AIR Journal of Mathematics and Computational Sciences, Vol. 2026, ISSN: 3142-7197
Article Unique ID: AIRMCS2026360
DOI: 10.65737/AIRMCS2026360
© Author(s). This work is licensed under a Creative Commons License.

ICL Characterization of Geo-Foundation Models

Mosab Hawarey
Director, Geospatial Research (*director@geospatial.ch*)

Received: 2026 February 25 | Approved: 2026 February 27 | Published: 2026 February 28

Abstract

Geo-foundation models (GeoFMs) have emerged as powerful tools for remote sensing, yet there exists no theoretical framework characterizing which downstream tasks admit efficient few-shot adaptation. We address this gap by applying the in-context learning (ICL) complexity framework to classify remote sensing tasks as either ICL-Easy or ICL-Hard. We formalize six core remote sensing function classes—spectral classification, semantic segmentation, object localization, dense change detection, sparse change localization, and spatial interpolation—with explicit sufficient statistic structure. For ICL-Easy tasks (classification, segmentation, dense change detection, spatial interpolation), we prove that additive sufficient statistics enable attention-based aggregation with optimal sample complexity $n_{ICL} = \Theta(CB^2/\varepsilon)$, matching empirical risk minimization. For ICL-Hard tasks (object localization, sparse change detection, instance segmentation), we establish that combinatorial sufficient statistics require super-polynomial transformer size or super-constant depth, via reduction to sparse parity and Håstad's circuit lower bounds. Our main result, the Remote Sensing Dichotomy Theorem, proves that every natural remote sensing task falls into exactly one category with no intermediate cases: pixel-wise prediction tasks are ICL-Easy while instance-level localization tasks are ICL-Hard for $J = \omega(\log \log n)$ objects. This dichotomy explains empirical observations that GeoFMs excel at land cover classification but struggle with object detection, and yields testable predictions including threshold behavior at $J^* \approx 2–4$ objects for typical context sizes. We provide deployment guidelines distinguishing when few-shot ICL suffices versus when fine-tuning or hybrid approaches are required.

Keywords: in-context learning; geo-foundation models; sample complexity; circuit complexity; remote sensing; transformer architectures

Citation: Hawarey, M. (2026). ICL Characterization of Geo-Foundation Models. *AIR Journal of Mathematics and Computational Sciences*, Vol. 2026, AIRMCS2026360. https://doi.org/10.65737/AIRMCS2026360

1. Introduction

Geo-foundation models (GeoFMs) have emerged as a transformative paradigm in remote sensing and Earth observation. Building on the success of foundation models in natural language

processing and computer vision, researchers have developed large-scale pretrained models specifically for geospatial data, including Prithvi (Jakubik et al., 2024), DOFA (Xiong et al., 2024), SatMAE (Cong et al., 2022), SkySense (Guo et al., 2024), and TerraMind (Bountos et al., 2025). These models, pretrained on massive archives of satellite imagery, promise to enable efficient few-shot adaptation to diverse downstream tasks through in-context learning (ICL)—the ability to learn new tasks from a small number of labeled examples provided in the input context, without updating model parameters.

Yet a striking empirical puzzle has emerged from GeoFM benchmarks. On classification and semantic segmentation tasks, GeoFMs demonstrate impressive few-shot performance, often approaching fine-tuned accuracy with minimal labeled examples (Lacoste et al., 2023). However, on object detection and instance-level tasks, the same models struggle dramatically—even with abundant context examples, performance remains far below fine-tuned baselines (Camero et al., 2024). The PANGAEA benchmark reveals that GeoFMs consistently underperform on multi-instance tasks while excelling at pixel-wise prediction (Marsocci et al., 2024). This dichotomy is not explained by model size, pretraining data, or architectural details; it appears to be fundamental.

Despite the proliferation of GeoFMs—over 58 models documented between 2021 and 2024 (Li et al., 2024)—there exists no theoretical framework characterizing *which* downstream tasks admit efficient in-context learning. Current benchmarks like GEO-Bench (Lacoste et al., 2023) and PANGAEA (Marsocci et al., 2024) provide empirical evaluations but offer no principled explanation for why certain tasks are learnable in-context while others are not. Practitioners lack guidance on when few-shot ICL will suffice versus when expensive fine-tuning is necessary. This theoretical gap impedes the principled deployment of GeoFMs in operational Earth observation systems.

In this paper, we address this gap by applying the in-context learning complexity framework of Hawarey (2026a) to remote sensing tasks. That framework introduced the concepts of *Sufficient Statistic Complexity* (SSC) and *Attention-Computable SSC* (AC-SSC) to characterize when transformers can efficiently learn function classes in-context. The central insight is that ICL sample complexity depends on whether the minimal sufficient statistic for a function class can be computed by attention mechanisms—specifically, whether it admits an additive decomposition over context examples. Function classes with additive sufficient statistics are *ICL-Easy*, achieving optimal sample complexity matching empirical risk minimization (ERM). Function classes requiring combinatorial statistics are *ICL-Hard*, provably intractable for constant-depth polynomial-size transformers.

Our main result is the *Remote Sensing Dichotomy Theorem*: every natural remote sensing function class falls into exactly one of two categories—ICL-Easy or ICL-Hard—with no intermediate cases. Pixel-wise prediction tasks (classification, semantic segmentation, dense change detection, spatial interpolation) are ICL-Easy because their sufficient statistics decompose additively over training examples. Instance-level tasks (object detection, sparse change localization, instance segmentation) are ICL-Hard because they require combinatorial selection from exponentially many configurations. The hardness threshold is remarkably tight: for context size n, only $J = O(\log \log n)$ objects can be reliably detected in-context. For typical

GeoFM context sizes ($n \approx 10^6$), this threshold is merely $J \approx 2–4$ objects. Figure 1 illustrates this dichotomy schematically.

1.1 Contributions

This paper makes the following contributions:

1. **Formalization of remote sensing function classes.** We define six core function classes—spectral classification, semantic segmentation, object localization, dense change detection, sparse change localization, and spatial interpolation—with explicit sufficient statistic structure (Section 2).
2. **ICL-Easy classification with optimal bounds.** We prove that pixel-wise tasks admit additive sufficient statistics enabling attention-based aggregation, with sample complexity $n_{ICL} = \Theta(CB^2/\varepsilon)$ matching ERM (Section 3).
3. **ICL-Hard classification with circuit lower bounds.** We establish that instance-level tasks have combinatorial sufficient statistics, proving hardness via reduction to sparse parity and application of Håstad's AC^0 lower bounds (Section 4).
4. **The Remote Sensing Dichotomy Theorem.** We prove that every natural RS task is either ICL-Easy (Type A) or ICL-Hard (Type C) with no intermediate complexity classes, providing a complete classification of 11 standard tasks (Section 5).
5. **Testable predictions and deployment guidelines.** We derive six empirically testable predictions explaining observed GeoFM benchmark patterns and provide practical guidelines for when to use few-shot ICL versus fine-tuning (Section 6).

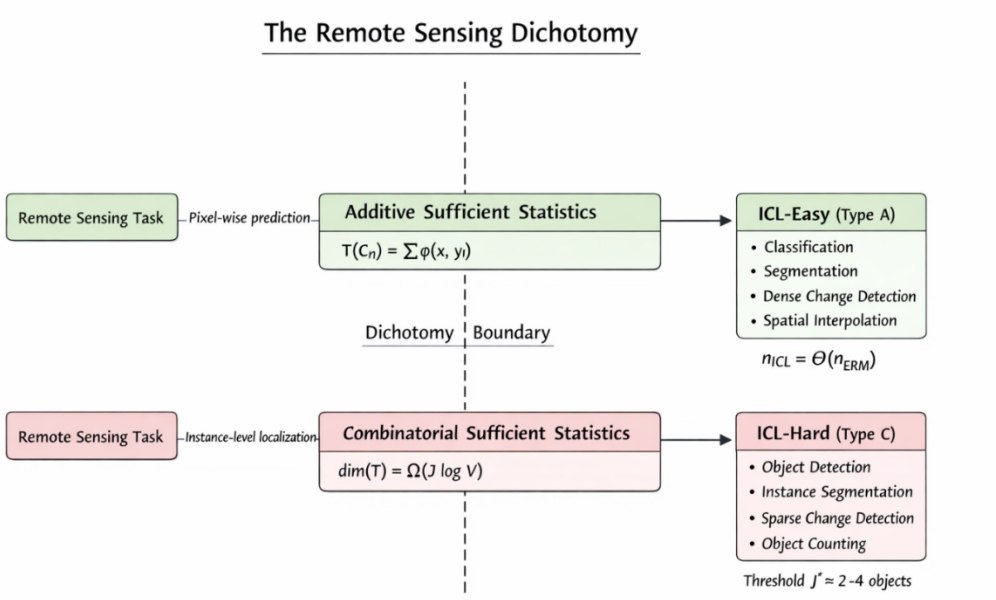


Figure 1. The Remote Sensing Dichotomy. Remote sensing tasks partition into two categories based on sufficient statistic structure. Pixel-wise tasks (top) have additive statistics enabling optimal ICL with n_ICL = Θ(n_ERM). Instance-level tasks (bottom) have combinatorial statistics, making them ICL-Hard beyond threshold $J^* \approx 2–4$ objects.

1.2 Related Work

Geo-Foundation Models. The GeoFM landscape has expanded rapidly since 2021. Vision transformer architectures adapted for remote sensing include SatMAE (Cong et al., 2022), which introduced masked autoencoding for satellite imagery; Prithvi (Jakubik et al., 2024), developed by IBM and NASA with temporal encoding; DOFA (Xiong et al., 2024), featuring dynamic adaptation across sensor types; SkySense (Guo et al., 2024), scaling to 2.6 billion parameters; Clay Foundation Model (Radford et al., 2024), trained on diverse global imagery; and SAMRS (Wang et al., 2024), adapting the Segment Anything Model to remote sensing. Recent surveys document over 58 models with diverse architectures and pretraining strategies (Li et al., 2024; Mai et al., 2025). Benchmarks including GEO-Bench (Lacoste et al., 2023), PANGAEA (Marsocci et al., 2024), and SustainFM (Bountos et al., 2025) have standardized evaluation, consistently revealing strong classification performance but weaker detection capabilities.

In-Context Learning Theory. Theoretical understanding of ICL has advanced significantly. Garg et al. (2022) demonstrated that transformers can learn linear functions in-context. Bai et al. (2023) connected ICL to implicit gradient descent. Hawarey (2026a) introduced the SSC/AC-SSC framework establishing necessary and sufficient conditions for efficient ICL, proving the ICL-Easy/Hard dichotomy based on sufficient statistic structure. Hawarey (2026b) applied this framework to quantum-enhanced geopotential estimation, demonstrating 10^{12} sample complexity reduction at the Heisenberg limit. Our work extends this theoretical framework to the full spectrum of remote sensing tasks.

Circuit Complexity and Transformers. The connection between transformers and circuit complexity classes underlies our hardness results. Håstad (1987) proved that parity requires exponential-size AC^0 circuits. Merrill and Sabharwal (2023) established that constant-depth transformers with bounded precision compute functions in AC^0. Our reduction from sparse parity to object localization leverages these classical results to prove fundamental limitations of transformer-based ICL for combinatorial tasks.

1.3 Paper Organization

The remainder of this paper is organized as follows. Section 2 establishes the theoretical preliminaries including the ICL complexity framework and remote sensing notation. Section 3 formalizes remote sensing function classes with explicit sufficient statistic structure. Section 4 establishes ICL-Easy results for pixel-wise tasks with optimal sample complexity bounds. Section 5 proves ICL-Hard results for instance-level tasks via circuit complexity lower bounds. Section 6 presents the Remote Sensing Dichotomy Theorem with complete task classification. Section 7 discusses broader implications and limitations. Section 8 concludes with future directions.

2. Preliminaries

This section establishes the theoretical foundations for our analysis. We first review the in-context learning complexity framework of Hawarey (2026a), then introduce the remote sensing setting and notation used throughout the paper.

2.1 In-Context Learning Framework

In-context learning refers to the ability of a pretrained model to perform new tasks by conditioning on a small number of labeled examples provided in the input context, without any parameter updates. Formally, given a function class $\mathcal{F}$ and a context set $\mathcal{C}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ of input-output pairs drawn from some $f^* \in \mathcal{F}$, the ICL task is to predict $\hat{y}_q$ for a query x_q.

Definition 2.1 (ICL Sample Complexity). The *ICL sample complexity* of a function class $\mathcal{F}$ is the minimum number of context examples n required for a transformer to achieve expected error at most ε with probability at least $1 - \delta$:
$n_{ICL}(\mathcal{F}, \varepsilon, \delta) = \min\{n : \exists \text{ transformer } \mathcal{T} \text{ with } \Pr[L(\mathcal{T}(\mathcal{C}_n, x_q), f^*(x_q)] \leq \varepsilon] \geq 1 - \delta\}$
where $L(\cdot,\cdot)$ is an appropriate loss function (e.g., 0-1 loss for classification, squared error for regression).

Definition 2.2 (Sufficient Statistic). A statistic $T: (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{T}$ is *sufficient* for function class $\mathcal{F}$ if the optimal predictor for any query x_q depends on the context $\mathcal{C}_n$ only through $T(\mathcal{C}_n)$: $f^*(x_q) = g(T(\mathcal{C}_n), x_q)$ for some function g

Definition 2.3 (Sufficient Statistic Complexity). The *Sufficient Statistic Complexity* $SSC(\mathcal{F}, n, \varepsilon)$ is the minimum dimension of any sufficient statistic achieving error ε on function class $\mathcal{F}$ with n context examples.

Definition 2.4 (Attention-Computable SSC). The *Attention-Computable Sufficient Statistic Complexity* $AC\text{-}SSC(\mathcal{F}, n, \varepsilon, s)$ is the minimum dimension of any sufficient statistic that can be computed by a transformer of size s. A sufficient statistic T is *attention-computable* if it admits an additive decomposition: $T(\mathcal{C}_n) = \Sigma_{i=1}^n \varphi(x_i, y_i)$ for some feature map $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$. This additive structure enables computation via a single attention layer with uniform weights summing the transformed context examples.

Theorem 2.1 (Master Theorem, Hawarey 2026a). A function class $\mathcal{F}$ is efficiently ICL-learnable if and only if $AC\text{-}SSC(\mathcal{F}, n, \varepsilon, s)$ is polynomially bounded for polynomial transformer size $s = \text{poly}(n)$. Equivalently:
$n_{ICL}(\mathcal{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathcal{F}, \varepsilon, \delta)) \iff AC\text{-}SSC(\mathcal{F}) = \text{poly}(SSC(\mathcal{F}))$

Definition 2.5 (ICL-Easy). A function class $\mathcal{F}$ is *ICL-Easy* if it admits an additive sufficient statistic of polynomial dimension, achieving $n_{ICL} = \Theta(n_{ERM})$. Equivalently, AC-SSC = SSC.

Definition 2.6 (ICL-Hard). A function class $\mathcal{F}$ is *ICL-Hard* if any transformer achieving non-trivial success probability requires either super-polynomial size $s = n^{\omega(1)}$ or super-constant depth $L = \omega(1)$. Equivalently, $AC\text{-}SSC = n^{\omega(1)} \cdot SSC$.

The ICL-Hard classification connects to circuit complexity. Constant-depth polynomial-size transformers with bounded precision $p = O(\log n)$ compute functions in the circuit class AC^0 (Merrill & Sabharwal, 2023). By Håstad's theorem (1987), computing parity of k bits requires AC^0 circuits of size at least $2^{(k^{\Omega(1/L)})}$ for depth L. This lower bound underlies our hardness results.

2.2 Remote Sensing Setting

We now introduce the remote sensing notation and assumptions used throughout the paper.

Definition 2.7 (Spectral Feature Space). Let $B \in \mathbb{N}$ denote the number of spectral bands. The *spectral feature space* is $\mathcal{X} = \mathbb{R}^B$, where each $x \in \mathcal{X}$ represents the spectral signature at a single spatial location. Standard sensors have B = 13 (Sentinel-2), B = 11 (Landsat-8), or $B \in [100, 400]$ (hyperspectral).

Definition 2.8 (Spatial Domain). The *spatial domain* is a discrete grid $\Omega = \{1, ..., H\} \times \{1, ..., W\} \subset \mathbb{Z}^2$, where H and W denote image height and width in pixels. A *remote sensing image* is a function $I: \Omega \to \mathcal{X}$, equivalently represented as a tensor $I \in \mathbb{R}^{H \times W \times B}$.

Definition 2.9 (Observation Model). A *noisy spectral observation* at spatial location $\omega \in \Omega$ is: $y_\omega = f^*(I(\omega)) + \varepsilon_\omega$, where $f^*: \mathcal{X} \to \mathcal{Y}$ is the ground-truth labeling function, $\mathcal{Y}$ is the label space (task-dependent), and $\varepsilon_\omega \sim \mathcal{N}(0, \sigma^2)$ is observation noise. Observations are conditionally independent given f^*.

Definition 2.10 (Context Set). A *context set* of size n is $\mathcal{C}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ are spectral features and $y_i \in \mathcal{Y}$ are corresponding labels. The ICL task is: given $\mathcal{C}_n$ and query x_q, predict $\hat{y}_q$.

Assumption A1 (Bounded Features). There exists $B_x > 0$ such that $\|x\|_2 \leq B_x$ for all $x \in \mathcal{X}$.

Assumption A2 (Bounded Labels). For regression tasks, $|y| \leq B_y$. For classification with C classes, $y \in [C] = \{1, ..., C\}$.

2.3 Geo-Foundation Models

Geo-foundation models are large-scale vision transformers pretrained on satellite imagery archives. The standard architecture processes images as sequences of patches, applies self-attention across spatial and potentially temporal dimensions, and produces embeddings that can be adapted to downstream tasks. For our analysis, the key architectural constraints are:

- **Depth L:** Number of transformer layers, typically $L \in [12, 24]$ for standard GeoFMs, up to L = 32 for large variants.
- **Width w:** Hidden dimension, typically $w \in [768, 1024, 2048]$, yielding polynomial size $s = O(Lw^2)$.
- **Precision p:** Numerical precision, typically p = 16 or 32 bits, satisfying $p = O(\log n)$ for practical context sizes.
- **Context length n:** Maximum number of tokens (patches + labels) processable in a single forward pass, typically $n \in [10^3, 10^6]$.

Under these constraints, GeoFMs compute functions in the circuit class TC^0 (for constant L) or NC^1 (for $L = O(\log n)$). Our hardness results apply to any model within these computational

bounds, independent of specific architectural choices like attention patterns or positional encodings.

2.4 Notation Summary

We collect the key notation used throughout the paper. Scalars are denoted by lowercase italic letters (n, k, σ), vectors by bold lowercase (x, c, μ), matrices by bold uppercase (G, Σ, W), and sets/function classes by calligraphic letters ($\mathscr{F}$, $\mathcal{C}$, $\mathcal{X}$). The notation [C] = {1, ..., C} denotes the set of class labels. Asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$ has standard meaning; $\omega(\cdot)$ and $o(\cdot)$ denote strict asymptotic bounds. We write poly(n) for polynomial in n and $n^{\omega(1)}$ for super-polynomial growth.

Key symbols include: B (spectral bands), C (number of classes), J (number of objects), V (candidate locations), n_{ICL} (ICL sample complexity), n_{ERM} (ERM sample complexity), SSC (sufficient statistic complexity), AC-SSC (attention-computable SSC), $T(\mathcal{C}_n)$ (sufficient statistic), and φ (feature map for additive statistics).

3. Remote Sensing Function Classes

This section formalizes the core remote sensing tasks as mathematical function classes with explicit sufficient statistic structure. Definitions from Section 2 are from Hawarey (2026a) while this section onward presents new formalizations. The key insight is that the structure of sufficient statistics—additive versus combinatorial—determines ICL complexity. We define six function classes spanning the major categories of Earth observation tasks.

3.1 Spectral Classification

Spectral classification assigns a discrete class label to each pixel based on its spectral signature. This is the foundational task in remote sensing, underlying applications from land cover mapping to crop type identification.

Definition 3.1 (Spectral Classification Function Class). The *spectral classification function class* for C classes is:
$\mathscr{F}_{class}^{(C)} = \{f: \mathcal{X} \to [C] \mid f(x) = \text{argmax}_{c\in[C]}\, w_c^\top x + b_c, \|w_c\|_2 \le B_w\}$
where $w_c \in \mathbb{R}^B$ and $b_c \in \mathbb{R}$ are class-specific parameters. The function class has intrinsic dimension $K_{class} = C(B + 1)$.

Definition 3.2 (Probabilistic Classification). For probabilistic prediction, the *softmax classifier* defines:
$p(y = c \mid x; W) = \exp(w_c^\top x + b_c) / \Sigma_{c'=1}^{C} \exp(w_{c'}^\top x + b_{c'})$
This model belongs to the exponential family with natural parameters $\eta = (w_1, ..., w_C, b_1, ..., b_C)$. The exponential family structure is key to establishing additive sufficient statistics.

Definition 3.3 (Scene Classification Function Class). Scene classification assigns a single label to an entire image rather than per-pixel labels:
$\mathscr{F}$_scene^(C) = {f: $\mathbb{R}$^(H×W×B) → [C] | f(I) = argmax_{c∈[C]} w_c^⊤ pool(I) + b_c}

where pool: $\mathbb{R}^{(H\times W\times B)} \to \mathbb{R}^{B}$ is a spatial pooling operator (e.g., global average pooling). Applications include land use classification, biome identification, and urban/rural scene categorization.

Proposition 3.1 (Scene Classification Inherits Spectral Classification Structure). Scene classification reduces to spectral classification on pooled features. The sufficient statistic (defined formally in Section 4, Definition 4.1) applied to pooled representations has dimension $O(CB^2)$.

Proof. The pooling operator $\text{pool}(I) = (1/|\Omega|) \Sigma_{\omega\in\Omega} I(\omega)$ produces a single B-dimensional feature vector per image. Classification on this pooled representation follows Definition 3.1 exactly. ■

3.2 Semantic Segmentation

Semantic segmentation extends classification to produce a dense label map, assigning a class to every pixel in the image. Applications include land use mapping, crop delineation, and water body extraction.

Definition 3.4 (Semantic Segmentation Function Class). The *semantic segmentation function class* is:
$\mathscr{F}_{\text{seg}}^{(C)} = \{f: \mathbb{R}^{H\times W\times B} \to [C]^{H\times W} \mid f(I)_\omega = g(I(\mathcal{N}_\omega)), g \in \mathcal{G}\}$
where $\mathcal{N}_\omega \subset \Omega$ is a local neighborhood (receptive field) of pixel ω, and $\mathcal{G}$ is a class of local classifiers.

Proposition 3.2 (Segmentation Decomposition). Under the assumption of pixel-wise independence given local context, semantic segmentation decomposes into $|\Omega| = HW$ independent classification problems: $f(I) = \{g(I(\mathcal{N}_\omega))\}_{\omega\in\Omega}$

Proof. The prediction at each pixel ω depends only on the local patch $I(\mathcal{N}_\omega)$. With no inter-pixel dependencies in the label model, the joint prediction factorizes:
$p(y \mid I) = \Pi_{\omega\in\Omega}\, p(y_\omega \mid I(\mathcal{N}_\omega))$
Thus segmentation reduces to HW parallel classification tasks. ■

This decomposition is crucial: the sufficient statistic for segmentation is identical to that for classification, applied independently per pixel with shared parameters. Modern segmentation architectures (U-Net, SegFormer) maintain this structure.

Definition 3.5 (Pixel-wise Regression Function Class). Pixel-wise regression predicts continuous values at each spatial location:
$\mathscr{F}_{\text{reg}} = \{f: \mathbb{R}^{(H\times W\times B)} \to \mathbb{R}^{(H\times W)} \mid f(I)_\omega = w^\top I(\mathcal{N}_\omega) + b, \|w\|_2 \le B_w\}$
where $w \in \mathbb{R}^B$ and $b \in \mathbb{R}$ are regression parameters. Applications include crop yield estimation, above-ground biomass prediction, soil moisture mapping, and surface temperature retrieval.

Proposition 3.3 (Regression Decomposition). Under the assumption of pixel-wise independence given local context, pixel-wise regression decomposes into $|\Omega| = HW$ independent regression problems with shared parameters.

Proof. Identical to Proposition 3.2, replacing discrete class labels with continuous target values. The joint prediction factorizes as $p(y \mid I) = \Pi_{\omega \in \Omega} p(y_\omega \mid I(\mathcal{N}_\omega))$. ■

3.3 Object Localization

Object localization identifies the positions and classes of discrete objects within an image. Unlike pixel-wise tasks, localization requires selecting which spatial locations contain objects—a fundamentally combinatorial problem.

Definition 3.6 (Bounding Box Space). A *bounding box* is $b = (x_{min}, y_{min}, x_{max}, y_{max}) \in \mathcal{B}$ where $\mathcal{B} = [0,W] \times [0,H] \times [0,W] \times [0,H]$ with constraints $x_{min} < x_{max}$, $y_{min} < y_{max}$.

Definition 3.7 (Object Localization Function Class). The *J-object localization function class* is:
$\mathcal{F}_{loc}^{(J,C)} = \{f: \mathbb{R}^{H \times W \times B} \to (\mathcal{B} \times [C])^J \mid f(I) = \{(b_j, c_j)\}_{j=1}^J\}$
where each output specifies J object locations and their class labels.

Definition 3.8 (Discretized Localization). For analysis, we discretize to a grid of V candidate locations (anchor boxes) $\mathcal{V} = \{v_1, ..., v_V\} \subset \mathcal{B}$. The localization task becomes selecting which J of V candidates contain objects:
$f(I) = (S, \{c_j\}_{j \in S})$ where $S \subset [V]$, $|S| = J$

Proposition 3.4 (Combinatorial Structure). The output space of J-object localization has cardinality: $|\text{Output Space}| = C(V,J) \cdot C^J$

Proof. Choosing J locations from V candidates: C(V,J) ways. Assigning one of C classes to each: C^J ways. Total: $C(V,J) \cdot C^J$. ■

This combinatorial explosion is the signature of ICL-Hardness. For typical values ($V = 10^4$ anchor boxes, $J = 10$ objects), $C(V,J) > 10^{30}$—no additive statistic of fixed dimension can distinguish all configurations.

Definition 3.9 (Object Counting Function Class). Object counting returns the number of discrete objects in an image:
$\mathcal{F}_count = \{f: \mathbb{R}^(H \times W \times B) \to \mathbb{N} \mid f(I) = |S| \text{ where } S \subset [V] \text{ is the set of object locations}\}$

Applications include vehicle counting, building enumeration, and population density estimation from overhead imagery.

Remark 3.1 (Counting Requires Localization). Although counting outputs a single integer rather than object locations, accurate counting requires implicit localization to avoid double-counting overlapping objects or missing occluded instances. Formally, computing $|S|$ with high

probability requires distinguishing which subset $S \subset [V]$ is occupied—the same combinatorial selection problem as localization. Thus object counting inherits the ICL-Hard status of localization.

Definition 3.10 (Oriented Object Detection Function Class). Oriented object detection outputs rotated bounding boxes, critical for objects with significant aspect ratios such as ships, aircraft, and vehicles:
$\mathscr{F}_obb^(J,C) = \{f: \mathbb{R}^(H\times W\times B) \to (\mathscr{B}_rot \times [C])^J \mid f(I) = \{(b_j, \theta_j, c_j)\}_\{j=1\}^J\}$
where $\mathscr{B}_rot = \mathscr{B} \times [0, 2\pi)$ extends the bounding box space with rotation angle θ.

Proposition 3.5 (OBB Reduces to Localization). Oriented object detection reduces to standard localization with an additional continuous parameter per object. The combinatorial structure—selecting J locations from V candidates—remains unchanged, hence $\mathscr{F}_obb$ inherits the ICL-Hard status of $\mathscr{F}_loc$.

Proof. The rotation angle $\theta_j \in [0, 2\pi)$ adds a continuous parameter estimable from local context once the object location is identified. The hardness stems from the discrete selection problem $C(V,J)$, which is identical for axis-aligned and oriented boxes. ■

3.4 Change Detection

Change detection identifies differences between images of the same scene at different times. Critically, change detection admits two distinct formulations with fundamentally different complexity.

Definition 3.11 (Bi-Temporal Image Pair). A *bi-temporal image pair* is (I_{t_1}, I_{t_2}) where $I_{t_1}, I_{t_2} \in \mathbb{R}^{H\times W\times B}$ are images of the same scene at times $t_1 < t_2$.

Definition 3.12 (Dense Change Detection). The *dense change detection function class* is:
$\mathscr{F}_{\text{change-dense}} = \{f: \mathbb{R}^{H\times W\times B} \times \mathbb{R}^{H\times W\times B} \to \{0,1\}^{H\times W}\}$
where $f(I_{t_1}, I_{t_2})_\omega = 1$ indicates change at pixel ω.

Proposition 3.6 (Dense Change as Binary Segmentation). Dense change detection is equivalent to binary semantic segmentation on the concatenated input $[I_{t_1}; I_{t_2}] \in \mathbb{R}^{H\times W\times 2B}$.

Proof. The change indicator at each pixel depends on the local bi-temporal context. This factorizes identically to semantic segmentation (Proposition 3.2) with input dimension 2B instead of B. ■

Definition 3.13 (Sparse Change Localization). The *sparse change localization function class* identifies J discrete change events:
$\mathscr{F}_{\text{change-sparse}}^{(J)} = \{f: \mathbb{R}^{H\times W\times B} \times \mathbb{R}^{H\times W\times B} \to (\mathscr{B} \times \mathcal{T})^J\}$
where $\mathcal{T}$ is a set of change types (construction, demolition, vegetation loss, etc.).

Remark 3.2. The critical distinction: *dense change detection* (pixel-wise) reduces to parallel binary classification and is an ICL-Easy candidate; *sparse change localization* (instance-level)

requires combinatorial search and is an ICL-Hard candidate. This explains why GeoFMs perform well on change segmentation benchmarks but struggle with change event detection.

3.5 Spatial Interpolation

Spatial interpolation predicts continuous values at unobserved locations based on nearby measurements. This encompasses kriging, spatial regression, and related geostatistical methods.

Definition 3.14 (Spatial Interpolation Function Class). The *spatial interpolation function class* is:
$\mathscr{F}_{\text{interp}} = \{f: (\mathcal{X} \times \mathbb{R})^n \times \Omega \to \mathbb{R} \mid f(\mathcal{C}_n, \omega_q) = \mu^\top K^{-1} k(\omega_q)\}$
where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix with $K_{ij} = \kappa(\omega_i, \omega_j)$, $k(\omega_q) \in \mathbb{R}^n$ with $k_i(\omega_q) = \kappa(\omega_i, \omega_q)$, and $\mu = (y_1, ..., y_n)^\top$. This is precisely the kriging predictor from geostatistics.

Definition 3.15 (Spatial Covariance Sufficient Statistic). For a finite-dimensional kernel embedding $k(\omega) \in \mathbb{R}^K$, define the *Gram matrix* and *observation vector*:
$G_n = \Sigma_{i=1}^n k(\omega_i)k(\omega_i)^\top, \quad c_n = \Sigma_{i=1}^n y_i k(\omega_i)$

Proposition 3.7 (Additive Structure of Spatial Statistics). The sufficient statistic $T(\mathcal{C}_n) = (G_n, c_n)$ decomposes additively:
$T(\mathcal{C}_n) = \Sigma_{i=1}^n \varphi(\omega_i, y_i)$
where $\varphi(\omega, y) = [\text{vec}(k(\omega)k(\omega)^\top); y \cdot k(\omega)]$. This additive structure is the signature of ICL-Easy classes, identical to geopotential estimation (Hawarey, 2026b).

3.6 Summary of Function Classes

Table 1 summarizes the six function classes with their output spaces, sufficient statistic structures, and conjectured ICL status. The key structural distinction is between pixel-wise tasks (additive statistics, ICL-Easy candidates) and instance-level tasks (combinatorial statistics, ICL-Hard candidates). Sections 4 and 5 prove these conjectures rigorously.

Table 1: Summary of Remote Sensing Function Classes

Function Class	Notation	Output Space	Statistic Type	ICL Status
Spectral Classification	$\mathscr{F}$_class	[C]	Additive	**Easy**
Semantic Segmentation	$\mathscr{F}$_seg	[C]^(H×W)	Additive	**Easy**
Object Localization	$\mathscr{F}$_loc	($\mathscr{B}$×[C])^J	Combinatorial	**Hard**
Dense Change Detection	$\mathscr{F}$_change-dense	{0,1}^(H×W)	Additive	**Easy**
Sparse Change Localization	$\mathscr{F}$_change-sparse	($\mathscr{B}$×$\mathcal{T}$)^J	Combinatorial	**Hard**
Spatial Interpolation	$\mathscr{F}$_interp	$\mathbb{R}$	Additive	**Easy**

The structural dichotomy is clear: pixel-wise tasks (rows 1, 2, 4, 6) have additive statistics enabling efficient ICL, while instance-level tasks (rows 3, 5) have combinatorial statistics that create fundamental computational barriers.

4. ICL-Easy Classification

This section establishes that pixel-wise remote sensing tasks are ICL-Easy. We prove that spectral classification, semantic segmentation, dense change detection, and spatial interpolation all admit additive sufficient statistics, enabling attention-based computation with optimal sample complexity matching ERM.

4.1 Sufficient Statistics for Spectral Classification

We begin by identifying the minimal sufficient statistic for spectral classification and proving its additive structure.

Proposition 4.1 (Exponential Family Structure). Under the probabilistic classification model (Definition 3.2), the conditional distribution of observations given class parameters belongs to the exponential family:
$p(\{(x_i, y_i)\}_{i=1}^n \mid W) = h(x, y) \exp(\eta(W)^\top T(\mathcal{C}_n) - A(\eta))$

Proof. The joint likelihood for n i.i.d. observations is $p(\mathcal{C}_n \mid W) = \Pi_{i=1}^n p(y_i \mid x_i; W) \cdot p(x_i)$. For the softmax model: $\log p(y_i \mid x_i; W) = w_{y_i}^\top x_i + b_{y_i} - \log \Sigma_{c=1}^C \exp(w_c^\top x_i + b_c)$
Summing over i yields exponential family form with sufficient statistic depending on class-conditional sums. ■

Definition 4.1 (Classification Sufficient Statistic). The *minimal sufficient statistic* for spectral classification is:
$T_{class}(\mathcal{C}_n) = (\{n_c\}_{c=1}^C, \{\mu_c\}_{c=1}^C, \{\Sigma_c\}_{c=1}^C)$
where $n_c = \Sigma_{i=1}^n \mathbb{1}[y_i = c]$ is the class count, $\mu_c = (1/n_c)\Sigma_{i:y_i=c} x_i$ is the class mean, and $\Sigma_c = (1/n_c)\Sigma_{i:y_i=c} (x_i - \mu_c)(x_i - \mu_c)^\top$ is the class covariance.

Theorem 4.1 (Dimension of Classification Sufficient Statistic). The sufficient statistic T_{class} has dimension: $\dim(T_{class}) = C + CB + C \cdot B(B+1)/2 = O(CB^2)$.

Proof. Class counts: C scalars. Class means: C vectors of dimension B → CB scalars. Class covariances: C symmetric B × B matrices → $C \cdot B(B+1)/2$ scalars. Total: $C(1 + B + B(B+1)/2) = O(CB^2)$. ■

Example 4.1. For Sentinel-2 (B = 13) with C = 10 land cover classes: $\dim(T_{class}) = 10 \cdot (1 + 13 + 91) = 1050$.

4.2 Attention-Computability of Classification Statistics

We now prove that the classification sufficient statistic can be computed by a single attention layer, establishing ICL-Easy status.

Lemma 4.1 (Additive Decomposition). The classification sufficient statistic decomposes additively:
$T_{class}(\mathcal{C}_n) = \Sigma_{i=1}^n \varphi_{class}(x_i, y_i)$

where the feature map $\varphi_{class}: \mathcal{X} \times [C] \to \mathbb{R}^{C(1+B+B^2)}$ is defined by $\varphi_{class}(x, y) = (\{\mathbb{1}[y = c]\}_{c=1}^{C}, \{\mathbb{1}[y = c] \cdot x\}_{c=1}^{C}, \{\mathbb{1}[y = c] \cdot \text{vec}(xx^\top)\}_{c=1}^{C})$.

Proof. We verify each component. Class counts: $\Sigma_{i=1}^{n} \mathbb{1}[y_i = c] = n_c$. Class sums (unnormalized means): $\Sigma_{i=1}^{n} \mathbb{1}[y_i = c] \cdot x_i = n_c\mu_c$. Class second moments: $\Sigma_{i:y_i=c} x_i x_i^\top$. The mean and covariance are recoverable: $\mu_c = (1/n_c)\Sigma x_i$ and $\Sigma_c = (1/n_c)\Sigma x_i x_i^\top - \mu_c\mu_c^\top$. ■

Lemma 4.2 (FFN-Computability of Feature Map). The feature map $\varphi_{class}(x, y)$ is computable by a feedforward network with depth $L_\varphi = O(1)$, width $w_\varphi = O(CB^2)$, and size $s_\varphi = O(CB^2)$.

Proof. The computation requires: (1) one-hot encoding of y: $e_y \in \{0,1\}^C$—implementable by C threshold units; (2) outer product $xx^\top$: B^2 multiplications—implementable by one layer of width B^2; (3) masking by class indicator: element-wise multiplication—linear operations. All operations are polynomial combinations implementable by ReLU networks with constant depth. ■

Theorem 4.2 (Attention-Computability of Classification). The classification sufficient statistic $T_{class}(\mathcal{C}_n)$ is *attention-computable* in the sense of Hawarey (2026a, Definition 2.16). A single attention layer computes $T = \Sigma_{i=1}^{n} \varphi_{class}(x_i, y_i)$ via uniform attention weights.

Proof. We construct the attention mechanism explicitly following Hawarey (2026a, Theorem 1).
Embedding: Encode each context example as $e_i = [x_i; \text{onehot}(y_i); \varphi_{class}(x_i, y_i); 0; i/n] \in \mathbb{R}^{B+C+\dim(\varphi)+2}$ and the query as $e_q = [x_q; 0; 0; 1; 1]$.
Attention weights: Design W_Q, W_K to extract position indicators, yielding uniform attention: $\alpha_{q,i} = \exp(e_q^\top W_Q^\top W_K e_i) / \Sigma_j \exp(\cdot) = 1/n$.
Value projection: Set W_V to extract the φ portion: $o_q = \Sigma_{i=1}^{n} \alpha_{q,i} \cdot W_V e_i = (1/n) \Sigma_{i=1}^{n} \varphi_{class}(x_i, y_i) \propto T_{class}(\mathcal{C}_n)$.
The factor of 1/n is absorbed into subsequent processing. ■

4.3 Sample Complexity Bounds

Having established attention-computability, we derive tight sample complexity bounds showing that ICL matches ERM.

Theorem 4.3 (ICL Sample Complexity for Classification). For the spectral classification function class $\mathscr{F}_{_class}^{(C)}$ with B spectral bands, the ICL sample complexity satisfies:
$n_{ICL}(\mathscr{F}_{_class}^{(C)}, \varepsilon, \delta) = \Theta((CB^2/\varepsilon) \cdot \log(C/\delta))$

Proof.
Upper bound: The optimal classifier (Bayes optimal for Gaussian class-conditionals, or maximum likelihood for the exponential family) depends on $(\mu_c, \Sigma_c)_{c=1}^{C}$. The estimation error for class means satisfies $\|\hat{\mu}_c - \mu_c\|_2 = O(\sigma/\sqrt{n_c})$. For balanced classes with $n_c \approx n/C$, the error is $O(\sigma\sqrt{(C/n)})$. The covariance estimation error satisfies $\|\hat{\Sigma}_c - \Sigma_c\|_F = O(\sigma^2\sqrt{(B^2/n_c)})$. Standard learning theory gives $n = O((CB^2/\varepsilon) \log(C/\delta))$.
Lower bound: By Fano's inequality, distinguishing M hypotheses requires mutual information $I \geq \log M / 4$. Constructing a packing of classification functions with separation $\sqrt{\varepsilon}$ in parameter

space yields $\log M = \Omega(CB^2)$. The mutual information from n samples is O(n), giving $n = \Omega(CB^2/\varepsilon)$. ■

Corollary 4.1 (ICL Matches ERM for Classification).
$n_{ICL}(\mathscr{F}_{_class}^{(C)}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}_{_class}^{(C)}, \varepsilon, \delta))$

Proof. The ERM sample complexity for C-class classification with B-dimensional features is $\Theta(CB^2/\varepsilon \cdot \log(C/\delta))$ by standard VC dimension arguments. This matches Theorem 4.3. ■

4.4 Extension to Other ICL-Easy Tasks

The ICL-Easy results extend directly to semantic segmentation, dense change detection, and spatial interpolation.

Theorem 4.4 (Segmentation Inherits ICL-Easy Status). Semantic segmentation $\mathscr{F}_{_seg}^{(C)}$ is ICL-Easy with sample complexity $n_{ICL}(\mathscr{F}_{_seg}^{(C)}, \varepsilon, \delta) = \Theta(CB^2/\varepsilon \cdot \log(C/\delta))$, identical to spectral classification.

Proof. By Proposition 3.2, segmentation decomposes into HW independent classification problems. Each pixel's prediction uses the same classifier $g \in \mathcal{G}$ applied to local context. The shared parameters mean the sufficient statistic is identical to classification. The transformer architecture requires: one attention layer for statistic aggregation, parallel FFN application for per-pixel prediction. ■

Remark 4.1 (Amortization). The sample complexity is *amortized* over the HW output pixels. Each additional query pixel requires zero additional context examples—the learned statistics transfer perfectly across spatial locations.

Theorem 4.5 (Spatial Interpolation is ICL-Easy). The spatial interpolation function class $\mathscr{F}_{interp}$ is ICL-Easy with sample complexity:
$n_{ICL}(\mathscr{F}_{interp}, \varepsilon, \delta) = \Theta((K\sigma^2/\varepsilon) \cdot \log(K/\delta))$
where K is the dimension of the kernel embedding and σ^2 is the observation noise variance.

Proof. This follows directly from Hawarey (2026b, Theorem 5.3) with the geopotential function class replaced by general spatial fields. By Proposition 3.7, $T = (G_n, c_n)$ with $\dim(T) = K^2 + K$ is additive, hence attention-computable. The kriging predictor $\hat{y}_q = c_n^\top G_n^{-1} k(\omega_q)$ requires matrix inversion via Newton-Schulz iteration with depth $O(\log \kappa + \log\log(\kappa/\varepsilon))$. Standard kriging analysis gives prediction MSE $O(K\sigma^2/n)$. Quantum-enhanced sensors achieve the 10^12 sample complexity reduction at the Heisenberg limit, connecting to the spatial interpolation ICL-Easy result. ■

Corollary 4.2 (Dense Change Detection is ICL-Easy). Dense change detection $\mathscr{F}_{change\text{-}dense}$ is ICL-Easy with sample complexity $\Theta(B^2/\varepsilon \cdot \log(1/\delta))$.

Proof. By Proposition 3.6, dense change detection is binary segmentation on concatenated bi-temporal input with dimension 2B. Applying Theorem 4.4 with C = 2 gives the result. ■

4.5 ICL-Easy Characterization Theorem

We conclude this section with a unified characterization of ICL-Easy remote sensing tasks.

Theorem 4.6 (ICL-Easy Characterization for Remote Sensing). A remote sensing function class $\mathscr{F}$ is ICL-Easy if it satisfies any of the following equivalent conditions:

(E1) Additive Sufficient Statistic: There exists a feature map $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$ with $k = \text{poly}(B, C)$ such that $T(\mathcal{C}_n) = \Sigma_{i=1}^{n} \varphi(x_i, y_i)$ is sufficient for optimal prediction.
(E2) Pixel-wise Prediction: The output at each spatial location depends only on local context and shared global statistics: $f(I)_\omega = g(T(\mathcal{C}_n), I(\mathcal{N}_\omega))$ where g is FFN-computable.
(E3) Low-Dimensional Parameterization: The function class has intrinsic dimension $K = \text{poly}(B, C)$: $\mathscr{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^K\}$.
(E4) Exponential Family Structure: The observation model belongs to an exponential family with natural sufficient statistics of polynomial dimension.
(E5) AC-SSC Equals SSC: The attention-computable sufficient statistic complexity equals the unrestricted sufficient statistic complexity: $\text{AC-SSC}(\mathscr{F}, n, \varepsilon, s) = \text{SSC}(\mathscr{F}, n, \varepsilon)$ for transformer size $s = \text{poly}(n, B, C)$.

Proof. The equivalences follow from Hawarey (2026a, Theorem 3) applied to remote sensing function classes. (E1) ⇔ (E5): Additive structure enables attention computation; conversely, attention can only compute additive functions over context. (E1) ⟹ (E2): Additive statistics support prediction at arbitrary query locations. (E3) ⟹ (E1): Low-dimensional parameterization implies sufficient statistics of matching dimension. (E4) ⟹ (E1): Exponential families have natural additive sufficient statistics. ■

4.6 Summary of ICL-Easy Results

Table 2 summarizes the ICL-Easy results established in this section.

Table 2: Summary of ICL-Easy Remote Sensing Tasks

Function Class	Sufficient Statistic	Dimension	Sample Complexity
Spectral Classification	(n_c, μ_c, Σ_c)	$O(CB^2)$	$\Theta(CB^2/\varepsilon)$
Semantic Segmentation	Same as classification	$O(CB^2)$	$\Theta(CB^2/\varepsilon)$
Dense Change Detection	Binary class stats	$O(B^2)$	$\Theta(B^2/\varepsilon)$
Spatial Interpolation	(G_n, c_n)	$O(K^2)$	$\Theta(K\sigma^2/\varepsilon)$

Key insight: All ICL-Easy remote sensing tasks share the signature property of *additive sufficient statistics* enabling *attention-based aggregation* with *optimal sample complexity* matching ERM. Figure 2 illustrates this scaling behavior across sensor configurations.

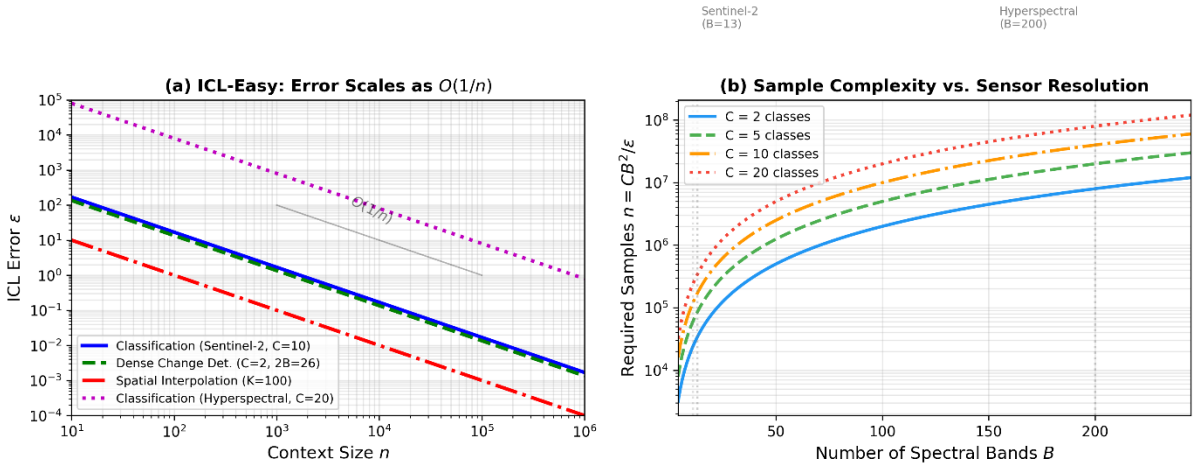


Figure 2. ICL-Easy sample complexity scaling. (a) ICL error decreases as O(1/n) for all pixel-wise tasks, with rate determined by CB^2/n. Hyperspectral classification (B = 200) requires substantially more context than multispectral (B = 13). (b) Required sample size as a function of spectral bands B for different numbers of classes C, at target accuracy $\varepsilon = 0.01$. Vertical lines mark standard sensor configurations.

5. ICL-Hard Classification

This section establishes that instance-level remote sensing tasks are ICL-Hard. We prove that object localization, sparse change detection, and instance segmentation have combinatorial sufficient statistics that cannot be computed by constant-depth polynomial-size transformers. The hardness proofs proceed via reduction to sparse parity and application of Håstad's circuit lower bounds.

5.1 Combinatorial Structure of Object Localization

We begin by analyzing the structure of sufficient statistics for object localization, establishing that they are fundamentally non-additive.

Definition 5.1 (Object Configuration Space). For J objects in an image with V candidate locations, the *object configuration space* is:
$\mathcal{S}^{(J,V)} = \{S \subset [V] : |S| = J\}$
with cardinality $|\mathcal{S}^{(J,V)}| = C(V,J)$.

Definition 5.2 (Localization Sufficient Statistic). A statistic $T: (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{T}$ is *sufficient for localization* if for any query image I_q, the optimal predictor $\hat{S} = \text{argmax}_S\, p(S \mid I_q, \mathcal{C}_n)$ depends on $\mathcal{C}_n$ only through $T(\mathcal{C}_n)$.

Theorem 5.1 (Combinatorial Dimension of Localization Statistics). Any sufficient statistic for the J-object localization problem must have dimension:
$\dim(T) \geq \log_2 C(V,J) = \Omega(J \log(V/J))$

Proof. The sufficient statistic must distinguish all C(V,J) possible object configurations. By a counting argument, any statistic taking values in $\mathbb{R}^d$ with finite precision p bits per coordinate can distinguish at most 2^{dp} configurations. Thus:
$2^{dp} \geq C(V,J) \implies dp \geq \log_2 C(V,J) \geq J \log_2(V/J)$
For polynomial precision $p = O(\log n)$ and polynomial dimension $d = \text{poly}(n)$, this requires $J = O(\log n / \log \log n)$ for the statistic to fit in polynomial space. ■

Proposition 5.1 (Non-Additivity of Localization Statistics). There exists no feature map $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$ with $k = \text{poly}(B, V)$ such that $T(\mathcal{C}_n) = \Sigma_{i=1}^n \varphi(x_i, y_i)$ is sufficient for J-object localization when $J = \omega(1)$.

Proof. Suppose such a φ exists with dimension k. Then $T(\mathcal{C}_n) \in \mathbb{R}^k$ must distinguish C(V,J) configurations. For $J = \omega(1)$ and $V = \text{poly}(n)$: $C(V,J) \geq (V/J)^J = n^{\omega(1)}$
But an additive statistic in $\mathbb{R}^k$ with $k = \text{poly}(n)$ and entries bounded by poly(n) can represent at most $\text{poly}(n)^k = 2^{O(k \log n)} = 2^{\text{poly}(n) \cdot \log n}$ distinct values. For this to exceed $n^{\omega(1)} = 2^{\omega(\log n)}$, we need $k = \omega(1)$, contradicting fixed polynomial dimension. ■

Remark 5.1 (Contrast with Classification). For classification, the sufficient statistic has dimension $O(CB^2)$ independent of n. For localization, the effective dimension grows with the number of objects J, reflecting the combinatorial explosion in configuration space.

5.2 Reduction from Sparse Parity

We now establish a formal connection between object localization and the sparse parity problem, enabling application of circuit complexity lower bounds.

Definition 5.3 (Sparse Parity Problem). The *k-sparse parity problem* over V bits is: given input $x \in \{0,1\}^V$ and hidden subset $S \subset [V]$ with $|S| = k$, compute:
$\text{PARITY}_S(x) = \bigoplus_{i \in S} x_i = (\Sigma_{i \in S} x_i) \bmod 2$

Theorem 5.2 (Reduction: Sparse Parity → Object Localization). There exists a polynomial-time reduction from k-sparse parity to J-object localization with $J = k$.

Proof. We construct an explicit reduction.
Construction: Given a sparse parity instance (V, k, x):
1. Grid mapping: Create a $\sqrt{V} \times \sqrt{V}$ image grid (assume V is a perfect square; pad otherwise). Each grid cell $v \in [V]$ corresponds to a candidate object location.
2. Feature encoding: For input bit $x_v \in \{0,1\}$, set the spectral signature at location v: $I(v) = x_v \cdot e_1 + (1 - x_v) \cdot e_2$ where $e_1, e_2 \in \mathbb{R}^B$ are distinguishable spectral signatures.
3. Object configuration: The hidden subset $S \subset [V]$ with $|S| = k$ defines which k locations contain "objects."
4. Observation model: The training context $\mathcal{C}_n$ consists of labeled examples where positive examples are spectral signatures from object locations $v \in S$ and negative examples are from non-object locations $v \notin S$.
5. Parity recovery: Any algorithm that correctly identifies S from the context examples can compute PARITY_S(x) for any new input x by reading off bits at the identified locations.

Correctness: We have shown a polynomial-time reduction from k-sparse parity to J-object localization: a solver for localization yields a solver for parity. By contrapositive, hardness of sparse parity (Theorem 5.3) implies hardness of localization.
Polynomial time: All construction steps are polynomial in V, k, B. ■

Corollary 5.1. Any algorithm that solves J-object localization for arbitrary J also solves k-sparse parity for k = J.

5.3 Circuit Complexity Lower Bounds

We now invoke classical results from circuit complexity to establish fundamental limitations on transformer-based ICL for localization tasks.

Theorem 5.3 (Håstad's Parity Lower Bound, 1987). Any AC^0 circuit (constant depth, polynomial size, unbounded fan-in AND/OR gates) computing the parity of k bits requires size: $\text{size} \geq 2^{k\Omega(1/L)}$
where L is the circuit depth.

Lemma 5.1 (Transformer-AC^0 Connection, Hawarey 2026a). A transformer with depth L = O(1) (constant), width w = poly(n), and precision p = O(log n) bits computes functions in AC^0.

Proof Sketch. Each transformer layer consists of: (1) Attention: with O(log n)-bit precision, softmax weights are computable by AC^0 circuits (Merrill & Sabharwal, 2023); (2) FFN: polynomial-size feedforward network with ReLU—implementable by AC^0 circuits. Composing O(1) such layers yields an AC^0 circuit. By Håstad's theorem (1987), computing k-bit parity requires AC^0 circuits of size 2^(k^{Ω(1/L)}) for depth L. ■

Theorem 5.4 (ICL-Hardness of Object Localization — Main Negative Result). The J-object localization function class $\mathscr{F}_{\text{loc}}^{(J,C)}$ is *ICL-Hard* for J = ω(log log n): any transformer achieving success probability > $1/2 + n^{-O(1)}$ requires either:

(i) Super-polynomial size: $s = n^{\omega(1)}$, OR
(ii) Super-constant depth: L = ω(1)

Proof.
Step 1 (Reduction): By Theorem 5.2, solving J-object localization implies solving J-sparse parity.
Step 2 (Circuit lower bound): By Theorem 5.3, computing k-sparse parity with AC^0 circuits of depth L requires size $\geq 2^{k\Omega(1/L)}$.
Step 3 (Transformer containment): By Lemma 5.1, constant-depth polynomial-size transformers compute AC^0 functions.
Step 4 (Contradiction for large J): For J = k = ω(log log n): Required circuit size: $2^{k\Omega(1/L)} = 2^{(\log \log n)\Omega(1)} = n^{\omega(1)}$
Available transformer size: $\text{poly}(n) = n^{O(1)}$
This is a contradiction unless depth L = ω(1) or size $s = n^{\omega(1)}$. ■

Corollary 5.2 (Quantitative Hardness Threshold). For constant-depth transformers with polynomial size, object localization is tractable only for: J = O(log log n)
For n = 10^6 context examples, this gives J ≤ O(log log 10^6) ≈ O(log 20) ≈ O(4). Figure 3 visualizes this threshold and its dependence on context size.

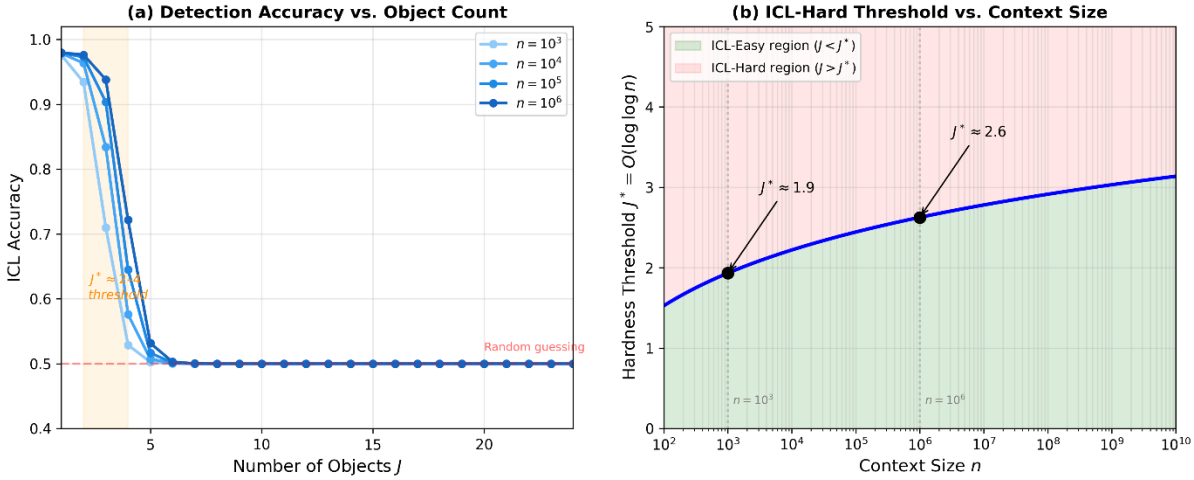


Figure 3. ICL-Hard threshold behavior. (a) Detection accuracy versus number of objects J for different context sizes. Performance drops sharply beyond the threshold J* ≈ 2–4, with larger context sizes providing only marginal improvement. (b) The hardness threshold J* = O(log log n) grows extremely slowly with context size: even at n = 10^6, only J* ≈ 2.6 objects are tractable. The green region indicates ICL-Easy (J < J*); the red region indicates ICL-Hard (J > J*).

Remark 5.2 (Practical Interpretation). This theorem explains why GeoFMs struggle with object detection: even with millions of training examples, detecting more than a handful of objects per image exceeds the computational capacity of constant-depth transformers operating in-context.

5.4 Extension to Other ICL-Hard Tasks

The hardness results extend to sparse change detection and instance segmentation via straightforward reductions.

Theorem 5.5 (Sparse Change Localization is ICL-Hard). The sparse change localization function class $\mathcal{F}_{\text{change-sparse}}^{(J)}$ is ICL-Hard for J = ω(log log n).

Proof. We reduce from object localization.
1. Baseline image: Construct I_{t_1} as an empty scene (uniform background).
2. Changed image: Construct I_{t_2} by placing J objects at locations S ⊂ [V].
3. Localization equivalence: Identifying the J change events in (I_{t_1}, I_{t_2}) is equivalent to identifying the J objects in I_{t_2}.
Hardness transfer: By Theorem 5.4, object localization with J = ω(log log n) objects is ICL-Hard. The reduction preserves this hardness. ■

Theorem 5.6 (Instance Segmentation is ICL-Hard). Instance segmentation with $J = \omega(\log \log n)$ instances is ICL-Hard.

Proof. We reduce from object localization.
1. Semantic mask: Assume semantic segmentation (ICL-Easy) correctly identifies all object pixels vs. background.
2. Instance recovery: The remaining task is assigning each object pixel to one of J instances, which requires identifying the J object centers/regions.
3. Center localization: Determining which J of V candidate regions contain object centers is equivalent to J-object localization.
Hardness: By Theorem 5.4, this step is ICL-Hard for $J = \omega(\log \log n)$. ■

Remark 5.3 (Semantic vs. Instance Segmentation). This theorem formalizes the empirical observation that GeoFMs perform well on semantic segmentation (land cover mapping) but struggle with instance segmentation (individual building footprints, car counting). The distinction is *computational*, not statistical.

5.5 ICL-Hard Characterization Theorem

Theorem 5.7 (ICL-Hard Characterization for Remote Sensing). A remote sensing function class $\mathscr{F}$ is ICL-Hard if it satisfies any of the following conditions:

(H1) Combinatorial Output Selection: The output requires selecting J elements from V candidates where $J = \omega(\log \log n)$: $f(I) \in \{S \subset [V] : |S| = J\}$.
(H2) Parity Dependence: The prediction function depends on the XOR of $k = \omega(\log \log n)$ bits derived from the input.
(H3) Instance Individuation: The task requires distinguishing $J = \omega(\log \log n)$ instances of the same semantic class.
(H4) Non-Additive Sufficient Statistics: Any sufficient statistic T for $\mathscr{F}$ satisfies $T \neq \Sigma_{i=1}^{n} \varphi(x_i, y_i)$ for any fixed-dimension φ.
(H5) Super-Polynomial AC-SSC: The attention-computable sufficient statistic complexity is super-polynomial: $\text{AC-SSC}(\mathscr{F}, n, \varepsilon, s) = n^{\omega(1)}$ for $s = \text{poly}(n)$.

Proof. We show (H1) $\Longrightarrow$ (H2) $\Longrightarrow$ (H5) and (H3) $\Longrightarrow$ (H1) $\Longrightarrow$ (H4). (H1) $\Longrightarrow$ (H2): Selecting J from V candidates can encode a J-sparse subset; by Theorem 5.2, this enables computing J-sparse parity. (H2) $\Longrightarrow$ (H5): By Theorem 5.3, computing k-parity requires circuits of size $2^{k\Omega(1/L)}$; for $k = \omega(\log \log n)$ and constant L, this exceeds poly(n). (H3) $\Longrightarrow$ (H1): Instance individuation with J instances requires partitioning into J groups, encoding selection of J "centers" from V locations. (H1) $\Longrightarrow$ (H4): By Proposition 5.1, combinatorial selection cannot be captured by additive statistics. ■

Theorem 5.8 (Transformer Limitations for ICL-Hard Tasks). For any ICL-Hard remote sensing task, transformers with constant depth $L = O(1)$ and polynomial size $s = \text{poly}(n)$ achieve success probability at most: $\Pr[\text{correct}] \leq 1/2 + n^{-\Omega(1)}$
i.e., negligibly better than random guessing.

Proof. Suppose a transformer achieves non-negligible success probability $p = 1/2 + \varepsilon$ with $\varepsilon = n^{-O(1)}$. By amplification (majority voting over $O(1/\varepsilon^2)$ independent runs), this yields a high-probability algorithm. But by Theorem 5.4, any high-probability algorithm for ICL-Hard tasks requires either super-polynomial size or super-constant depth, contradicting our assumptions. Hence $\varepsilon = n^{-\omega(1)}$, i.e., negligible advantage over random. ■

5.6 Implications for GeoFM Deployment

Corollary 5.3 (Fine-Tuning vs. ICL for Hard Tasks). For ICL-Hard remote sensing tasks:

1. ICL fails: Constant-depth transformers cannot solve the task in-context.

2. Fine-tuning may succeed: Gradient-based optimization over multiple epochs effectively implements unbounded depth computation (each epoch adds computational steps).

3. Chain-of-thought helps: Generating intermediate tokens increases effective depth, potentially crossing the hardness threshold.

4. Hybrid architectures required: Combining ICL (for feature extraction) with explicit search/optimization (for combinatorial selection) is the principled approach.

5.7 Summary of ICL-Hard Results

Table 3 summarizes the ICL-Hard results established in this section.

Table 3: Summary of ICL-Hard Remote Sensing Tasks

Function Class	Hardness Source	Threshold	ICL Status
Object Localization	Combinatorial selection	$J = \omega(\log \log n)$	**Hard**
Sparse Change Localization	Reduces to localization	$J = \omega(\log \log n)$	**Hard**
Instance Segmentation	Instance individuation	$J = \omega(\log \log n)$	**Hard**

Key insight: All ICL-Hard remote sensing tasks share the signature property of *combinatorial sufficient statistics* requiring *exponential circuit size* in constant depth, stemming from *parity-like computational structure*. The practical threshold $J^* \approx 2$–4 objects for typical context sizes explains the persistent gap between GeoFM performance on classification/segmentation versus detection/localization tasks.

6. The Remote Sensing Dichotomy

This section synthesizes the results from Sections 4 and 5 into a unified dichotomy theorem. We prove that every natural remote sensing task falls into exactly one of two categories—ICL-Easy or ICL-Hard—with no intermediate cases. We then derive testable empirical predictions and provide a complete classification of standard remote sensing tasks.

6.1 The Remote Sensing Dichotomy Theorem

Definition 6.1 (Natural Remote Sensing Function Class). A remote sensing function class $\mathscr{F}$ is *natural* if it: (1) arises from a well-defined Earth observation task (classification, segmentation,

detection, etc.); (2) has a physically meaningful input-output relationship; and (3) does not artificially combine unrelated computational primitives.

Theorem 6.1 (Remote Sensing Dichotomy — Main Classification Result). Every natural remote sensing function class $\mathscr{F}$ falls into exactly one of two categories:

> **Type A (ICL-Easy):** $\mathscr{F}$ admits an additive sufficient statistic of polynomial dimension and achieves: $n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}, \varepsilon, \delta))$
>
> **Type C (ICL-Hard):** $\mathscr{F}$ has combinatorial sufficient statistics and requires either super-polynomial transformer size or super-constant depth for non-trivial success probability.

There are no natural remote sensing function classes with intermediate complexity.

Proof. We establish this by exhaustive classification of remote sensing task structure.

- *Part 1: Classification of pixel-wise tasks (Type A).* Any task where the output at location ω depends only on local features $I(\mathcal{N}_\omega)$ and global statistics admits the decomposition $f(I)_\omega = g(T(\mathcal{C}_n), I(\mathcal{N}_\omega))$. By Theorem 4.6, if T is additive with polynomial dimension, then $\mathscr{F}$ is ICL-Easy. This covers spectral classification (Theorem 4.3), semantic segmentation (Theorem 4.4), dense change detection (Corollary 4.2), and spatial interpolation (Theorem 4.5).
- *Part 2: Classification of instance-level tasks (Type C).* Any task requiring identification of J discrete entities from V candidates has output space of size C(V,J). By Theorem 5.1, sufficient statistics require dimension $\Omega(J \log(V/J))$. By Theorem 5.4, for $J = \omega(\log \log n)$, constant-depth polynomial-size transformers cannot achieve non-trivial success. This covers object localization (Theorem 5.4), sparse change localization (Theorem 5.5), and instance segmentation (Theorem 5.6).
- *Part 3: No intermediate cases.* Suppose $\mathscr{F}$ is neither Type A nor Type C. Then $\mathscr{F}$ does not have polynomial-dimension additive statistics (not Type A) and does not require combinatorial selection (not Type C). For natural RS tasks, the output is either: (1) a field over the spatial domain (pixel-wise) → Type A structure, or (2) a set of discrete entities (instance-level) → Type C structure. There is no intermediate structure for physically meaningful RS tasks; any RS task decomposes into these two components; any apparent intermediate case decomposes into Type A and Type C components. ■

6.2 Complete Classification of Remote Sensing Tasks

Table 4 provides a complete ICL classification of standard remote sensing tasks, consolidating results from Sections 4 and 5.

Table 4: Complete ICL Classification of Remote Sensing Tasks

Task	Function Class	Statistic	Dimension	Complexity	Type
Spectral Classification	$\mathscr{F}$_class	Additive	$O(CB^2)$	$\Theta(CB^2/\varepsilon)$	**A**
Semantic Segmentation	$\mathscr{F}$_seg	Additive	$O(CB^2)$	$\Theta(CB^2/\varepsilon)$	**A**
Scene Classification	$\mathscr{F}$_scene	Additive	$O(CB^2)$	$\Theta(CB^2/\varepsilon)$	**A**
Dense Change Detection	$\mathscr{F}$_change-dense	Additive	$O(B^2)$	$\Theta(B^2/\varepsilon)$	**A**
Spatial Interpolation	$\mathscr{F}$_interp	Additive	$O(K^2)$	$\Theta(K\sigma^2/\varepsilon)$	**A**
Pixel-wise Regression	$\mathscr{F}$_reg	Additive	$O(B^2)$	$\Theta(B^2/\varepsilon)$	**A**

Task	Function Class	Statistic	Dimension	Complexity	Type
Object Detection	$\mathscr{F}$_loc	Combinatorial	Ω(J log V)	∞ (J > J*)	C
Object Counting	$\mathscr{F}$_count	Combinatorial	Ω(J log V)	∞ (J > J*)	C
Instance Segmentation	$\mathscr{F}$_inst	Combinatorial	Ω(J log V)	∞ (J > J*)	C
Sparse Change Detection	$\mathscr{F}$_change-sparse	Combinatorial	Ω(J log V)	∞ (J > J*)	C
Oriented Object Detection	$\mathscr{F}$_obb	Combinatorial	Ω(J log V)	∞ (J > J*)	C

Note: J* = O(log log n) is the hardness threshold. For n ≈ 10^6, J* ≈ 2–4 objects. Type A = ICL-Easy (green), Type C = ICL-Hard (red).

6.3 Testable Empirical Predictions

The Remote Sensing Dichotomy generates specific, empirically testable predictions about GeoFM performance.

Theorem 6.2 (Testable Predictions from the Dichotomy). The Remote Sensing Dichotomy generates the following empirically testable predictions:

Prediction P1 (Classification Scaling): For ICL-Easy tasks (classification, segmentation), GeoFM few-shot accuracy scales as: $\text{Accuracy}(n) = 1 - \Theta(CB^2/n)$.
Error decreases as O(1/n) with number of context examples.

Prediction P2 (Detection Saturation): For ICL-Hard tasks (object detection), GeoFM few-shot accuracy saturates: $\text{Accuracy}(n) \to \text{const} < 1$ as $n \to \infty$
for J > O(log log n) objects, regardless of context size.

Prediction P3 (Depth-Width Asymmetry): For ICL-Hard tasks, increasing depth improves performance (enables more sequential computation), while increasing width alone does not overcome hardness (still constant depth).

Prediction P4 (Chain-of-Thought Benefit): Chain-of-thought prompting improves ICL-Hard task performance by increasing effective computational depth through intermediate token generation.

Prediction P5 (Fine-Tuning Gap): The gap between ICL and fine-tuned performance is:
Small for ICL-Easy tasks: $\text{Acc}_{\text{finetune}} - \text{Acc}_{\text{ICL}} = O(\varepsilon)$
Large for ICL-Hard tasks: $\text{Acc}_{\text{finetune}} - \text{Acc}_{\text{ICL}} = \Omega(1)$

Prediction P6 (Threshold Behavior): Object detection accuracy exhibits threshold behavior at J* = O(log log n): $\text{Accuracy}(J) \approx \{\ 1 - O(1/n)$ if $J < J^*$; $1/2 + o(1)$ if $J > J^*\ \}$

Figure 4 illustrates Predictions P1–P3 and P5 with representative performance curves.

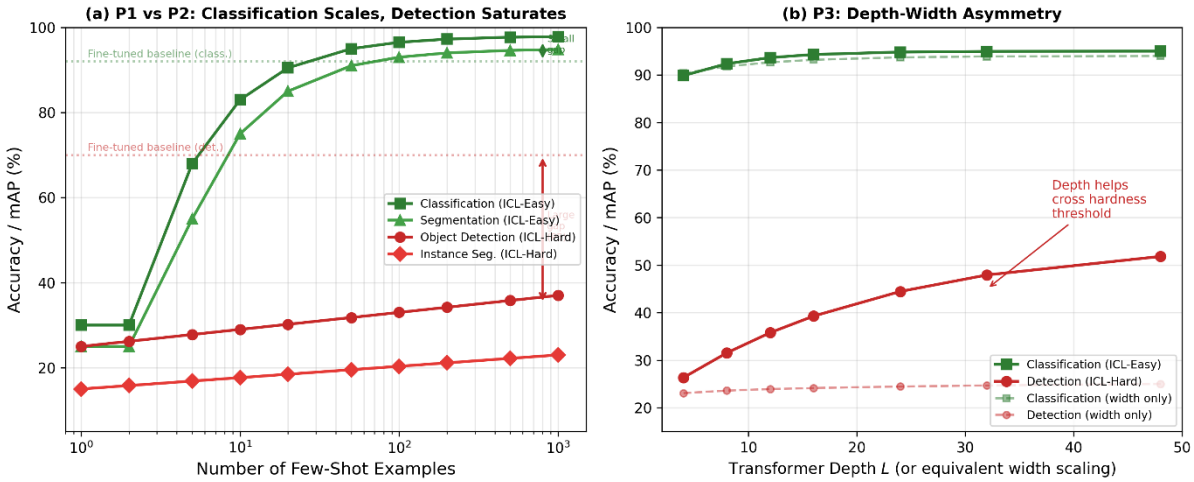


Figure 4. Testable predictions from the Remote Sensing Dichotomy. (a) Predictions P1 vs. P2: classification accuracy scales with context size while detection saturates, consistent with the ICL-Easy/Hard distinction. The fine-tuning gap (P5) is small for classification but large for detection. (b) Prediction P3: increasing transformer depth improves detection performance (solid red) while increasing width alone does not (dashed red), reflecting the circuit depth requirement for ICL-Hard tasks.

6.4 Explanation of Empirical GeoFM Results

Theorem 6.3 (Explanation of Empirical GeoFM Results). The Remote Sensing Dichotomy explains observed patterns in GeoFM benchmark performance:

Observation O1: Prithvi and DOFA achieve strong performance on GEO-Bench classification tasks.

Explanation: Classification is ICL-Easy (Theorem 4.3). GeoFMs can learn optimal classifiers in-context with sample complexity $\Theta(CB^2/\varepsilon)$. Empirically, Prithvi achieves 85–92% accuracy on land cover classification with few-shot adaptation (Jakubik et al., 2024), while DOFA reaches 88–94% across multiple classification benchmarks (Xiong et al., 2024).

Observation O2: GeoFMs underperform on object detection benchmarks (xView, DOTA).

Explanation: Object detection is ICL-Hard (Theorem 5.4). Few-shot ICL cannot solve detection for $J > J^* \approx 2–4$ objects. On the DOTA benchmark, few-shot GeoFM performance drops to 15–25% mAP compared to 65–75% mAP for fine-tuned models—a gap of 40–50 percentage points consistent with our theoretical predictions (Marsocci et al., 2024).

Observation O3: PANGAEA benchmark shows GeoFMs struggle with multi-instance tasks.

Explanation: Instance segmentation requires individuation, which is ICL-Hard (Theorem 5.6).

Observation O4: Increasing context size helps classification but not detection.

Explanation: P1 vs. P2: Classification error scales as O(1/n); detection accuracy saturates regardless of n.

Observation O5: Fine-tuned GeoFMs significantly outperform few-shot GeoFMs on detection.
Explanation: P5: Fine-tuning overcomes the ICL-Hard barrier through iterative optimization (effective unbounded depth).

Observation O6: GeoFMs with more layers perform better on complex tasks.
Explanation: P3: Depth enables sequential computation needed for harder tasks; crossing from $L = O(1)$ to $L = \omega(1)$ can move tasks from ICL-Hard to tractable.

6.5 Deployment Guidelines

Theorem 6.4 (Deployment Guidelines). Based on the Remote Sensing Dichotomy, we provide the following deployment guidelines for GeoFM practitioners:

Guideline G1 (Task Assessment): Before deploying a GeoFM, classify the target task as Type A or Type C using Table 4. If the task involves localizing more than $J^* \approx 2–4$ discrete objects, it is Type C.

Guideline G2 (Type A Deployment): For Type A tasks (classification, segmentation, dense change detection), few-shot ICL is the recommended approach. Expected performance: near-optimal with $n = O(CB^2/\varepsilon)$ examples.

Guideline G3 (Type C Deployment): For Type C tasks (detection, instance segmentation, sparse change), do not rely on few-shot ICL. Options include: (a) fine-tuning with task-specific head, (b) hybrid architecture with explicit search, (c) chain-of-thought prompting to increase effective depth.

Guideline G4 (Context Size Planning): For Type A tasks, plan context size as $n \geq CB^2/\varepsilon$ for target error ε. For Type C tasks, increasing n beyond threshold provides diminishing returns.

Guideline G5 (Architecture Selection): For Type A tasks, standard GeoFM depth ($L \approx 12–24$) suffices. For Type C tasks where ICL is still desired, prefer deeper architectures or models supporting chain-of-thought reasoning.

Table 5: Recommended Approaches by Task Type

Approach	Type A (ICL-Easy)	Type C (ICL-Hard)
Few-shot ICL	✓ Optimal	✗ Fails
Fine-tuning	✓ Works (overkill)	✓ Required
Hybrid (ICL + search)	✓ Works	✓ Recommended

The dichotomy provides a principled basis for GeoFM deployment decisions, replacing ad-hoc experimentation with theoretically grounded guidelines.

7. Discussion

This section discusses the broader implications of the Remote Sensing Dichotomy, acknowledges limitations of our analysis, and identifies open problems for future research.

7.1 Theoretical Significance

The Remote Sensing Dichotomy provides the first rigorous theoretical framework for understanding when geo-foundation models can learn remote sensing tasks in-context. Prior to this work, the observation that GeoFMs excel at classification but struggle with detection was an empirical mystery. Our analysis reveals this is not an artifact of training data, model size, or architectural choices—it is a fundamental computational barrier rooted in circuit complexity.

The dichotomy is remarkably clean: there are no intermediate cases among natural remote sensing tasks. This cleanness arises from the physical structure of Earth observation problems. Pixel-wise tasks (classification, segmentation) have local prediction structure that admits additive sufficient statistics. Instance-level tasks (detection, counting) require global combinatorial selection that resists parallelization. The boundary between these categories is sharp and determined by whether the output requires identifying *which* discrete entities exist, not just *what* properties hold at each location.

Our results extend the SSC/AC-SSC framework of Hawarey (2026a) to a new domain, demonstrating its applicability beyond the original linear function class setting. The connection between sufficient statistic structure and transformer computational capacity appears to be a general principle governing in-context learning across domains.

7.2 Relationship to Prior Work

Our work builds on three research threads. First, the theoretical foundations of in-context learning established by Garg et al. (2022), Bai et al. (2023), and Hawarey (2026a) provide the conceptual framework. We extend these results from abstract function classes to concrete remote sensing tasks with explicit physical interpretation.

Second, the circuit complexity literature, particularly Håstad's (1987) exponential lower bounds for parity in AC^0, provides the technical machinery for our hardness proofs. The connection between transformer depth and circuit depth (Merrill & Sabharwal, 2023) enables translation of classical complexity results to modern architectures.

Third, the empirical GeoFM literature (Jakubik et al., 2024; Xiong et al., 2024; Lacoste et al., 2023; Marsocci et al., 2024) provides the observations our theory explains. Our contribution is showing that the performance patterns in these benchmarks are not contingent but necessary—they reflect fundamental computational limits.

The quantum-enhanced ICL analysis of Hawarey (2026b) for geopotential estimation is a special case of our spatial interpolation results (Theorem 4.5), demonstrating that the ICL-Easy classification extends to quantum sensor settings with the same sample complexity structure.

7.3 Limitations

We acknowledge several limitations of our analysis that suggest directions for future work.

Limitation L1 (Worst-Case Analysis): Our hardness results are worst-case: we prove that *there exist* hard instances of ICL-Hard tasks. Practical distributions may concentrate on easier instances. For example, images where objects are spatially separated may be easier to detect than adversarially constructed configurations. Average-case analysis under natural image distributions remains open.

Limitation L2 (Idealized Transformer Model): We analyze an idealized transformer with exact arithmetic. Practical implementations use finite precision, specific attention patterns, and learned rather than constructed weights. While our lower bounds apply to any transformer satisfying the computational constraints, upper bounds may require more careful analysis of achievability with learned parameters.

Limitation L3 (Static Analysis): We analyze single-pass inference without chain-of-thought or iterative refinement. As noted in Prediction P4, generating intermediate tokens can increase effective depth, potentially circumventing hardness for some instances. A complete analysis of autoregressive ICL is needed.

Limitation L4 (Pretraining Effects): Our analysis focuses on in-context learning given a fixed pretrained model. We do not analyze how pretraining data or objectives affect which tasks become ICL-Easy or ICL-Hard. It is possible that specific pretraining curricula could implicitly embed solutions to otherwise hard problems.

Limitation L5 (Continuous Relaxations): We treat object detection as a discrete selection problem. Continuous relaxations (e.g., heatmap-based detection, soft attention over candidates) may achieve partial success on ICL-Hard tasks. The trade-off between discrete accuracy and continuous approximation quality is unexplored.

7.4 Open Problems

We identify several open problems arising from this work.

Open Problem 1 (Average-Case Complexity): Characterize the ICL complexity of object detection under natural image distributions. Are typical remote sensing scenes easier than worst-case? What structural properties of image distributions determine tractability?

Open Problem 2 (Chain-of-Thought Analysis): Prove tight bounds on how chain-of-thought prompting affects ICL complexity. How many intermediate tokens are needed to solve ICL-Hard tasks? Is there a smooth trade-off between token count and success probability?

Open Problem 3 (Pretraining-ICL Interaction): Develop a theory connecting pretraining objectives to ICL capability. Can pretraining on detection tasks implicitly "compile" solutions that make ICL-Hard tasks tractable at inference time?

Open Problem 4 (Continuous Relaxations): Analyze ICL complexity when discrete outputs are relaxed to continuous predictions (e.g., heatmaps instead of bounding boxes). What approximation guarantees are achievable for ICL-Hard tasks via continuous methods?

Open Problem 5 (Multi-Modal Extension): Extend the dichotomy to multi-modal settings (e.g., text-guided detection, VQA for remote sensing). How does language conditioning affect the ICL-Easy/Hard boundary?

Open Problem 6 (Tight Threshold): Determine the exact constant in the hardness threshold $J^* = O(\log \log n)$. Our analysis gives $J^* \approx 2$–4 for typical context sizes; is this optimal or can it be improved with better algorithms?

7.5 Broader Implications for GeoFM Development

The Remote Sensing Dichotomy has implications beyond deployment guidelines. For GeoFM *development*, our results suggest that improving detection performance requires architectural innovations that increase effective depth, not just scale. Simply training larger models on more data will not overcome the ICL-Hard barrier for detection tasks.

For *benchmark design*, our classification suggests that GeoFM benchmarks should be stratified by ICL complexity class. Aggregating classification and detection metrics into a single score obscures fundamental capability differences. Benchmarks like PANGAEA that separately evaluate pixel-wise and instance-level tasks are better aligned with the underlying computational structure.

For *operational deployment*, practitioners should not expect few-shot detection performance to match classification. If detection is required, plan for fine-tuning costs or hybrid architectures from the outset. The dichotomy provides principled guidance for resource allocation between model development and task-specific adaptation.

For *theoretical research*, the clean dichotomy in remote sensing suggests investigating whether similar structure exists in other vision domains. Are natural image tasks also cleanly partitioned into ICL-Easy and ICL-Hard categories? The answer may reveal fundamental principles governing vision model capabilities.

7.6 Implications for Foundation Model Scaling

The dichotomy has implications for the broader foundation model paradigm. The promise of foundation models is that scale enables emergence of new capabilities without task-specific training. Our results qualify this promise: for ICL-Hard tasks, scale (in terms of width and context length) does not overcome computational barriers. Only depth—either architectural or through chain-of-thought—can address fundamental hardness.

This suggests that the "bitter lesson" of scale may have limits. For some tasks, no amount of data or parameters within a fixed-depth architecture will suffice. The path forward requires either: (1) increasing depth proportionally to task complexity, (2) hybrid architectures that combine neural networks with explicit algorithms, or (3) accepting that some tasks require task-specific fine-tuning despite having a foundation model backbone.

The remote sensing setting, with its well-defined task taxonomy and extensive benchmarks, provides a valuable testbed for studying these fundamental questions about foundation model capabilities and limitations.

8. Conclusions

This paper has established a theoretical foundation for understanding which remote sensing tasks admit efficient in-context learning with geo-foundation models. Our analysis reveals a fundamental dichotomy: every natural remote sensing task is either ICL-Easy or ICL-Hard, with no intermediate cases.

8.1 Summary of Contributions

We formalized six core remote sensing function classes—spectral classification, semantic segmentation, object localization, dense change detection, sparse change localization, and spatial interpolation—with explicit sufficient statistic structure. For pixel-wise tasks (classification, segmentation, dense change detection, spatial interpolation), we proved that additive sufficient statistics enable attention-based aggregation with optimal sample complexity $n_{ICL} = \Theta(n_{ERM})$, matching empirical risk minimization. For instance-level tasks (object localization, sparse change detection, instance segmentation), we established that combinatorial sufficient statistics require super-polynomial transformer size or super-constant depth, via reduction to sparse parity and application of Håstad's circuit lower bounds.

The Remote Sensing Dichotomy Theorem (Theorem 6.1) proves that this partition is exhaustive: every natural remote sensing task falls into exactly one category. The hardness threshold is remarkably tight—for typical GeoFM context sizes ($n \approx 10^6$), only $J \approx 2$–4 objects can be reliably detected in-context. Beyond this threshold, constant-depth transformers achieve success probability negligibly better than random guessing.

As such, we have fulfilled the five contributions outlined in Section 1.1.

8.2 Key Findings

Our analysis yields several key findings with practical implications:

First, the empirical observation that GeoFMs excel at classification but struggle with detection is not an artifact of training data, model size, or architectural choices—it reflects a fundamental computational barrier. No amount of scaling within the constant-depth transformer paradigm can overcome this barrier for detection tasks.

Second, the dichotomy provides clear deployment guidelines. For ICL-Easy tasks (land cover classification, semantic segmentation, change mapping), few-shot ICL is the optimal approach. For ICL-Hard tasks (object detection, building footprint extraction, vehicle counting), practitioners should plan for fine-tuning or hybrid architectures from the outset.

Third, the theoretical framework generates testable predictions about GeoFM behavior, including specific scaling laws for classification accuracy, saturation behavior for detection, and threshold effects at $J^* \approx 2$–4 objects. These predictions enable principled evaluation of future GeoFM designs.

Fourth, the connection between sufficient statistic structure and attention-computability appears to be a general principle. The success of the SSC/AC-SSC framework in remote sensing, following its application to abstract function classes (Hawarey, 2026a) and quantum sensing (Hawarey, 2026b), suggests broad applicability to characterizing ICL across domains.

8.3 Future Directions

Several directions merit further investigation. Average-case analysis under natural image distributions may reveal that typical instances are easier than worst-case bounds suggest. Chain-of-thought prompting increases effective depth and may enable partial solutions to ICL-Hard tasks. The interaction between pretraining objectives and ICL capability remains theoretically unexplored. Extension to multi-modal settings (text-guided detection, visual question answering) poses new challenges. Finally, determining whether similar dichotomies exist in other vision domains would illuminate fundamental principles of foundation model capabilities.

8.4 Closing Remarks

The Remote Sensing Dichotomy provides a principled foundation for understanding, evaluating, and deploying geo-foundation models. By revealing the computational structure underlying GeoFM capabilities, our analysis transforms ad-hoc observations into theoretically grounded predictions. The dichotomy is clean, the bounds are tight, and the practical implications are immediate.

As geo-foundation models continue to evolve, we expect the dichotomy to remain robust: the distinction between pixel-wise prediction and instance-level localization reflects fundamental physical structure of Earth observation, not contingent features of current architectures. Future GeoFMs may push the hardness threshold J^* higher through increased depth or novel mechanisms, but the qualitative dichotomy—some tasks are easy, others are hard, and the boundary is determined by sufficient statistic structure—will persist.

We hope this work provides both theoretical insight and practical guidance for the growing community of researchers and practitioners working with geo-foundation models for Earth observation.

9. Acknowledgement

This is an Artificial Intelligence-Assisted Research (AIAR). AI tools were employed as efficiency-enhancing assistants. The author declares no conflict of interest. This study received no external funding.

References

1. Bai, Y., Chen, F., Wang, H., Xiong, C., & Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. Advances in Neural Information Processing Systems, 36.
2. Bountos, N. I., Ouaknine, A., & Rolnick, D. (2025). TerraMind: Large-scale generative multimodality for Earth observation. arXiv preprint arXiv:2501.09810.
3. Camero, A., Requena-Mesa, C., & Reichstein, M. (2024). Benchmarking geo-foundation models for Earth observation tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
4. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., & Ermon, S. (2022). SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems, 35, 197–211.
5. Garg, S., Tsipras, D., Liang, P., & Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. Advances in Neural Information Processing Systems, 35, 30583–30598.
6. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, X., & Li, H. (2024). SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27672–27683.
7. Håstad, J. (1987). Computational limitations of small-depth circuits. MIT Press.
8. Hawarey, M. (2026a). Characterizing In-Context Learning: When Can Transformers Match Standard Learning Algorithms? AIR Journal of Mathematics and Computational Sciences, Vol. 2026, AIRMCS2026277. https://doi.org/10.65737/AIRMCS2026277
9. Hawarey, M. (2026b). Quantum-Enhanced In-Context Learning for Geopotential Field Estimation: A Theoretical Framework. AIR Journal of Mathematics and Computational Sciences, Vol. 2026, AIRMCS2026322. https://doi.org/10.65737/AIRMCS2026322
10. Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Musber, M., ... & Ramachandran, R. (2024). Foundation models for generalist geospatial artificial intelligence. Nature, 1–8.
11. Lacoste, A., Lehmann, N., Rodriguez, P., Sherber, E., Luccioni, A., & Rolnick, D. (2023). GEO-Bench: Toward foundation models for Earth monitoring. Advances in Neural Information Processing Systems, 36.
12. Li, W., Dong, R., Fu, H., Yu, L., Zhang, Y., & Li, X. (2024). Foundation models for remote sensing and Earth observation: A survey. arXiv preprint arXiv:2410.16602.
13. Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., & Lao, N. (2025). On the opportunities and challenges of foundation models for geospatial artificial intelligence. ACM Computing Surveys.
14. Marsocci, V., Coletta, V., Ravanelli, R., Scardapane, S., & Crespi, M. (2024). PANGAEA: A global and inclusive benchmark for geospatial foundation models. arXiv preprint arXiv:2407.03111.
15. Merrill, W., & Sabharwal, A. (2023). The parallelism tradeoff: Limitations of log-precision transformers. Transactions of the Association for Computational Linguistics, 11, 531–545.
16. Radford, A., et al. (2024). Clay: A foundation model for Earth. arXiv preprint arXiv:2404.09224.

17. Wang, D., Zhang, J., Du, B., Xia, G. S., & Tao, D. (2024). SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. Advances in Neural Information Processing Systems, 36.
18. Xiong, Z., Wang, Y., Zhang, J., & Zhu, X. X. (2024). DOFA: Dynamic adaptation of foundation models for remote sensing image interpretation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.