
Satellite Communications Systems: Systems, Techniques and Technology.
John Wiley & Sons, 6 edition, 2020.

[51] Robin M. Green.
Spherical Astronomy.
Cambridge University Press, 1985.

[52] J.D. Kraus and R.J. Marhefka.
Antennas for All Applications.
McGraw-Hill series in electrical engineering. McGraw-Hill, 2002.

[53] J.C. Slater and N.H. Frank.
Introduction to Theoretical Physics: by John C. Slater and Nathaniel H. Frank.
International series in pure and applied physics. McGraw-Hill, 1933.

[54] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton.
Fake it till you make it: face analysis in the wild using synthetic data alone.
In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3661–3671. IEEE, 2021.

[55] Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar.
Synthetic data, real errors: how (not) to publish and use synthetic data, 2023.

[56] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko.
Visda: A synthetic-to-real benchmark for visual domain adaptation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018.

[57] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund.
Sim-to-real transfer in deep reinforcement learning for robotics: a survey.
In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020.

[58] Antonio Torralba and Alexei A. Efros.
Unbiased look at dataset bias.
In *CVPR 2011*, pages 1521–1528, 2011.

[59] Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino.
Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning.
IEEE Access, 9:153171–153187, 2021.

[60] United States Space Command.
Space-Track: Satellite Database, 2024.
Accessed: 2024-10-31.

[61] John L. Gooban.
Rendezvous and proximity operations of the space shuttle.
Source of Acquisition: NASA Johnson Space Center, 2005.
United Space Alliance, LLC, Houston, Texas, 77058.

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox.
U-net: Convolutional networks for biomedical image segmentation.
In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[64] Jason Ramapuram, Maurits Diephuis, Frantzeska Lavda, Russ Webb, and Alexandros Kalousis.
Variational saccading: Efficient inference for large resolution images.
In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 119. BMVA Press, 2019.

[65] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy.
The impact of positional encoding on length generalization in transformers.
In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[66] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde.
On the computational complexity of self-attention.
In Shipra Agrawal and Francesco Orabona, editors, *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pages 597–619. PMLR, 2023.

[67] Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett.
Fspool: Learning set representations with featurewise sort pooling.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[68] Ben Hayes, Charalampos Saitis, and György Fazekas.
The responsibility problem in neural networks with unordered targets.
In Krystal Maughan, Rosanne Liu, and Thomas F. Burns, editors, *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023.

[69] Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett.
Deep set prediction networks.
In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3207–3217, 2019.

[70] Huilin Qu, Congqiao Li, and Sitian Qian.
Particle transformer for jet tagging.
In *International Conference on Machine Learning*, pages 18281–18292. PMLR, 2022.

[71] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe.
Curriculum learning: A survey.
Int. J. Comput. Vis., 130(6):1526–1565, 2022.

Sidney Besnard Sidney Besnard received his M.S. degree in imagery and data analysis from the University of Caen, France, in 2022. He is currently a Ph.D. student at Safran Data Systems in Colombelles, France, in collaboration with the GREYC, CNRS UMR 6072 of ENSI-CAEN, supervised by Frederic Jurie (GREYC), Jalal Fadili (GREYC) and Sterdenn Daumont (SDS). Since 2022, his research in machine learning has been applied to satellite detection and characterization systems based on passive RF systems that generate large quantities of raw data. His research focuses on integrating the physical knowledge of the system into neural networks, with the goal of creating neural networks that are less demanding in real data for training and improving the generalization step between generated and real data. Additionally, his work can be extended to nonlinear inverse problems, which can be regularized through the physical knowledge of the system.

Jalal Fadili Jalal Fadili is a Full Professor at Ecole National Supérieure d'Ingénieurs de Caen, and Junior member of Institut Universitaire de France since Oct. 2013. He holds several scientific management positions (editorial activities, national excellence research networks). He also held visiting positions at several universities (QUT-Australia, Stanford, Cal- Tech, EPFL-Switzerland, MIT). In the last decade, he has been an invited or plenary speaker at various international events. His research interests include mathematical signal and image processing, mathematical statistics, inverse problems, variational methods and regularization theory, and non-smooth optimization. His areas of application include medical and astronomical imaging. He has published more than 170 papers in the leading journals and conferences of these fields, 7 book chapters and 2 books.

Frederic Jurie Frederic Jurie is a full Professor at the University of Caen Normandy with a background in computer science research. He holds a PhD in Machine Learning and Computer Vision, and has published in these fields. His career includes roles at CNRS and INRIA Rhône-Alpes, as well as directing the GREYC research laboratory (Normandy, France). His work focuses on Computer Vision and Machine Learning, particularly image and video interpretation. He has contributed to visual tracking, bags of visual words, metric learning, shape models, and image descriptors. He has also ventured into Deep Learning, developing new perception algorithms. Since 2019, F. Jurie has held a chair in Artificial Intelligence, now promoting the application of Machine Learning across various scientific disciplines, embracing the concept of "AI for science".

Steredenn Daumont Steredenn Daumont received the Engineering degree in electronics from École Nationale Supérieure des Sciences Appliquées et de Technologie, Lannion, France and the Ph.D. degree in signal processing and telecommunication from the Institute of Electronics and Telecommunications of Rennes, University of Rennes 1, Signal Communications and Embedded Electronics (SCEE) Research Team, Supelec, Rennes, France, in 2006 and 2009, respectively. From 2006 to 2009, she worked on blind sources separation of MIMO signals and on PAPR, in the SCEE team. Since 2010, she has been working with Zodiac Data Systems (now Safran Data Systems). Her research interests are in signal processing for communications, blind sources separation, and characterization of satellite signals.