# FLAME: Physics-Guided Neural Operators for Onboard Satellite Methane Detection in Hyperspectral Imagery

**Junhyuk Heo** [1] **Junghwan Park** [1] **Sangcheol Sim** [1] **Beomkyu Choi** [1] **Woojin Cho** [1]

## Abstract

Methane is a major driver of near-term climate change, and rapidly identifying its emission sources is a critical climate intervention. Spaceborne hyperspectral imagery is the primary tool for this task, but the volume of data produced by each sensor makes ground-based detection impractical and necessitates onboard detection. Classical methods incur prohibitive computational cost on onboard hardware, while deep learning models are fast but fall short on detection quality. We propose FLAME, a physics-guided neural operator that builds the physics of methane absorption directly into its architecture. On the methane detection benchmark, FLAME achieves the highest detection accuracy among all evaluated methods, reduces the pixel-level false positive rate by nearly $3\times$ over the strongest neural baseline, uses the fewest parameters among learned baselines, and runs within the latency budget of onboard satellite hardware. Codes are available at https://github.com/ROKMC1250/FLAME.

## 1. Introduction

Methane ($CH_4$) is the second largest anthropogenic driver of radiative forcing after carbon dioxide, with a global warming potential roughly 80 times that of $CO_2$ over a 20 year horizon (Masson-Delmotte et al., 2021; Etminan et al., 2016; Saunois et al., 2019). Because its atmospheric lifetime is only about a decade, near-term methane reductions translate directly into slowing of global warming (Shindell et al., 2012; Ocko et al., 2021), making mitigation one of the most cost-effective short-term climate levers. A disproportionate share of anthropogenic methane comes from a small number of point sources: large ultra-emitters in the oil and gas sector account for a substantial fraction of sectoral emissions (Lauvaux et al., 2022; Alvarez et al., 2018; Duren et al., 2019). Identifying and localising these sources is therefore high-leverage, since a single repaired leak can eliminate emissions equivalent to those of thousands of vehicles.

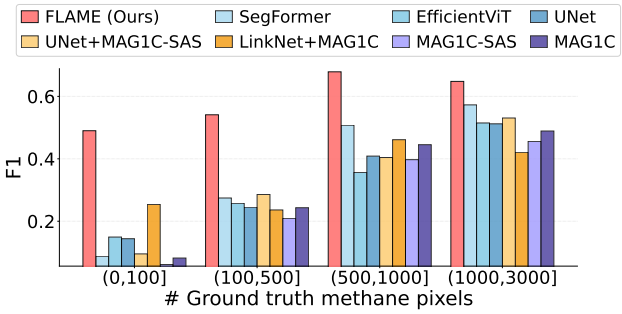


*Figure 1.* **F1 score across plume sizes.** F1 score is measured on the STARCOP dataset, binned by the number of ground-truth plume pixels per tile. FLAME leads across all plume sizes, with the largest margin in the small-plume regime where weak signals are most easily confused with background.

Spaceborne hyperspectral imagery offers a scalable tool for such detection. Methane's distinctive shortwave-infrared absorption features between 2122 and 2488 $nm$ are captured by imaging spectrometers such as AVIRIS-NG and NASA's EMIT (Thompson et al., 2016), and upcoming missions including NASA's Surface Biology and Geology (SBG) (Cawse-Nicholson et al., 2021) and ESA's Copernicus Hyperspectral Imaging Mission (CHIME) (Nieke & Rast, 2018) will extend this toward near-global coverage. The resulting data volumes, however, create a delivery bottleneck: hyperspectral tiles reach hundreds of gigabytes per day per sensor, while downlink bandwidth is limited to a few ground-station passes (Langer et al., 2023). Raw radiance is therefore processed on the ground days after acquisition, by which time an ultra-emission event may have vented for its entire duration.

Onboard processing directly addresses this latency bottleneck: detection on the spacecraft means only plume-containing tiles need full downlinking, and tasking signals can be issued in minutes. This is no longer speculative. ESA's Φ-Sat missions have demonstrated onboard deep-learning inference in orbit (Giuffrida et al., 2021; Esposito et al., 2019), and embedded GPU platforms such as NVIDIA Jetson are being evaluated as payload computers (Furano

[1]TelePIX, Seoul, Republic of Korea. Correspondence to: Woojin Cho <woojin.py@gmail.com>.

Accepted by ICML 2026 AI for Science Workshop.

et al., 2020; Yost & Weston, 2024). The design constraint thus shifts to accuracy under tight power, memory, and runtime budgets.

Existing detection methods sit uneasily within these constraints. Classical matched filters model methane absorption through the Beer–Lambert law and remain the standard tool for hyperspectral methane retrieval (Foote et al., 2020), but their reliance on tile- or column-wide background statistics leads to unstable residuals over heterogeneous scenes and to costly iterative covariance estimation. Purely data-driven segmentation networks are fast, but they discard the physical structure of the problem and in practice over-detect bright or textured surfaces that share no absorption signature with methane. Hybrid two-stage pipelines retain the physics but inherit the runtime of the matched filter they sit on top of. None of these options is simultaneously accurate, fast, and physically grounded.

We argue that the right place to inject physics is inside the network, not around it. We propose FLAME (Fourier Learned Absorption Matched Estimator), a physics-guided neural operator that preserves the functional form of the log-domain matched filter while replacing the quantities that make it fragile, namely the tile-wide background and the global covariance, with pixel-wise estimates produced by a Fourier-based neural operator. The detection score itself is computed by a parameter-free inner-product layer, so the Beer–Lambert structure is baked into the architecture rather than learned from data. A scheduled auxiliary loss initially aligns the learned score with a classical matched-filter product and then decays, allowing the model to surpass its physics teacher while retaining its inductive bias.

**Contributions.**

- **Physics-guided neural operator.** We propose a new approach to methane detection that integrates the physical signal model into a neural operator, recasting retrieval as a function-to-function mapping rather than coupling physics and neural networks externally as in prior work.
- **Strong detection performance with a compact model.** On the STARCOP benchmark, FLAME achieves the highest F1 across all evaluated methods with the fewest parameters among learned baselines, and reduces the pixel-level false positive rate by nearly $3\times$ over the strongest neural baseline.
- **Onboard satellite deployment analysis.** We characterise inference time, power draw, and thermal behaviour on three NVIDIA Jetson modules representative of current and next-generation onboard satellite payload computers.

## 2. Related Works

**Onboard Satellite Processing.** Onboard machine learning for Earth observation has moved from concept to orbit in recent years. The Φ-Sat-1 mission demonstrated onboard inference of a cloud-filtering CNN on a dedicated AI accelerator (Giuffrida et al., 2021), and in-orbit hyperspectral experiments confirmed feasibility on CubeSat-class platforms (Esposito et al., 2019). Subsequent work has benchmarked deep-learning stacks on representative edge hardware (Ziaja et al., 2021) and examined the radiation, power, and thermal constraints that shape payload computer design (Furano et al., 2020; Yost & Weston, 2024). The broader vision of responsive constellations, in which onboard detections trigger follow-up tasking across multiple assets, has been articulated for both disaster response and greenhouse gas monitoring (Růžička et al., 2022; Parr et al., 2024). Within methane detection specifically, onboard constraints have driven work on accelerating classical retrievals (Herec et al., 2025) and on compact learned models with reduced input bandwidth (Růžička & Markham, 2025). FLAME is designed with these constraints as first-class concerns, trading matched-filter iteration for a single-pass neural operator.

**Methane Detection from Hyperspectral Imagery.** The matched filter has been the standard tool for hyperspectral methane retrieval for over a decade. This is because methane is characterized by narrow absorption features in the SWIR range, which hyperspectral sensors can resolve with dense contiguous bands, whereas multispectral imagery averages radiance over a small number of broad bands and can blur these spectral cues. As illustrated in Figure 2, $CH_4$ exhibits distinctive absorption structure across this region. The matched filter models each pixel as a tile mean perturbed by a methane-specific absorption signature under the Beer–Lambert law, whitens the residual by the tile covariance, and projects onto the methane target spectrum (Thompson et al., 2015; Manolakis et al., 2013; Funk et al., 2002). MAG1C (Foote et al., 2020) is the current workhorse in this family, combining an albedo correction with an iteratively-reweighted $\ell_1$ penalty at the cost of repeated covariance inversions, and variants extend it with scene-specific enhancement spectra (Ayasse et al., 2019). Learning-based methods extend the matched-filter family in two directions. Two-stage pipelines such as HyperSTARCOP (Růžička et al., 2023) refine matched-filter score maps with a U-Net (Ronneberger et al., 2015), and lightweight variants pair LinkNet (Chaurasia & Culurciello, 2017) with accelerated filters such as MAG1C-SAS (Herec et al., 2025). Such pipelines inherit the detection ceiling of the underlying filter, since any plume the filter misses cannot be recovered downstream. End-to-end alternatives such as MethaneMapper (Kumar et al., 2023) and HyperspectralViTs (Růžička

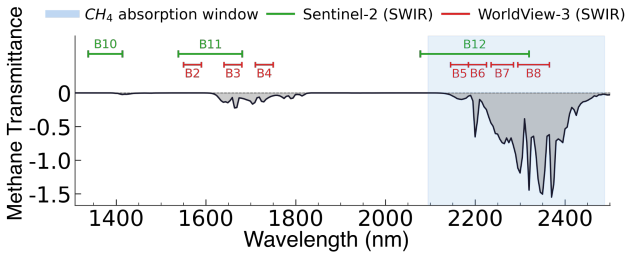


*Figure 2.* Methane absorption spectrum in the SWIR window relevant to plume detection. The broad SWIR bands of multispectral sensors (Sentinel-2, WorldView-3) cannot resolve the narrow absorption features near $2122 - 2488\,nm$.

& Markham, 2025), which adapt SegFormer (Xie et al., 2021) and EfficientViT (Cai et al., 2023) to raw hyperspectral input, remove the matched filter entirely but discard the absorption prior encoded in the Beer–Lambert law. FLAME keeps the matched-filter score as an architectural component while replacing its fragile tile-level quantities with learned pixel-wise estimates.

**Neural Operators for Satellite Data.** Neural operators learn mappings between function spaces rather than between fixed-dimensional tensors, which aligns well with the continuous spatial and spectral structure of remote-sensing data (Kovachki et al., 2023). The Fourier Neural Operator (FNO) parameterises a global convolution in the Fourier domain and has achieved strong results on parametric PDEs (Li et al., 2020) and global weather forecasting (Pathak et al., 2022; Kurth et al., 2023). U-FNO (Wen et al., 2022) adds a U-shaped multi-scale backbone to spectral mixing, producing compact models that capture both global and local structure. Applications to remote sensing so far include tasks such as soil organic carbon estimation (Wong et al., 2023) and climate downscaling (Yang et al., 2024). Our motivation for using a neural operator is that hyperspectral scenes exhibit long-range spatial correlations that global spectral convolutions capture more naturally than local kernels, and that the two unobserved quantities in the log-domain matched filter are both maps from an input function to an output function over the same domain, which matches the native form of neural-operator approximation. FLAME instead integrates the matched-filter signal model into the operator's output, preserving its inner-product structure as an architectural invariant.

## 3. Preliminaries

This section reviews the matched filter for methane detection. We first introduce the Beer–Lambert model in the radiance domain and the classical matched filter derived from its linearization, and then present the log-domain formulation that eliminates the linear approximation. The log-domain derivation motivates the FLAME architecture introduced in Section 4.

### 3.1. Matched Filter under the Beer–Lambert Law

A hyperspectral sensor measures radiance across $p$ discrete SWIR bands, so we work with $\mathbb{R}^p$-valued spectra throughout. Let $L_i \in \mathbb{R}^p$ denote the radiance observed at pixel $i$ over the $p$ bands, and let $L_i^{\mathcal{B}} \in \mathbb{R}^p$ denote the corresponding methane-free background radiance that would be observed in the absence of a plume. Under the Beer–Lambert law, the two quantities are related through the methane column enhancement $\alpha_i \geq 0$ and the fixed unit absorption spectrum $s \in \mathbb{R}^p$, as follows.

$$L_i = L_i^{\mathcal{B}} \odot \exp(\alpha_i\, s), \tag{1}$$

where $\odot$ denotes element-wise multiplication and $\exp(\cdot)$ is applied element-wise. Following the sign convention of Foote et al. (2020), we define $s$ such that $s_j < 0$ in bands of strong methane absorption, which absorbs the sign of the linearization into the spectrum. Classical matched filters operate directly in the radiance domain. Assuming that $\alpha_i\, s$ is small element-wise, a first-order Taylor expansion of Eq. (1) yields the following approximation, where $L_i^{\mathcal{B}}$ is replaced by the tile-shared sample mean $\mu \in \mathbb{R}^p$ in the second line.

$$\begin{aligned} L_i &\approx L_i^{\mathcal{B}} + \alpha_i\,(L_i^{\mathcal{B}} \odot s), \\ L_i &\approx \mu + \alpha_i\,(\mu \odot s) + \eta, \end{aligned} \tag{2}$$

where $\eta$ is an additive noise term. Defining the target spectrum $t = \mu \odot s$, where $C \in \mathbb{R}^{p\times p}$ is the background covariance estimated empirically from the tile, the matched-filter estimate of $\alpha_i$ is given as follows.

$$\hat{\alpha}_i^{\mathrm{MF}} = \frac{(L_i - \mu)^\top C^{-1}(\mu \odot s)}{(\mu \odot s)^\top C^{-1}(\mu \odot s)}. \tag{3}$$

MAG1C (Foote et al., 2020) extends this estimator with two additional components: an albedo correction $r_i = L_i^\top \mu / \mu^\top \mu$ that compensates for the multiplicative nature of the Beer-Lambert attenuation across pixels of varying brightness, and an iteratively reweighted $\ell_1$ regularization that suppresses already-detected methane pixels during background re-estimation. At iteration $k$, the update is given as follows.

$$\hat{\alpha}_i^{(k)} = \max\left(\frac{(L_i - \mu^{(k)})^\top v^{(k)} - w_i^{(k)}}{r_i^{(k)} \cdot m^{(k)}},\ 0\right), \tag{4}$$

where $v^{(k)} = (C^{(k)})^{-1}(\mu^{(k)} \odot s)$ is the whitened target and $m^{(k)} = (\mu^{(k)} \odot s)^\top v^{(k)}$ is the normalization constant, both re-estimated after subtracting previous methane estimates, and $w_i^{(k)} = 1/(r_i^{(k)}(\hat{\alpha}_i^{(k-1)} + \delta))$ is an adaptive sparsity term with stabilization constant $\delta > 0$.

Three limitations follow from this formulation. First, a single tile-level $\mu$ cannot represent heterogeneous surfaces. Second, computing $v^{(k)}$ requires an explicit covariance inversion $(C^{(k)})^{-1}$, which dominates the runtime on onboard satellite hardware. Third, iteration is unavoidable because the initial $\mu$ is contaminated by methane pixels and must be re-estimated after suppressing previous detections.

### 3.2. Log-Domain Matched Filter

The linear approximation in Eq. (2) can be avoided by operating in log space. Taking the element-wise natural logarithm of Eq. (1) yields the following exact linear relation, to which we attach an additive noise term $\eta_i$ that captures sensor noise and unmodeled deviations.

$$\ell_i = \ell_i^{\mathcal{B}} + \alpha_i\, s + \eta_i, \tag{5}$$

where $\ell_i = \log L_i \in \mathbb{R}^p$ is the log-radiance and $\ell_i^{\mathcal{B}} = \log L_i^{\mathcal{B}} \in \mathbb{R}^p$ is the log-background, with $\log(\cdot)$ applied element-wise. Unlike Eq. (2), this relation holds without approximation and remains valid for strong plumes in which $\alpha_i\, s$ is not small. Crucially, the Beer–Lambert attenuation is additive in the log domain rather than multiplicative, so the background radiance no longer scales the target spectrum; as a consequence, the albedo correction $r_i$ present in Eq. (3) is not needed. The covariance of $\eta_i$, denoted $\mathcal{C}_i \in \mathbb{R}^{p\times p}$, is in general pixel-dependent because sensor noise scales with radiance and its log-space counterpart is therefore heteroscedastic.

Under the observation model in Eq. (5), the weighted least-squares estimator of $\alpha_i$ is given as follows.

$$\hat{\alpha}_i^{\log} = \frac{(\ell_i - \ell_i^{\mathcal{B}})^\top\, \mathcal{C}_i^{-1}\, s}{s^\top\, \mathcal{C}_i^{-1}\, s}. \tag{6}$$

The component that determines whether a pixel is detected is the inner product in the numerator, which can be written band-wise as follows.

$$(\ell_i - \ell_i^{\mathcal{B}})^\top\, \mathcal{C}_i^{-1}\, s = \sum_{j=1}^{p} \left[\ell_i - \ell_i^{\mathcal{B}}\right]_j \cdot \left[\mathcal{C}_i^{-1}\, s\right]_j, \tag{7}$$

where $[\,\cdot\,]_j$ denotes the $j$-th SWIR-band component. This expression is a spectrally weighted projection of the log-residual onto the methane spectrum, whitened by the inverse noise covariance. Two unobserved quantities appear: the log-background $\ell_i^{\mathcal{B}}$, and the whitened target $\mathcal{C}_i^{-1}s$. FLAME replaces both with learned estimates while preserving the inner-product structure of Eq. (7).

## 4. Proposed Methods: FLAME

Building on the log-domain formulation in Section 3.2, FLAME estimates the log-background and the whitened target in Eq. (7) using a shared neural operator followed by two lightweight heads.

### 4.1. Formulation

We view each hyperspectral tile as a spectral field $\ell : \Omega \to \mathbb{R}^p$ defined over the spatial domain $\Omega \subset \mathbb{R}^2$, with $\ell(x) \in \mathbb{R}^p$ the log-radiance vector at spatial location $x$. FLAME replaces the two unobserved quantities in Eq. (7) with predictions of the same functional form—a log-background field $\hat{\ell}^{\mathcal{B}} : \Omega \to \mathbb{R}^p$ and a spectral-weight field $\hat{w} : \Omega \to \mathbb{R}^p$—so that detection is cast as a mapping between spectral fields over the same spatial domain. On the discrete $H \times W$ grid, we identify fields with their evaluations at pixel locations and write $\ell_i = \ell(x_i)$, $\hat{\ell}_i^{\mathcal{B}} = \hat{\ell}^{\mathcal{B}}(x_i)$, and $\hat{w}_i = \hat{w}(x_i)$. The backbone $\Phi(\,\cdot\,;\theta_\Phi)$ is a neural operator that maps the input field to a latent feature field; its evaluation at pixel $i$ is given as follows.

$$\mathbf{z}_i = \big[\Phi(\ell;\theta_\Phi)\big](x_i) \in \mathbb{R}^d, \tag{8}$$

with feature dimension $d$. Two point-wise heads $f_{bg}$ and $f_{sw}$ then lift the feature vector to the two physical fields, evaluated pixel-wise as follows.

$$\hat{\ell}_i^{\mathcal{B}} = f_{bg}(\mathbf{z}_i;\theta_{bg}) \in \mathbb{R}^p, \quad \hat{w}_i = f_{sw}(\mathbf{z}_i;\theta_{sw}) \in \mathbb{R}^p. \tag{9}$$

Here $\theta_\alpha = (\theta_\Phi, \theta_{bg}, \theta_{sw})$ collects the parameters that enter the physics-guided score, $\theta_{sg}$ parameterizes the segmentation head introduced in Section 4.2, and $\Theta = (\theta_\alpha, \theta_{sg})$ denotes the full parameter set. The head $f_{bg}$ approximates the log-background $\ell_i^{\mathcal{B}}$ pixel-wise. The head $f_{sw}$ approximates the whitened target through the parameterization $\mathcal{C}_i^{-1}s \to \hat{w}_i \odot s$, in which the fixed absorption spectrum $s$ is retained as a multiplicative physical prior. Substituting Eq. (9) and this parameterization into the log-MF numerator in Eq. (7) yields the FLAME detection score, which we refer to as the physics-guided score and define as follows.

$$\hat{\alpha}_i(\ell;\theta_\alpha) = \big(\ell_i - \hat{\ell}_i^{\mathcal{B}}\big)^\top \big(\hat{w}_i \odot s\big). \tag{10}$$

Although $\hat{\ell}_i^{\mathcal{B}}$ and $\hat{w}_i$ appear without explicit arguments in Eq. (10), both depend on the entire input field $\ell$ through $\Phi$, giving the score a non-local receptive field. The full pipeline runs in a single forward pass that avoids the iterative covariance inversion of Eq. (4), and strictly contains the classical log-domain matched filter as a special case, as shown in Appendix A.

### 4.2. Heads and Score Layer

**Log-background head.** The log-background head $f_{bg}$ is implemented as a single $1 \times 1$ convolution, $\hat{\ell}_i^{\mathcal{B}} = W_{bg}\mathbf{z}_i + b_{bg}$ with $W_{bg} \in \mathbb{R}^{p\times d}$, $b_{bg} \in \mathbb{R}^p$, and parameters $\theta_{bg} = (W_{bg}, b_{bg})$. The bias $b_{bg}$ is initialized to the training-set mean log-spectrum, so the model starts from the shared-mean solution assumed by the classical matched filter and then departs to a pixel-wise estimate as learning proceeds.

**Spectral weight head.** The spectral weight head $f_{sw}$ is also a $1 \times 1$ convolution followed by a softplus activation

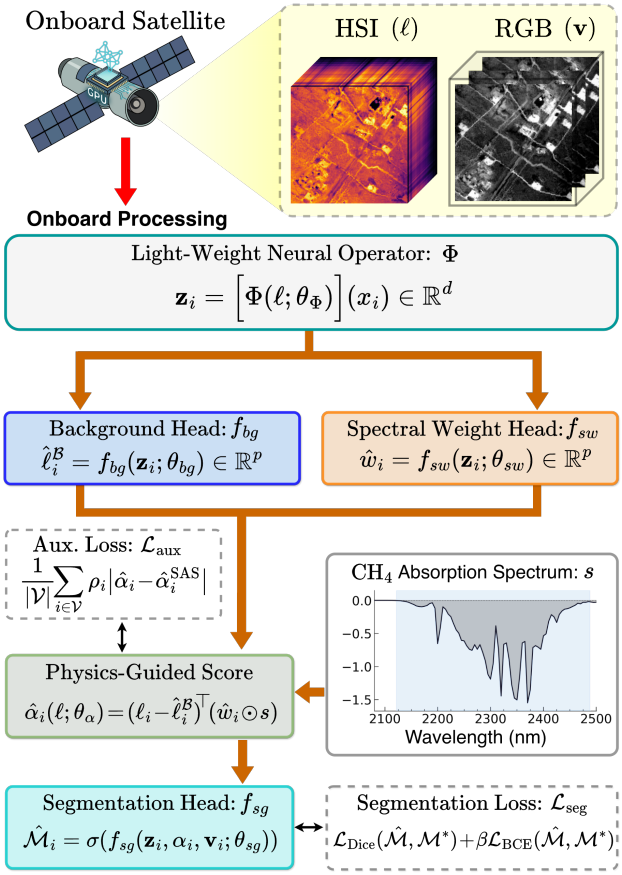


*Figure 3.* **Overview of FLAME.** A neural operator predicts pixel-wise backgrounds and spectral weights, combined with the fixed $CH_4$ prior to yield a physics-guided score and plume mask.

to enforce non-negativity, $\hat{w}_i = \text{softplus}(W_{sw}\mathbf{z}_i + b_{sw})$ with $\theta_{sw} = (W_{sw}, b_{sw})$. $\hat{w}_i$ represents a pixel-adaptive generalization of the diagonal part of $\mathcal{C}_i^{-1}$. Off-diagonal correlations in $\mathcal{C}_i$ are not modeled inside $f_{sw}$, which outputs a vector; instead, they are absorbed by the backbone $\Phi$, which mixes all bands globally before $f_{sw}$ operates on $\mathbf{z}_i$. The explicit factor $s$ in Eq. (10) guarantees that the effective spectral weighting aligns with the methane absorption prior regardless of the learned $\hat{w}_i$.

**Score layer.** The score layer is parameter-free and implements Eq. (10) directly. Because Eq. (5) is exact, the score remains structurally unbiased across the full range of plume intensities, including strong emissions for which the radiance-domain linearization in Eq. (2) would underestimate $\alpha_i$. For notational convenience, throughout the remainder of the paper we let $\hat{\alpha}_i$ denote the score after a fixed normalization by a constant $\tau > 0$ and clipping to $[0, \tau_{\max}]$, where $\tau$ is determined from log-domain MAG1C-SAS statistics on the training set. The same normalization is applied to the MAG1C-SAS reference $\hat{\alpha}_i^{\text{SAS}}$ used for auxiliary supervision in Section 4.3.

**Segmentation head.** The segmentation head $f_{sg}$ incorporates spatial context and implicitly handles the albedo correction $r_i$, the normalization $m$, and the sparsity suppression that appear in Eq. (4). It consumes the concatenation $\mathbf{u}_i = [\mathbf{z}_i, \hat{\alpha}_i, \mathbf{v}_i] \in \mathbb{R}^{d+4}$ of the backbone features, the normalized score, and the normalized RGB triplet $\mathbf{v}_i \in \mathbb{R}^3$. $f_{sg}$ consists of two $3 \times 3$ convolutional blocks with batch normalization and ReLU, followed by a $1 \times 1$ projection to a single logit. The predicted plume probability is $\hat{\mathcal{M}}_i = \sigma(f_{sg}(\mathbf{u}_i;\theta_{sg}))$, where $\sigma$ denotes the sigmoid function.

### 4.3. Training

All parameters $\Theta = (\theta_\alpha, \theta_{sg})$ are trained jointly with the objective

$$\mathcal{L}(\Theta) = \mathcal{L}_{\text{seg}}(\Theta) + \gamma(t) \cdot \mathcal{L}_{\text{aux}}(\theta_\alpha), \tag{11}$$

where $t$ indexes the training epoch. The segmentation loss $\mathcal{L}_{\text{seg}}$ supervises the predicted plume mask, while the auxiliary loss $\mathcal{L}_{\text{aux}}$ aligns the physics-guided score with a classical matched-filter teacher and is annealed by $\gamma(t)$.

**Segmentation loss.** We combine Dice loss with positive-weighted binary cross-entropy.

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}}(\hat{\mathcal{M}}, \mathcal{M}^*) + \beta \cdot \mathcal{L}_{\text{BCE}}(\hat{\mathcal{M}}, \mathcal{M}^*), \tag{12}$$

with $\beta = \min(n_-/n_+, 50)$, where $n_+$ and $n_-$ are the positive and negative pixel counts.

**Curriculum auxiliary supervision.** Since the score map is initially uninformative, we stabilize early training by matching $\hat{\alpha}_i$ against precomputed MAG1C-SAS outputs $\hat{\alpha}_i^{\text{SAS}}$ as follows.

$$\mathcal{L}_{\text{aux}}(\theta_\alpha) = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \rho_i \cdot \left|\hat{\alpha}_i(\ell;\theta_\alpha) - \hat{\alpha}_i^{\text{SAS}}\right|, \tag{13}$$

where $\mathcal{V}$ is the set of valid pixels and $\rho_i = 1 + \lambda_\rho \hat{\alpha}_i^{\text{SAS}}$ with $\lambda_\rho = 10$. The auxiliary weight follows a half-period cosine decay $\gamma(t) = \frac{1}{2}(1+\cos(\pi \min(t/T, 1)))$ with $T = 10$. This induces two phases: in Phase 1 $\mathcal{L}_{\text{aux}}$ drives optimization, aligning the backbone with the matched filter, and in Phase 2 $\mathcal{L}_{\text{seg}}$ takes over, driving the model beyond the MAG1C-SAS teacher.

## 5. Experiments

We evaluate FLAME along four axes. First, we benchmark detection accuracy and efficiency against classical, two-stage, and end-to-end baselines on the STARCOP test set. Second, we qualitatively characterize FLAME against representative baselines across plume regimes. Third, we

*Table 1.* **Quantitative Comparison on the STARCOP test set**. Score columns are reported as mean ± standard deviation. Best results are in bold, second best are underlined. For Pixel FPR, Time, and Params, lower is better. Pixel FPR is reported as $\times 10^{-4}$. Backbones and decoders are abbreviated in the Method column as MobileNetV2 (MNv2), ResNet18 (R18), ConvUp (CU), and ConvUpStride (CUS). Inference time is measured on an RTX 4090 with batch size 1.

| | Method | F1 | IoU | Precision | Recall | Pixel FPR | Time (ms) | Params |
|---|---|---|---|---|---|---|---|---|
| Classical | CEM | 0.178 | 0.098 | 0.114 | 0.399 | 74.0 | 19.6 | – |
| | MF | 0.178 | 0.098 | 0.115 | 0.398 | 74.0 | 21.0 | – |
| | ACE | 0.154 | 0.084 | 0.111 | 0.254 | 49.0 | 32.5 | – |
| | MAG1C-tile | 0.300 | 0.176 | 0.228 | 0.437 | 35.0 | 343.7 | – |
| | MAG1C-SAS | 0.284 | 0.166 | 0.194 | 0.528 | 52.0 | 116.2 | – |
| Two-stage | UNet+MAG1C-tile | 0.477 ± 0.036 | 0.314 ± 0.032 | 0.320 ± 0.034 | <u>0.946</u> ± 0.013 | 49.0 ± 7.0 | 348.1 | 6.60M |
| | UNet+MAG1C-SAS | 0.441 ± 0.020 | 0.283 ± 0.017 | 0.300 ± 0.022 | 0.844 ± 0.042 | 48.0 ± 8.0 | 120.6 | 6.60M |
| | LinkNet+MAG1C-tile | 0.402 ± 0.026 | 0.252 ± 0.021 | 0.256 ± 0.023 | **0.947** ± 0.018 | 67.0 ± 9.0 | 347.5 | <u>0.85M</u> |
| | LinkNet+MAG1C-SAS | 0.397 ± 0.017 | 0.248 ± 0.013 | 0.260 ± 0.017 | 0.846 ± 0.027 | 58.0 ± 7.0 | 120.0 | <u>0.85M</u> |
| End-to-end | UNet (MNv2) | 0.421 ± 0.007 | 0.267 ± 0.006 | 0.337 ± 0.023 | 0.570 ± 0.042 | 27.0 ± 5.0 | <u>4.9</u> | 6.65M |
| | UNet (R18) | 0.309 ± 0.020 | 0.183 ± 0.014 | 0.218 ± 0.015 | 0.535 ± 0.032 | 46.0 ± 3.0 | **4.1** | 14.54M |
| | SegFormer (base) | 0.393 ± 0.055 | 0.246 ± 0.044 | 0.286 ± 0.085 | 0.736 ± 0.129 | 52.0 ± 23.0 | 6.3 | 3.82M |
| | SegFormer (CU) | <u>0.515</u> ± 0.009 | <u>0.347</u> ± 0.008 | <u>0.417</u> ± 0.012 | 0.679 ± 0.059 | <u>23.0</u> ± 3.0 | 7.3 | 4.30M |
| | SegFormer (CUS) | 0.449 ± 0.022 | 0.290 ± 0.019 | 0.314 ± 0.038 | 0.831 ± 0.116 | 46.0 ± 14.0 | 20.3 | 4.30M |
| | EfficientViT (base) | 0.345 ± 0.035 | 0.209 ± 0.026 | 0.222 ± 0.021 | 0.785 ± 0.135 | 66.0 ± 12.0 | 7.1 | 4.81M |
| | EfficientViT (CU) | 0.404 ± 0.003 | 0.253 ± 0.002 | 0.262 ± 0.003 | 0.880 ± 0.009 | 59.0 ± 1.0 | 7.6 | 4.85M |
| | EfficientViT (CUS) | 0.414 ± 0.077 | 0.264 ± 0.060 | 0.279 ± 0.063 | 0.829 ± 0.041 | 55.0 ± 17.0 | 8.2 | 4.85M |
| | FLAME (Ours) | **0.608** ± 0.005 | **0.437** ± 0.005 | **0.651** ± 0.056 | 0.576 ± 0.035 | **8.0** ± 2.0 | 6.2 | **0.78M** |

ablate the physics-guided score to isolate its contribution. Fourth, we assess deployability by profiling inference on hardware representative of onboard satellite platforms.

### 5.1. Experimental Setup

**Dataset.** We evaluate on the STARCOP dataset (Růžička et al., 2023), which pairs AVIRIS-NG hyperspectral tiles from the Permian Basin campaign with pixel-level methane plume annotations curated on $512 \times 512$ chips. We use 72 SWIR bands in the 2122–2488 nm range together with three RGB bands, follow the original train/test split, and report all metrics on the held-out test set of 342 tiles.

**Evaluation Metrics.** We report pixel-level recall, precision, F1, and IoU, together with the pixel-level false positive rate to quantify spurious detections on regions without plumes. Inference time is measured on a single NVIDIA RTX 4090 with batch size 1.

**Implementation Details.** We train FLAME with a curriculum that first aligns the physics-guided score with a matched-filter teacher and then transitions to the segmentation objective, and report mean and standard deviation over 3 random seeds. Full optimizer settings, training schedule, and loss configuration are provided in Appendix B.

**Baselines.** We compare against three categories of methods. Classical matched filters include CEM (Kraut et al., 2005), MF (Manolakis & Shaw, 2002; Funk et al., 2002), ACE (Chang & Heinz, 2000), MAG1C (Foote et al., 2020), and MAG1C-SAS (Herec et al., 2025). Two-stage pipelines combine a matched filter with a segmentation network: the UNet (Ronneberger et al., 2015) variants paired with MAG1C products follow the HyperSTARCOP design introduced in (Růžička et al., 2023), while the LinkNet (Chaurasia & Culurciello, 2017) variants paired with MAG1C-SAS are taken from (Herec et al., 2025). End-to-end models process raw hyperspectral data without a matched filter, and we evaluate SegFormer (Xie et al., 2021) and EfficientViT (Cai et al., 2023) under the HyperspectralViTs adaptations proposed in (Růžička & Markham, 2025), along with UNet (Ronneberger et al., 2015) baselines. Classical matched-filter scores are reproduced from (Herec et al., 2025), while all learned baselines are trained under the same protocol as FLAME.

### 5.2. Comparison with Existing Methods

We compare FLAME against the classical, two-stage, and end-to-end baselines introduced in Section 5.1.

**Detection performance.** FLAME achieves the highest F1 of 0.608 and IoU of 0.437 in Table 1, improving over the strongest two-stage pipeline UNet+MAG1C-tile by 13.1 F1 points while eliminating the matched-filter computation, and over the strongest end-to-end baseline SegFormer with the ConvUp decoder by 9.3 F1 points. Its precision of 0.651, the highest among all methods, indicates substantially cleaner masks than any learned alternative. FLAME's recall of 0.576 is lower than that of the two-stage pipelines, which exceed 0.84, but the high recall of those pipelines is

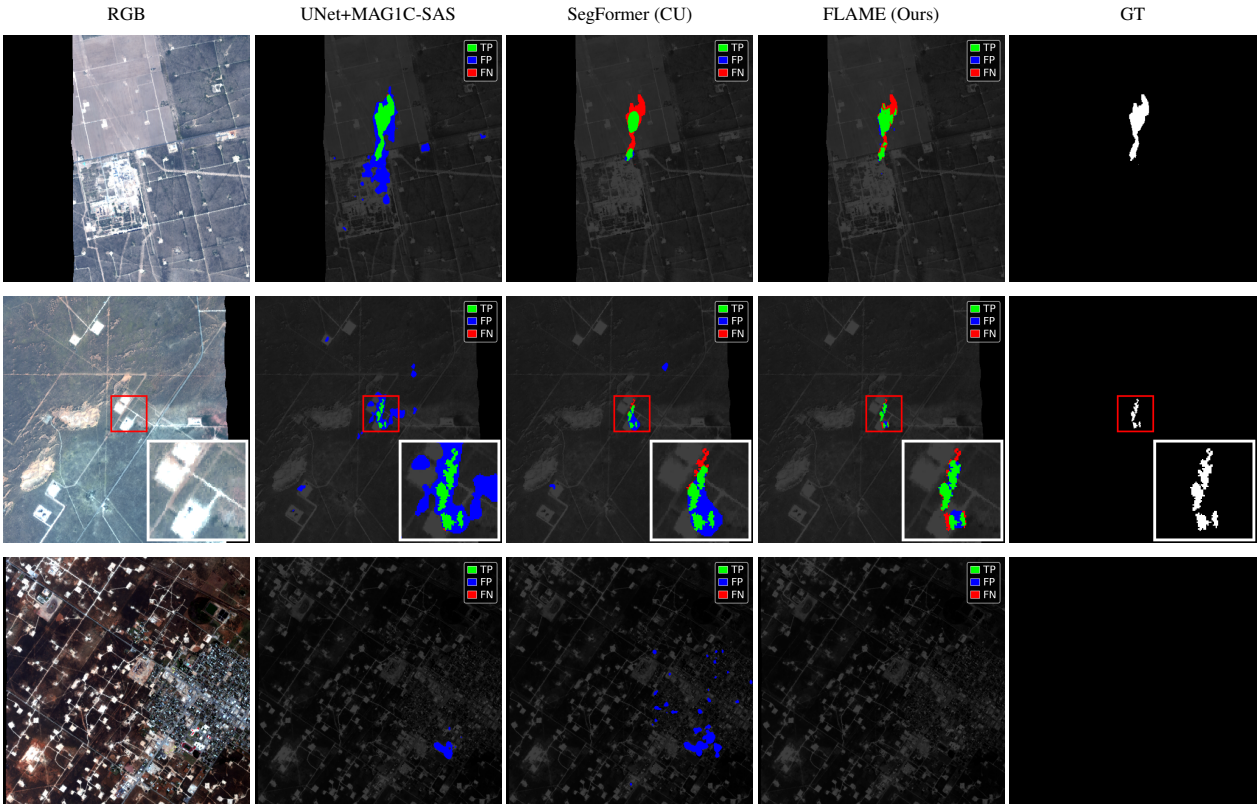


*Figure 4.* **Qualitative Comparison on the STARCOP test set**. Three test tiles representing a strong-plume regime (top), a weak-plume regime (middle), and a plume-free regime (bottom), shown for UNet+MAG1C-SAS, SegFormer with the ConvUp decoder, and FLAME. Prediction panels overlay true positives in green, false positives in blue, and false negatives in red on a darkened RGB backdrop. The ground-truth column shows the binary plume mask.

achieved by predicting overly large plume masks rather than by recovering more true methane pixels, as the qualitative comparison in Figure 4 makes visible.

**False positive behavior.** FLAME's pixel-level false positive rate of $8 \times 10^{-4}$ is the lowest among learned methods, roughly threefold below the next best end-to-end model SegFormer with the ConvUp decoder at $2.3 \times 10^{-3}$ and an order of magnitude below two-stage pipelines, whose rates span $4.8 \times 10^{-3}$ to $6.7 \times 10^{-3}$. This advantage is preserved across plume sizes, with the smallest degradation on tiles containing fewer than 100 plume pixels. The corresponding per-bin F1 in Figure 1 shows that FLAME also leads in F1 across all plume-size bins, so the low FPR reflects cleaner predictions rather than conservative thresholding.

**Efficiency trade-off.** FLAME runs in 6.2 ms per tile on an RTX 4090, competitive to the fastest end-to-end models and roughly 18 times faster than MAG1C-SAS. With 0.78M parameters, it is smaller than LinkNet's 0.85M and roughly an order of magnitude smaller than U-Net variants. A bubble chart visualization of the joint speed, accuracy, and size trade-off is provided in Appendix D.

### 5.3. Qualitative Analysis

Figure 4 compares FLAME against the two-stage baseline UNet+MAG1C-SAS and the end-to-end baseline SegFormer on three tiles spanning a strong plume, a weak plume, and an urban scene without plumes.

In the strong-plume tile in the top row, all three methods localize the plume core, but the failure modes differ. UNet+MAG1C-SAS scatters false positives across the surrounding area, SegFormer under-detects portions of the plume itself, shown as red false negatives, and FLAME produces a tight mask that aligns with the plume boundary.

The contrast widens in the weak-plume tile in the middle row, where the two baselines exhibit opposing failure modes. UNet+MAG1C-SAS spreads false positives across the surrounding region, while SegFormer misses much of the plume. The inset of the middle row makes this gap visible at higher resolution, where FLAME tracks the fine, fragmented structure of the plume that the baselines either oversmooth into surrounding clutter or fail to detect. This pattern is consistent with FLAME's pixel-wise modeling of the background and noise statistics. Classical pipelines

*Table 2.* **Effect of the physics-guided score**. The w/o physics-guide row ablates the score layer, with backbone features passed directly to the segmentation head. The w/ physics-guide row corresponds to the full FLAME model. All metrics are pixel-level on the STARCOP test set from a single training seed.

| Variant | F1 | IoU | Precision | Recall |
|---|---|---|---|---|
| w/o physics-guide | 0.409 | 0.257 | 0.437 | 0.384 |
| w/ physics-guide | **0.603** | **0.432** | **0.596** | **0.611** |

share a single background and covariance across the tile, so pixels whose true background deviates from the tile mean leave residuals that are easily mistaken for weak methane signals. FLAME instead predicts $\hat{\ell}_i^B$ and $\hat{w}_i$ per pixel, so background variations are absorbed into the background model itself and no longer leak into the score for weak plumes. The log-domain formulation further removes the first-order Taylor approximation of the Beer–Lambert law that classical matched filters rely on, ensuring that strong plumes are also treated under the same exact signal model as weak ones. The urban tile without plumes in the bottom row confirms this. Both baselines generate spatially extended false detections on building rooftops and other man-made structures whose spectral signatures are easily confused with methane absorption, while FLAME suppresses the response and yields a clean prediction.

These observations explain the F1 gain and pixel-FPR reduction reported in Table 1, as FLAME suppresses the excessive false positives of two-stage pipelines while avoiding the incomplete plume masks often produced by end-to-end models. Further qualitative results, including per-pixel probability maps for every method, are provided in Appendix H.

### 5.4. Effect of the Physics-Guided Score

The physics-guided score in Eq. (10) is the core inductive bias of FLAME. To isolate its contribution, we compare the full model against an ablation that removes the score layer entirely, so the backbone features feed directly into the segmentation head without the physics signal $\hat{\alpha}_i$.

Table 2 shows that adding the physics-guided score raises F1 from 0.409 to 0.603 and IoU from 0.257 to 0.432. The improvement is larger than the gap between FLAME and any learned baseline in Table 1, indicating that the parameter-free log-domain inner product, rather than the backbone capacity alone, is what drives FLAME's detection performance. Appendix G further decomposes the score into its learned components and analyzes how it is shaped by the training curriculum and refined by the segmentation head.

### 5.5. Onboard Satellite Deployment

We profile FLAME on three NVIDIA Jetson platforms representative of the hardware currently considered for onboard satellite deployment: Jetson Orin NX 16GB, Jetson AGX Orin 64GB, and Jetson AGX Thor. Each platform runs FLAME with identical trained weights on a fixed set of STARCOP tiles in FP32 precision, with the full measurement protocol described in Appendix E.

*Table 3.* **Onboard inference on Jetson platforms**. Time is reported per tile, and TDP is the catalog power cap of each module.

| Platform | Time (ms) | Power (W) | Temp (°C) | TDP (W) |
|---|---|---|---|---|
| Jetson Orin NX | 225.5 | 9.5 | 51.8 | 15 |
| Jetson AGX Orin | 117.6 | 7.0 | 47.5 | 30 |
| Jetson AGX Thor | 36.4 | 59.5 | 41.5 | 120 |

Table 3 reports the measurements. Jetson Orin NX processes one $512 \times 512$ tile in 225.5 ms at 9.5 W, within the 15 W cap of this module. Jetson AGX Orin reduces the latency to 117.6 ms at 7.0 W of actual draw, less than a quarter of the 30 W cap, so the workload leaves substantial compute headroom on this class of module. Jetson AGX Thor reaches 36.4 ms at 59.5 W, roughly half of the 120 W cap. All three platforms thermally equilibrate between 41.5 and 51.8 °C, indicating sustained operation rather than transient peak performance. These power envelopes fall within the budgets typical of small-satellite payload computers, making FLAME directly deployable on the onboard hardware class that current and upcoming hyperspectral missions are expected to carry.

## 6. Conclusion

In this work, we proposed FLAME, a physics-guided neural operator for onboard methane plume detection in hyperspectral imagery. FLAME preserves the log-domain matched-filter structure of methane retrieval while replacing fragile tile-level statistics with pixel-wise log-background and spectral-weight estimates learned by a compact Fourier-based neural operator. By embedding the Beer–Lambert absorption prior directly into a parameter-free score layer, FLAME combines the physical reliability of classical matched filters with the speed and adaptability of learned segmentation models. Experiments on the STARCOP benchmark show that this design improves detection accuracy, substantially reduces false positives, and remains lightweight enough for onboard satellite deployment. These results suggest that physics-guided neural operators provide a practical path toward accurate, efficient, and deployable methane monitoring from future hyperspectral satellite missions.

**Impact Statements** This work targets faster onboard detection of methane point sources from satellite hyperspectral imagery. Since methane is a potent short-lived greenhouse gas driven by a few emitters, locating these sources cuts the delay from emission to response from days to minutes.

## References


Alvarez, R. A., Zavala-Araiza, D., Lyon, D. R., Allen, D. T., Barkley, Z. R., Brandt, A. R., Davis, K. J., Herndon, S. C., Jacob, D. J., Karion, A., et al. Assessment of methane emissions from the us oil and gas supply chain. *Science*, 361(6398):186–188, 2018.

Ayasse, A. K., Dennison, P. E., Foote, M., Thorpe, A. K., Joshi, S., Green, R. O., Duren, R. M., Thompson, D. R., and Roberts, D. A. Methane mapping with future satellite imaging spectrometers. *Remote Sensing*, 11(24):3054, 2019.

Cai, H., Li, J., Hu, M., Gan, C., and Han, S. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17302–17313, 2023.

Cawse-Nicholson, K., Townsend, P. A., Schimel, D., Assiri, A. M., Blake, P. L., Buongiorno, M. F., Campbell, P., Carmon, N., Casey, K. A., Correa-Pabón, R. E., et al. Nasa's surface biology and geology designated observable: A perspective on surface imaging algorithms. *Remote sensing of environment*, 257:112349, 2021.

Chang, C.-I. and Heinz, D. C. Constrained subpixel target detection for remotely sensed imagery. *IEEE transactions on geoscience and remote sensing*, 38(3):1144–1159, 2000.

Chaurasia, A. and Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pp. 1–4. IEEE, 2017.

Duren, R. M., Thorpe, A. K., Foster, K. T., Rafiq, T., Hopkins, F. M., Yadav, V., Bue, B. D., Thompson, D. R., Conley, S., Colombi, N. K., et al. California's methane super-emitters. *Nature*, 575(7781):180–184, 2019.

Esposito, M., Conticello, S. S., Pastena, M., and Domínguez, B. C. In-orbit demonstration of artificial intelligence applied to hyperspectral and thermal sensing from space. In *CubeSats and SmallSats for remote sensing III*, volume 11131, pp. 88–96. SPIE, 2019.

Etminan, M., Myhre, G., Highwood, E. J., and Shine, K. P. Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters*, 43(24):12–614, 2016.

Foote, M. D., Dennison, P. E., Thorpe, A. K., Thompson, D. R., Jongaramrungruang, S., Frankenberg, C., and Joshi, S. C. Fast and accurate retrieval of methane concentration from imaging spectrometer data using sparsity prior. *IEEE Transactions on Geoscience and Remote Sensing*, 58(9):6480–6492, 2020.

Funk, C. C., Theiler, J., Roberts, D. A., and Borel, C. C. Clustering to improve matched filter detection of weak gas plumes in hyperspectral thermal imagery. *IEEE transactions on geoscience and remote sensing*, 39(7):1410–1420, 2002.

Furano, G., Meoni, G., Dunne, A., Moloney, D., Ferlet-Cavrois, V., Tavoularis, A., Byrne, J., Buckley, L., Psarakis, M., Voss, K.-O., et al. Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. *IEEE Aerospace and Electronic Systems Magazine*, 35(12):44–56, 2020.

Giuffrida, G., Fanucci, L., Meoni, G., Batič, M., Buckley, L., Dunne, A., Van Dijk, C., Esposito, M., Hefele, J., Vercruyssen, N., et al. The $\phi$-sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Herec, J., Růžička, V., and Pitoňák, R. Optimizing methane detection on board satellites: Speed, accuracy, and low-power solutions for resource-constrained hardware. In *2025 European Data Handling & Data Processing Conference (EDHPC)*, pp. 1–9. IEEE, 2025.

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Kraut, S., Scharf, L. L., and Butler, R. W. The adaptive coherence estimator: A uniformly most-powerful-invariant adaptive detection statistic. *IEEE Transactions on Signal Processing*, 53(2):427–438, 2005. doi: 10.1109/TSP.2004.840823.

Kumar, S., Arevalo, I., Iftekhar, A., and Manjunath, B. Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17609–17618, 2023.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Langer, D. D., Orlandić, M., Bakken, S., Birkeland, R., Garrett, J. L., Johansen, T. A., and Sørensen, A. J. Robust and reconfigurable on-board processing for a hyperspectral

imaging small satellite. *Remote Sensing*, 15(15):3756, 2023.

Lauvaux, T., Giron, C., Mazzolini, M., d'Aspremont, A., Duren, R., Cusworth, D., Shindell, D., and Ciais, P. Global assessment of oil and gas methane ultra-emitters. *Science*, 375(6580):557–561, 2022.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Manolakis, D. and Shaw, G. Detection algorithms for hyperspectral imaging applications. *IEEE signal processing magazine*, 19(1):29–43, 2002.

Manolakis, D., Truslow, E., Pieper, M., Cooley, T., and Brueggeman, M. Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms. *IEEE Signal Processing Magazine*, 31(1):24–33, 2013.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., et al. *Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2021.

Nieke, J. and Rast, M. Towards the copernicus hyperspectral imaging mission for the environment (chime). In *Igarss 2018-2018 ieee international geoscience and remote sensing symposium*, pp. 157–159. IEEE, 2018.

Ocko, I. B., Sun, T., Shindell, D., Oppenheimer, M., Hristov, A. N., Pacala, S. W., Mauzerall, D. L., Xu, Y., and Hamburg, S. P. Acting rapidly to deploy readily available methane mitigation measures by sector can immediately slow global warming. *Environmental Research Letters*, 16(5):054042, 2021.

Parr, J., Acciarini, G., Bridges, C., Mateo-Garcia, G., Portales-Julia, E., Purcell, C., Ruzicka, V., Spalding, A., and Veitch-Michaelis, J. Live twinning: A vision of ml enabled assets in leo for rapid response to natural catastrophes. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 685–688. IEEE, 2024.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Růžička, V. and Markham, A. Hyperspectralvits: General hyperspectral models for on-board remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Růžička, V., Vaughan, A., De Martini, D., Fulton, J., Salvatelli, V., Bridges, C., Mateo-Garcia, G., and Zantedeschi, V. Ravæn: unsupervised change detection of extreme events using ml on-board satellites. *Scientific reports*, 12(1):16939, 2022.

Růžička, V., Mateo-Garcia, G., Gómez-Chova, L., Vaughan, A., Guanter, L., and Markham, A. Semantic segmentation of methane plumes with hyperspectral machine learning models. *Scientific Reports*, 13(1):19999, 2023.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., et al. The global methane budget 2000–2017. *Earth System Science Data Discussions*, 2019:1–136, 2019.

Shindell, D., Kuylenstierna, J. C., Vignati, E., Van Dingenen, R., Amann, M., Klimont, Z., Anenberg, S. C., Muller, N., Janssens-Maenhout, G., Raes, F., et al. Simultaneously mitigating near-term climate change and improving human health and food security. *Science*, 335 (6065):183–189, 2012.

Thompson, D., Leifer, I., Bovensmann, H., Eastwood, M., Fladeland, M., Frankenberg, C., Gerilowski, K., Green, R., Kratwurst, S., Krings, T., et al. Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane. *Atmospheric Measurement Techniques*, 8(10):4383–4397, 2015.

Thompson, D., Thorpe, A., Frankenberg, C., Green, R., Duren, R., Guanter, L., Hollstein, A., Middleton, E., Ong, L., and Ungar, S. Space-based remote imaging spectroscopy of the aliso canyon ch4 superemitter. *Geophysical Research Letters*, 43(12):6571–6578, 2016.

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., and Benson, S. M. U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.

Wong, K. C., Klein, L., Da Silva, A. F., Wang, H., Singh, J., and Syeda-Mahmood, T. Image-based soil organic carbon remote sensing from satellite images with fourier neural operator and structural similarity. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1305–1308. IEEE, 2023.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Yang, Q., Hernandez-Garcia, A., Harder, P., Ramesh, V., Sattigeri, P., Szwarcman, D., Watson, C. D., and Rolnick, D. Fourier neural operators for arbitrary resolution climate data downscaling. *Journal of Machine Learning Research*, 25(420):1–30, 2024.

Yost, B. and Weston, S. State-of-the-art small spacecraft technology. Technical report, 2024.

Ziaja, M., Bosowski, P., Myller, M., Gajoch, G., Gumiela, M., Protich, J., Borda, K., Jayaraman, D., Dividino, R., and Nalepa, J. Benchmarking deep learning for on-board space applications. *Remote Sensing*, 13(19):3981, 2021.

## A. FLAME Generalizes the Log-Domain Matched Filter

In Section 4 we claimed that FLAME strictly contains the classical log-domain matched filter as a special case of its hypothesis class. This appendix formalizes the claim and exhibits an explicit parameter setting that realizes the reduction.

**Proposition A.1** (FLAME generalizes the log-domain matched filter)**.** *Fix $\bar{\ell} \in \mathbb{R}^p$ and a vector $\omega^2$ with strictly positive entries, and define the diagonal log-noise covariance $C = \mathrm{diag}(\omega^2)$. There exists a score-level parameter setting $\theta_\alpha = (\theta_\Phi, \theta_{bg}, \theta_{sw})$ such that, for every log-radiance tile $\ell$ and every pixel $i$,*

$$\hat{\alpha}_i(\ell; \theta_\alpha) \;=\; (\ell_i - \bar{\ell})^\top C^{-1} s, \tag{14}$$

*which is the detection-relevant inner product in the numerator of Eq. (6) under a tile-shared background $\ell_i^B \equiv \bar{\ell}$ and a pixel-independent diagonal covariance $C_i \equiv C$.*

*Proof.* We construct an explicit parameter setting that realizes the reduction. Set

$$W_{bg} = 0, \quad b_{bg} = \bar{\ell}, \quad W_{sw} = 0, \quad b_{sw} = \log\big(\exp(1/\omega^2) - \mathbf{1}\big), \tag{15}$$

where all operations on $\omega^2$ are element-wise. The choice of $b_{sw}$ is well-defined because softplus is a continuous, strictly increasing bijection from $\mathbb{R}$ onto $(0, \infty)$ and $\omega^2$ has strictly positive entries. This setting yields

$$\hat{\ell}_i^B = \bar{\ell}, \qquad \hat{w}_i = 1/\omega^2, \tag{16}$$

for every pixel. Any $\theta_\Phi$ is admissible because both heads are independent of $z_i$. Substituting Eq. (16) into Eq. (10) gives

$$\hat{\alpha}_i(\ell; \theta_\alpha) = (\ell_i - \bar{\ell})^\top \big((1/\omega^2) \odot s\big), \tag{17}$$

and since $C^{-1} s = (1/\omega^2) \odot s$, this is Eq. (14). □

**Strict containment.** The inclusion is strict because any $W_{bg} \neq 0$ produces a log-background that varies across pixels and cannot be reproduced by any tile-shared $\bar{\ell}$, and any $W_{sw} \neq 0$ yields a pixel-dependent spectral weight that no single diagonal $C$ can match. Together with the single-pass evaluation of the score layer in Eq. (10), these two extensions address the three limitations identified in Section 3.1 in a one-to-one manner, with a learned $\hat{\ell}_i^B$ replacing the tile-shared background, a learned $\hat{w}_i$ replacing the pixel-independent covariance, and a single forward pass replacing the iterative refinement of Eq. (4).

**Scalar normalization and exactness.** The full estimator of Eq. (6) differs from Eq. (14) by the data-independent scalar $1/(s^\top C^{-1} s)$, which rescales the detection map uniformly and does not affect which pixels are detected. In FLAME this global rescaling is applied through the fixed $\tau$-normalization of Section 4.2. Crucially, Proposition A.1 rests on the exact identity in Eq. (5) rather than on the first-order linearization in Eq. (2), so the matched-filter behavior FLAME inherits at $\theta_\alpha$ remains unbiased across the full range of plume intensities, including the strong ultra-emitters for which the radiance-domain linearization is known to underestimate $\alpha_i$.

## B. Dataset and Implementation Details

**Dataset.** We followed the train/test partition released with STARCOP (Růžička et al., 2023), using 3,425 tiles for training and 342 tiles for held-out evaluation, with no separate validation split. Plume tiles and tiles without plumes are nearly balanced at the tile level, but the positive class is sparse at the pixel level, with $0.26\%$ of training pixels carrying the plume label, rising to $0.52\%$ within tiles that contain plumes. The test split follows the same regime. All tiles are stored as $512 \times 512$ rasters, and AVIRIS-NG scene crops smaller than this footprint are zero-padded, with the padded margins excluded from training and evaluation through a per-pixel valid mask. Table 4 summarizes the tile- and pixel-level class balance of each split.

*Table 4.* Tile- and pixel-level statistics for the STARCOP partition used in this work. The Plume % column reports the fraction of pixels labelled positive among pixels that are not padded, and the value in parentheses restricts this fraction to tiles that contain plumes.

| Split | Tiles | Plume tiles | No-plume tiles | Plume % (within plume tiles) |
|---|---|---|---|---|
| Train | 3,425 | 1,712 | 1,713 | 0.26% (0.52%) |
| Test | 342 | 166 | 176 | 0.21% (0.44%) |

**Model architecture.** The FLAME backbone is a U-FNO (Wen et al., 2022) with hidden width $d = 14$ and $L_F = 3$ FNO blocks followed by $L_U = 3$ U-FNO blocks. Each spectral mixing layer retains $m_1 = m_2 = 12$ Fourier modes per spatial axis. Table 5 summarizes these settings.

*Table 5.* Architectural hyperparameters of FLAME.

| Hyperparameter | Value |
|---|---|
| Backbone hidden width $d$ | 14 |
| Fourier modes $m_1, m_2$ | 12, 12 |
| FNO blocks $L_F$ | 3 |
| U-FNO blocks $L_U$ | 3 |

**Training.** We trained FLAME for 50 epochs with AdamW (Loshchilov & Hutter, 2017), using initial learning rate $\eta_0 = 2 \times 10^{-3}$ and weight decay $10^{-4}$, under a cosine schedule decaying to $\eta_{\min} = 10^{-6}$, with gradient clipping to $\ell_2$-norm 1.0 and mixed-precision training. The global batch size was 24 and inputs were processed at the native $512 \times 512$ resolution. We applied random horizontal and vertical flips and uniform $\{0, 90, 180, 270\}^\circ$ rotations jointly to the input maps, the binary label, and the cached MAG1C-SAS teacher map, and the valid mask $m$ was broadcast against every loss term and metric counter so that padded margins did not contribute to gradient or evaluation. The BCE component of the segmentation loss is reweighted per minibatch with $\beta = \min(N_-/N_+, 50)$, where $N_+$ and $N_-$ are the positive and negative pixel counts under the valid mask, and the Dice term is masked by $m$ before summing. We set the score normalization constant to $\tau = 1750$. For FLAME and every learned baseline we used a fixed decision rule throughout training and inference, $\sigma(\hat{y}) > 0.5$ followed by a $3 \times 3$ cross-shaped morphological opening, without per-model threshold tuning. All learned models were trained on three NVIDIA RTX 4090 GPUs.

**Baselines.** We compare against three categories of methods. The classical matched filters use the canonical method-specific thresholds of (Herec et al., 2025), with $0.004$ for CEM (Kraut et al., 2005) and MF (Manolakis & Shaw, 2002; Funk et al., 2002), $0.03$ for ACE (Chang & Heinz, 2000), and $300$ for both MAG1C (Foote et al., 2020) and MAG1C-SAS (Herec et al., 2025). The two-stage pipelines, namely UNet paired with MAG1C and LinkNet paired with MAG1C-SAS, retain the HyperSTARCOP normalization protocol of (Herec et al., 2025). The end-to-end baselines are UNet with a MobileNet-V2 encoder (Sandler et al., 2018), SegFormer-B0 (Xie et al., 2021), and EfficientViT-B1 (Cai et al., 2023). All learned baselines share FLAME's training data, augmentation pipeline, valid-pixel mask, decision threshold, and three-seed protocol, with each retaining the optimizer recipe and learning-rate scale of its underlying architecture. Table 6 lists the resulting settings.

**Identical evaluation.** Every row of Table 1 is evaluated on the same 342-tile STARCOP test partition described in Table 4, with the same pixel-level metric definitions for recall, precision, F1, IoU, and FPR computed against all valid background pixels, and with the same per-row threshold rule, namely the canonical thresholds of (Herec et al., 2025) for classical filters

*Table 6.* Hyperparameters of the end-to-end learned baselines trained in our environment. All models are initialized randomly and share FLAME's training recipe except for the initial learning rate.

| Method | Backbone | Initial LR | Epochs | Batch |
|---|---|---|---|---|
| UNet | MobileNet-V2 | $2.0 \times 10^{-3}$ | 50 | 24 |
| SegFormer | MiT-B0 | $1.2 \times 10^{-4}$ | 50 | 24 |
| EfficientViT | EfficientViT-B1 | $1.0 \times 10^{-3}$ | 50 | 24 |
| FLAME (ours) | UFNO + dual heads | $2.0 \times 10^{-3}$ | 50 | 24 |

and the $\sigma > 0.5$ rule followed by morphological opening for segmentation networks. As a self-consistency check, we verified that our local re-runs of the classical filters reproduce the reference F1 values of (Herec et al., 2025) within $\pm 0.005$.

## C. Neural Operator Backbone Comparison

*Table 7.* Neural operator backbone comparison within the FLAME physics-head framework. The physics heads, segmentation head, and training schedule are kept fixed, so this comparison isolates the effect of the operator backbone from the physics-guided formulation. Pixel FPR is reported as $\times 10^{-4}$.

| Variant | F1 | IoU | Precision | Recall | Pixel FPR | Params (M) | Time (ms) |
|---|---|---|---|---|---|---|---|
| FLAME (UFNO) | **0.603** | **0.432** | 0.596 | 0.611 | 10.0 | 0.78 | 6.2 |
| FNO | 0.531 | 0.361 | 0.604 | 0.474 | 7.0 | 0.79 | **5.6** |
| F-FNO | 0.356 | 0.216 | 0.371 | 0.342 | 14.0 | 2.39 | 18.1 |
| Tucker-FNO | 0.571 | 0.399 | 0.506 | **0.654** | 15.0 | 2.16 | 16.9 |
| UNO | 0.263 | 0.151 | 0.240 | 0.291 | 22.0 | 2.54 | 8.2 |
| WNO | 0.517 | 0.349 | **0.629** | 0.440 | **6.0** | **0.77** | 26.7 |

**Backbone comparison.** We compare different neural-operator backbones while keeping the physics heads, segmentation head, and training schedule fixed. As shown in Table 7, UFNO leads in F1 and IoU while remaining among the most parameter-efficient and fastest options. WNO matches its parameter count and achieves the highest precision and lowest pixel FPR, but its inference time is roughly four times higher. FNO is the fastest at the cost of seven F1 points, and the larger backbones, namely F-FNO, Tucker-FNO, and UNO, increase parameter count and latency without a corresponding improvement in detection performance. UFNO therefore offers the best joint trade-off across detection accuracy, parameter count, and inference speed, which is why we adopt it as the FLAME backbone.

**Backbone configurations.** Table 8 lists the architectural hyperparameters of each backbone evaluated in Table 7. All variants share the FLAME physics heads, segmentation head, training schedule, and input pipeline, so the metrics in Table 7 isolate the effect of the backbone design. For FLAME (UFNO), the value $3 + 3$ in the Layers column denotes three FNO blocks followed by three U-FNO blocks.

*Table 8.* Architectural configuration of each neural operator backbone in Table 7.

| Variant | Layers | Width | Modes / sub-bands | Notes |
|---|---|---|---|---|
| FLAME (UFNO) | $3 + 3$ | 14 | $m_1 = m_2 = 12$ | SE block and mini-UNet local path |
| FNO | 4 | 14 | $m_1 = m_2 = 32$ | canonical formulation |
| F-FNO | 8 | 48 | $m = 32$ per axis | factorized 1D spectral convolution |
| Tucker-FNO | 4 | rank 64 | $m = 32$ per axis | Tucker decomposition |
| UNO | 7 | $w = 8, f = 0.75$ | $m = [24, 14, 8, 8, 8, 14, 22]$ | U-shape channel schedule |
| WNO | 3 | 14 | db6, level 6 | DWT sub-band weighting |

## D. Additional Quantitative Results

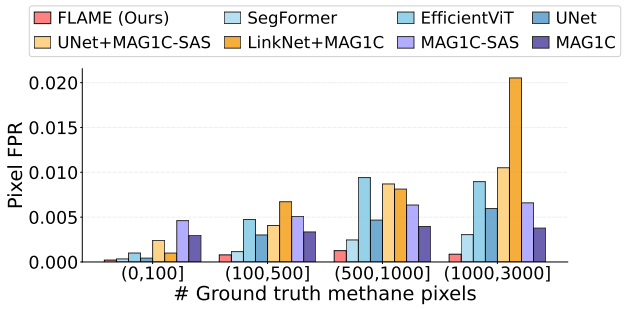


*Figure 5.* Pixel-level false positive rate across test tiles grouped by the number of ground-truth methane pixels.

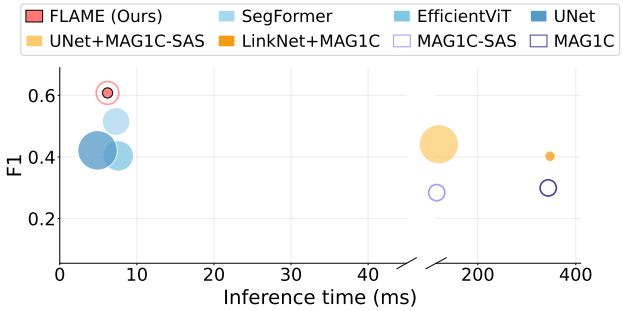


*Figure 6.* Comparison of speed, accuracy, and size across the evaluated methods. Marker size indicates the number of trainable parameters.

**False-positive behavior across plume sizes.** The plume-size analysis in the main text reports F1 scores across bins defined by the number of ground-truth methane pixels. Here we provide the complementary false-positive analysis. This analysis is important because methane plume detection is highly imbalanced at the pixel level, and an apparent gain in F1 can be caused by overly large predicted masks that improve recall while increasing false alarms over background regions. Figure 5 reports the pixel-level false positive rate across the same plume-size bins used in the main text. FLAME maintains a low false-positive rate across plume regimes, including larger plume cases where several segmentation baselines tend to over-extend the plume mask into surrounding background. This behavior indicates that the physics-guided score improves plume localization without simply expanding predictions around candidate plume regions. The result also supports the qualitative observations in Appendix H, where FLAME produces more compact predictions around plume boundaries and fewer activations on scenes without plumes.

**Speed, accuracy, and size trade-off.** Figure 6 compares the evaluated methods in terms of inference latency, detection accuracy, and parameter count. This analysis complements the main hardware profiling results by showing where each method lies in the broader accuracy-efficiency landscape. Classical matched-filter methods have no trainable parameters but require expensive statistical estimation, while two-stage pipelines inherit the cost of their methane enhancement product. End-to-end learned models are fast, but they do not explicitly preserve the methane absorption prior. FLAME occupies a favorable trade-off region by combining single-pass inference with a compact physics-guided architecture.

## E. Onboard Satellite Deployment Details

This appendix expands on the onboard satellite deployment results in Section 5.5.

**Hardware.** Table 9 reports the configuration of the three NVIDIA Jetson modules used in Section 5.5. The three modules span the compute and power tiers that current and next-generation onboard satellite payload computers are expected to provide. Jetson Orin NX 16GB represents the low-power tier suitable for cubesat-class platforms, Jetson AGX Orin 64GB represents the production-grade mid-range tier suitable for current onboard inference, and Jetson AGX Thor represents the next-generation high-performance tier. The TDP cap column reports the configured power budget under which the latency, power, and temperature measurements in Table 3 were taken.

*Table 9.* Hardware configuration of the NVIDIA Jetson platforms used in Section 5.5.

| Platform | CPU | GPU | Memory | TDP cap |
|---|---|---|---|---|
| Jetson Orin NX | 8× A78AE @ 2.0 GHz | Ampere | 16 GB | 15 W |
| Jetson AGX Orin | 12× A78AE @ 2.2 GHz | Ampere | 64 GB | 30 W |
| Jetson AGX Thor | 14× Neoverse-V3AE @ 2.6 GHz | Blackwell | 128 GB | 120 W |

**Measurement protocol.** All measurements use the same trained FLAME weights and a fixed batch of five STARCOP test tiles. For latency, we issue 50 untimed warm-up forwards and then 200 timed repetitions over the five-tile set per platform, and report the mean per-tile wall-clock time. For power and temperature, on-device sensors are sampled every 100 ms throughout the timed window, and we report mean and peak values per power rail and thermal zone. The values in Table 3 correspond to the total module power and to the GPU thermal zone. Each platform's TDP cap is set to the value listed in Table 9 before profiling.

## F. Limitations and Future Work

**Single-dataset evaluation.** All results are reported on the STARCOP benchmark, following the evaluation protocol of MAG1C-SAS (Herec et al., 2025). The physics-guided score in Eq. (10) depends only on the methane absorption spectrum and the log-radiance field, both of which are sensor-specific but conceptually transferable, so we expect the architecture to adapt to other SWIR spectrometers given matched training data.

**Absence of in-orbit validation.** The onboard satellite deployment characterization in Section 5.5 is performed on NVIDIA Jetson platforms under laboratory conditions. Ground-based profiling does not capture the environmental stressors encountered in low Earth orbit, including total ionizing dose, single event effects, and thermal cycling induced by repeated eclipse transitions. The vacuum environment also eliminates atmospheric convection, so in-orbit heat dissipation must rely entirely on conduction to the spacecraft thermal control system and radiative emission to deep space, which can shift the steady-state operating temperatures reported in Table 3. Radiation-induced bit flips can degrade neural network inference in ways that depend on weight encoding and arithmetic precision, and the FP32 PyTorch configuration we profile has not been evaluated under such conditions. Validation on radiation-tolerant payload computers, and ultimately an in-orbit demonstration, is left for future work.

## G. Analyses of the Physics-Guided Score

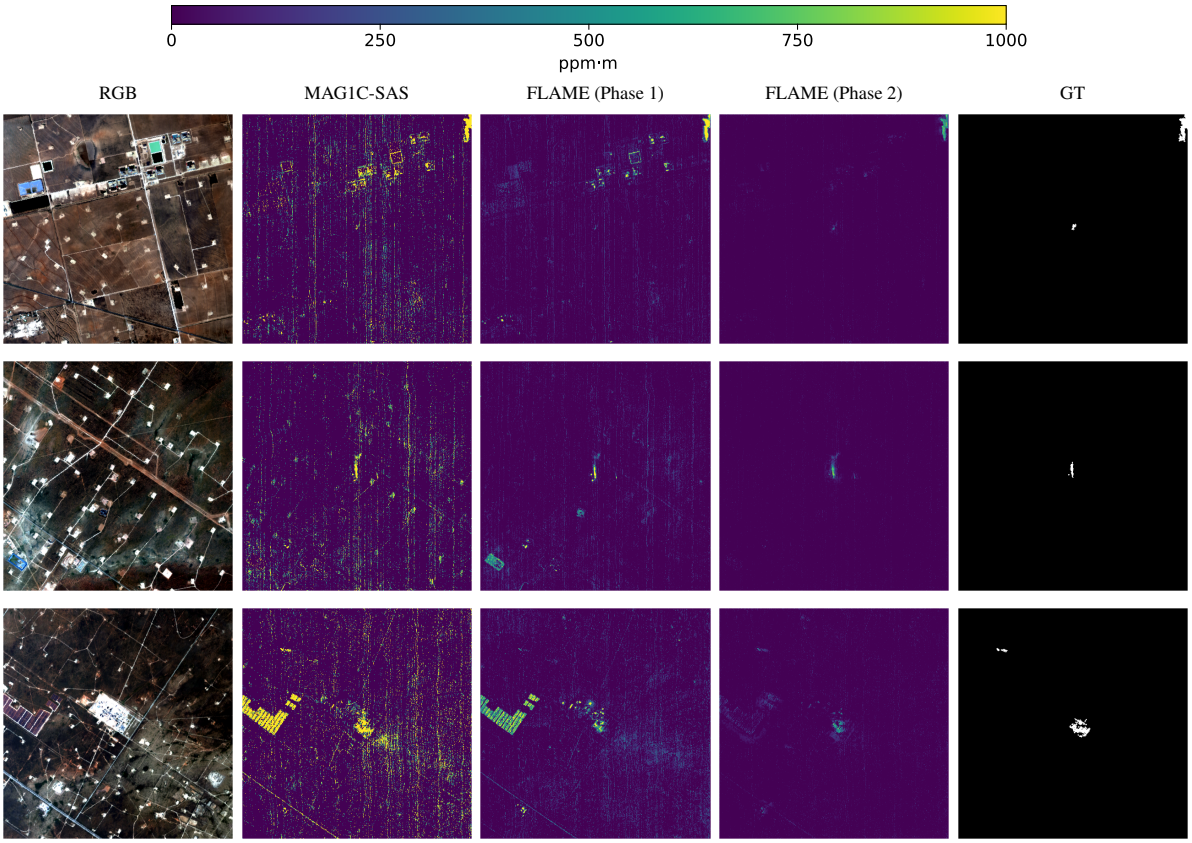


*Figure 7.* Evolution of the FLAME physics score across the two-phase training curriculum described in the main text. Columns show the RGB tile, the MAG1C-SAS reference, the FLAME physics score after Phase 1, the FLAME physics score after Phase 2, and the ground-truth plume mask.

**Curriculum evolution.** As described in the main text, the curriculum first aligns the physics score with the MAG1C-SAS teacher and then shifts the optimization toward the segmentation objective. This design gives the score layer a stable physics-aligned initialization while allowing the final model to depart from the teacher when ground-truth supervision provides a better plume boundary. Figure 7 visualizes this transition, with the Phase 1 and Phase 2 columns corresponding to the model state at epoch $T = 10$, when $\gamma(t)$ first reaches zero, and at the final epoch 50, respectively. The Phase 1 physics score follows the MAG1C-SAS response, while the Phase 2 physics score becomes more selective around annotated plume regions. This selectivity matters because MAG1C-SAS responds to a wide range of albedo variations and surface confounders, and the Phase 2 physics score suppresses these responses while preserving the plume signal, which directly contributes to the false-positive reduction reported in the main text.

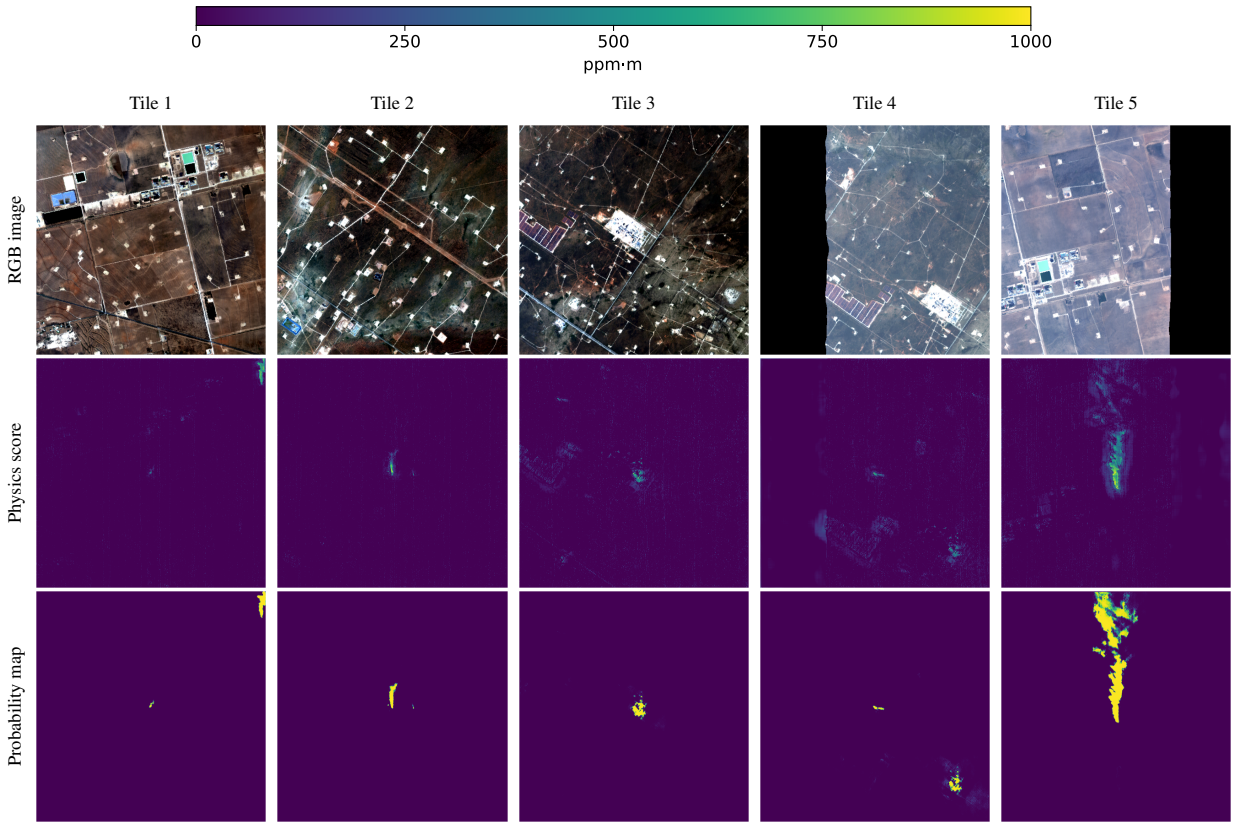


*Figure 8.* Comparison between the FLAME physics score and the probability map produced by the segmentation head. Each column corresponds to a single test tile.

**Score-to-mask refinement.** The physics score is an intermediate methane evidence map, while the segmentation head converts it into the probability map that feeds the final binary mask through thresholding. Figure 8 compares the physics score and the probability map on the same tiles. The segmentation head sharpens the response and calibrates the probability while preserving the spatial structure of the physics score. This indicates that the location and shape of the final binary mask are already determined at the physics score stage, and that the segmentation head acts as a refinement step rather than an independent decision module, which is consistent with the score-layer ablation in the main text.

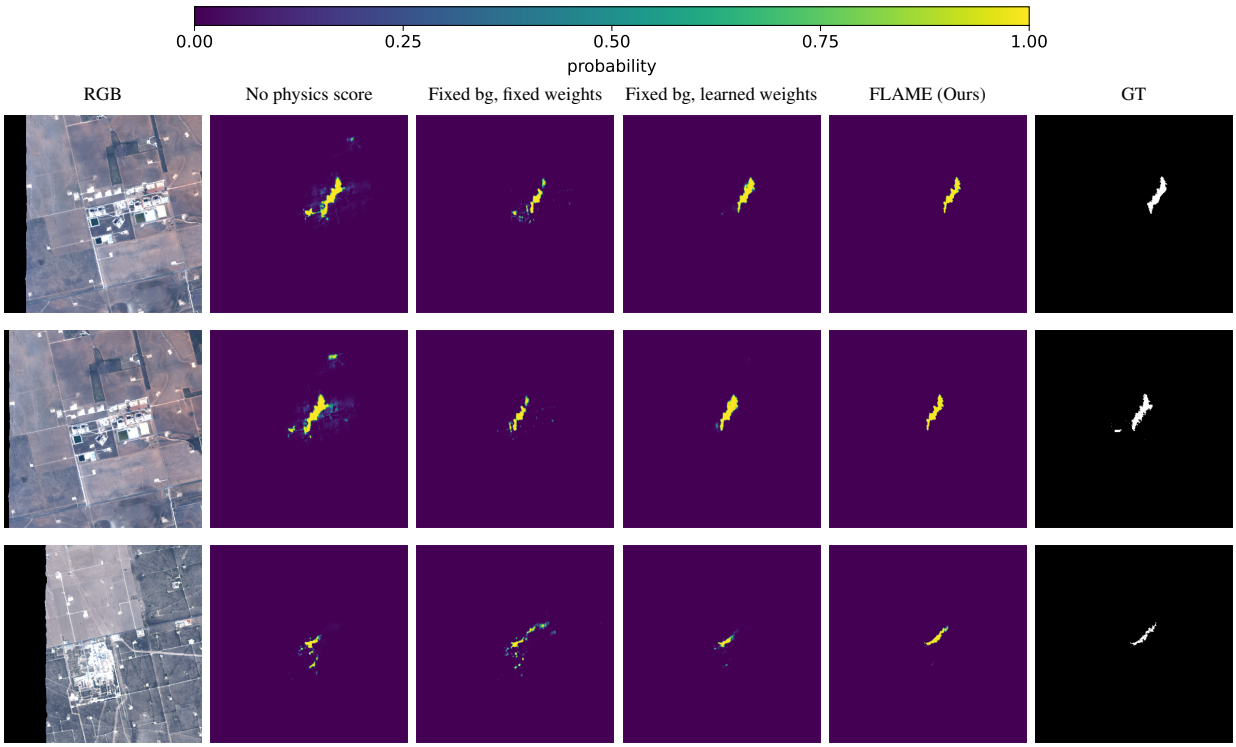


*Figure 9.* Probability maps produced by physics-integration variants of FLAME. Columns show RGB, no physics score, fixed background with fixed spectral weights, fixed background with learned spectral weights, full FLAME, and the ground-truth plume mask, where bg denotes background.

*Table 10.* Component-wise ablation of the physics score. The full FLAME model learns both the pixel-wise log-background and the pixel-wise spectral weighting, while the variants progressively remove these learned components.

| Variant | Recall | Precision | F1 | IoU |
|---|---|---|---|---|
| No physics score | 0.3841 | 0.4367 | 0.4087 | 0.2568 |
| Fixed background, fixed weights | 0.4340 | 0.5543 | 0.4868 | 0.3217 |
| Fixed background, learned weights | 0.4988 | **0.6702** | 0.5720 | 0.4005 |
| FLAME (Ours) | **0.6110** | 0.5957 | **0.6032** | **0.4319** |

**Component-wise ablation.** The score-layer ablation in the main text compares FLAME with and without the physics score. Here we further decompose the physics score itself to isolate the contribution of each learned physical field. We analyze four variants. The first variant uses no physics score and trains the neural backbone as a segmentation model, which corresponds to the w/o physics-guide condition in the main text. The second variant uses a fixed background and fixed spectral weights, recovering a closed-form matched-filter score. The third variant fixes the background but learns the spectral weights, isolating the contribution of pixel-adaptive spectral weighting. The fourth variant learns both the pixel-wise log-background and the pixel-wise spectral weights, which is the full FLAME model and corresponds to the w/ physics-guide condition in the main text. Table 10 reports the corresponding detection metrics, and Figure 9 shows the probability maps. Each added component yields a clear F1 gain. Introducing a physics score lifts F1 from 0.4087 to 0.4868, learning the spectral weights raises it to 0.5720, and additionally learning the log-background reaches 0.6032. The largest jump comes from learning the spectral weights, but the background head is still needed to recover the recall lost in the third variant, where precision is high but plume coverage remains incomplete. This pattern supports the use of both learned physical fields rather than either component alone.

## H. Additional Qualitative Results

This appendix complements the qualitative analysis in Section 5.3.

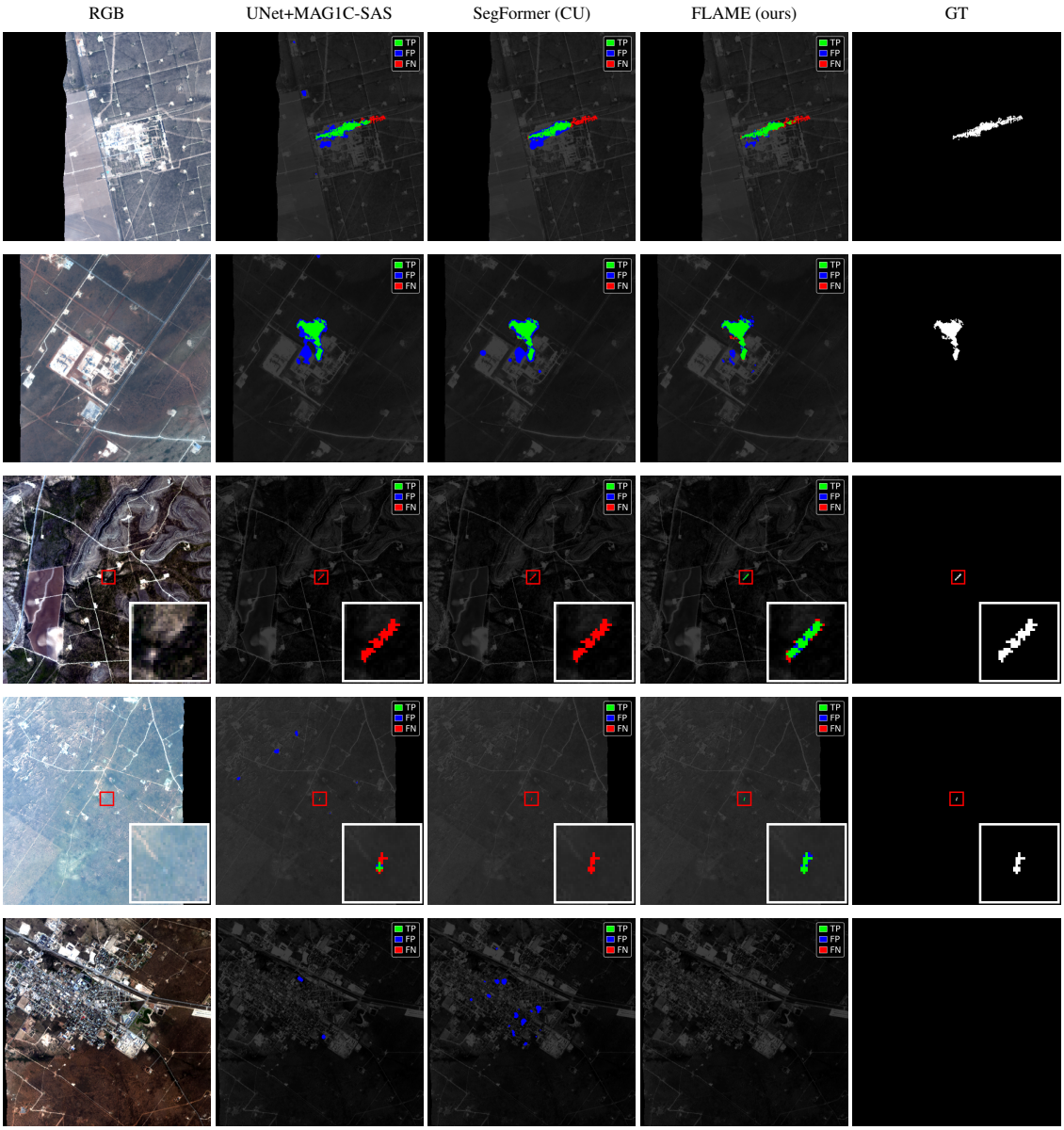


*Figure 10.* Additional qualitative comparison on STARCOP test tiles. Columns show RGB, UNet+MAG1C-SAS, SegFormer with the ConvUp decoder, FLAME, and the ground-truth mask. Prediction panels overlay true positives in green, false positives in blue, and false negatives in red on a darkened RGB backdrop. From top to bottom, rows show two strong-plume tiles, a weak-plume tile with a zoomed inset of the plume region, an over-extension case, and a plume-free urban tile.

**Qualitative Comparison** Across the regimes shown in Figure 10, the dominant failure mode of UNet+MAG1C-SAS is over-prediction around the plume, visible as broad blue halos that extend the predicted mask beyond the annotated region; SegFormer with the ConvUp decoder either misses parts of the plume body, leaving red false negatives along the plume, or scatters small detections in plume-free areas. FLAME produces masks that remain close to the annotated plume in the strong and weak regimes, recovers the elongated plume morphology in the zoomed weak-plume tile, and stays nearly empty on the plume-free urban scene where surface clutter triggers detections in the other models.

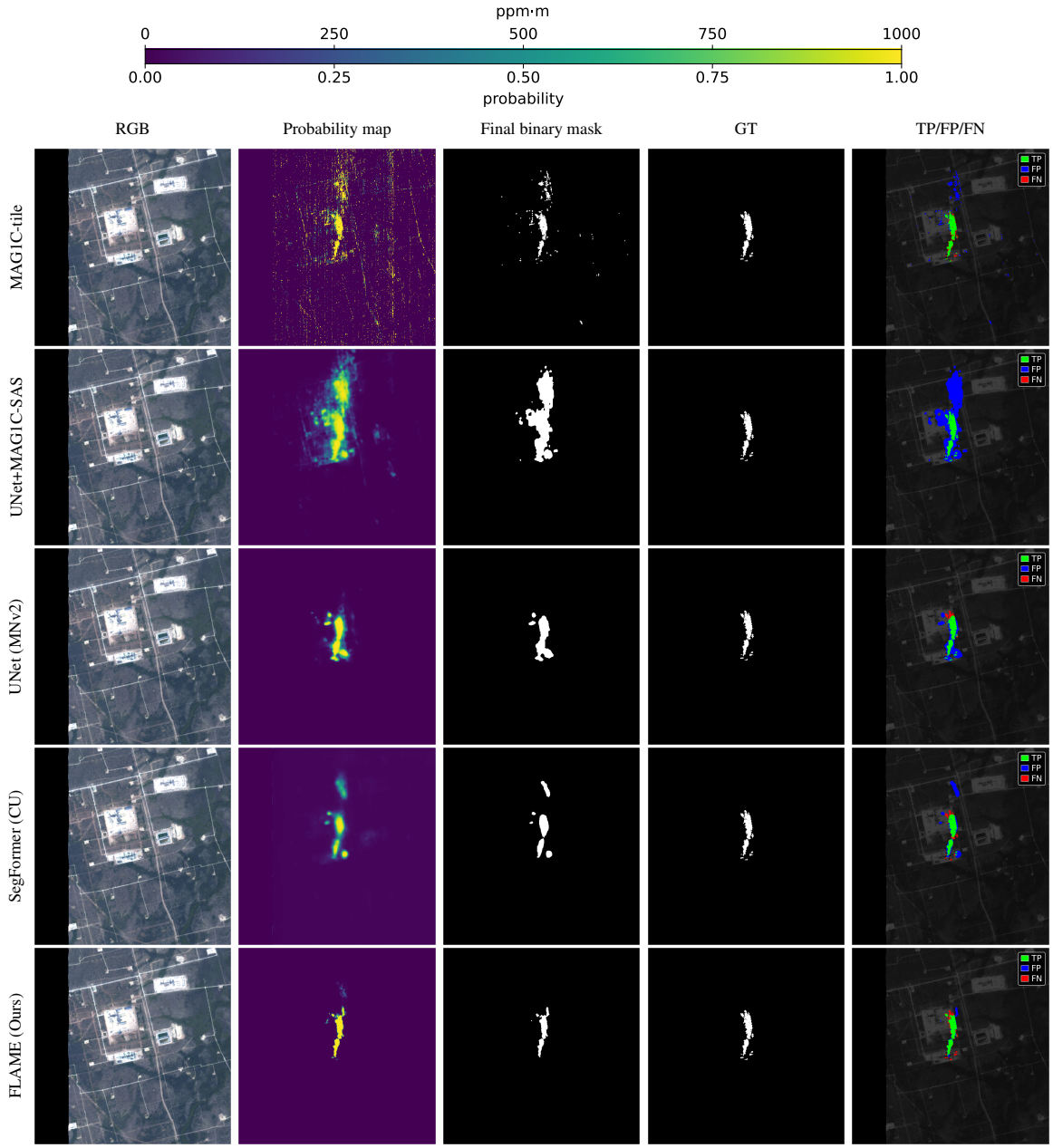


*Figure 11.* Per-pixel outputs on a strong-plume test tile. Columns show the RGB tile, the probability map, the final binary mask, the ground-truth plume mask, and the TP/FP/FN overlay. From top to bottom, rows show MAG1C-tile, UNet+MAG1C-SAS, UNet with the MobileNetV2 encoder, SegFormer with the ConvUp decoder, and FLAME. For the MAG1C-tile row, the probability-map column instead reports the matched-filter enhancement in ppm·m using the top color bar; all learned models use sigmoid probabilities with the bottom color bar.

**Detailed Strong-Plume Case** In Figure 11, the MAG1C-tile score activates broadly across the tile, with strong responses on roads and field boundaries that share a directional radiance pattern with the plume. UNet+MAG1C-SAS inherits parts of this background structure and produces a wide diffuse score around the plume body. The end-to-end UNet and SegFormer scores are concentrated near the plume but leak into adjacent surfaces, which translates into the over-extended binary masks visible in the third and fourth rows. The FLAME score collapses onto a thin region that closely follows the annotated plume, and its binary mask correspondingly contains the smallest amount of background activation among the methods shown.

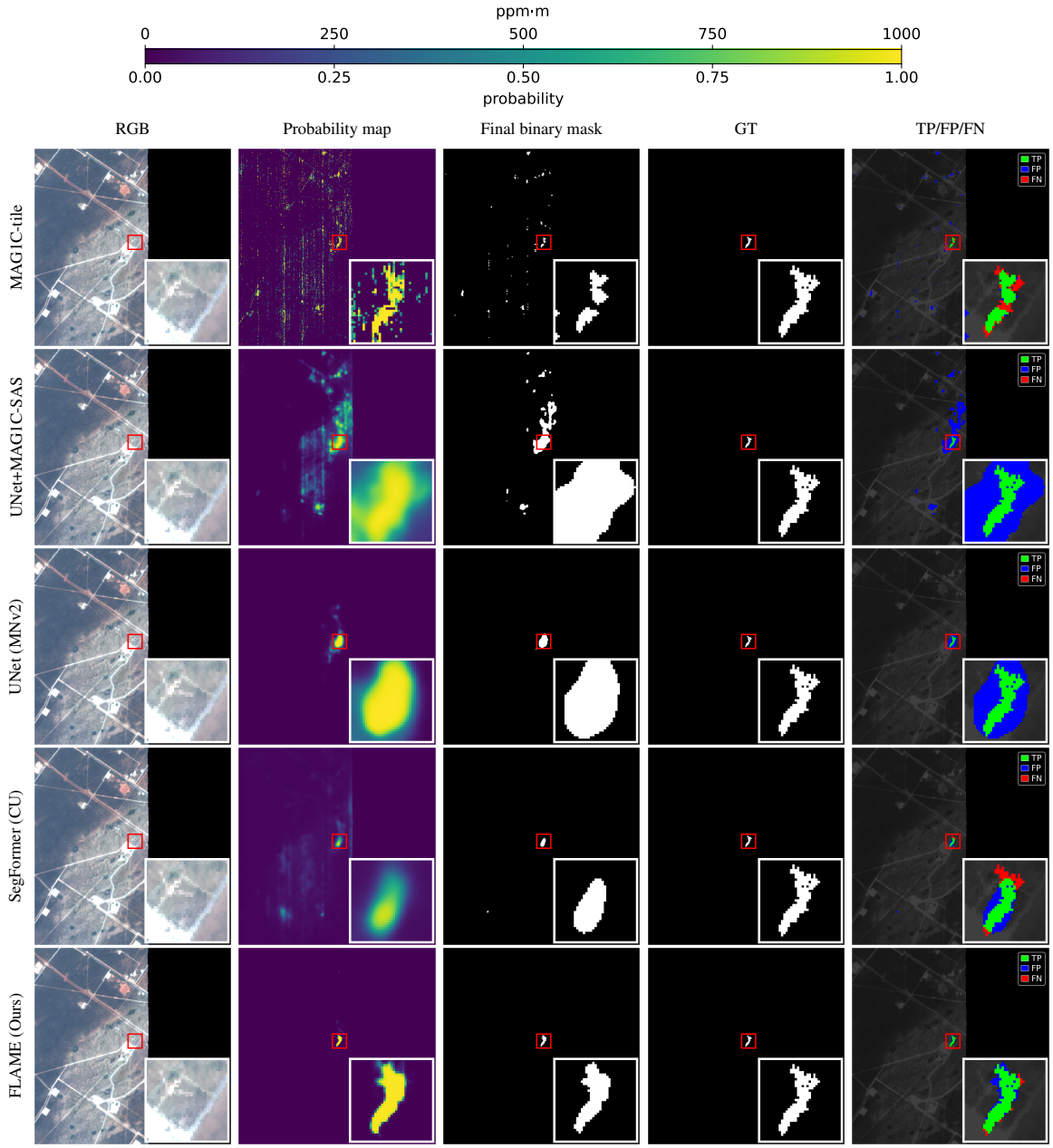


*Figure 12.* Per-pixel outputs on a weak-plume test tile. Row and column layouts follow Figure 11. Each panel includes a zoomed inset of the plume region marked by the red rectangle.

**Detailed Weak-Plume Case** Figure 12 shows the same comparison on a tile where the methane signal is too faint to be reliably distinguished from radiance variations in the score map alone. MAG1C-tile leaves the plume region nearly indistinguishable from the surrounding clutter. UNet+MAG1C-SAS produces a strong but spatially diffuse response over the plume area, and the resulting mask spreads well beyond the annotated region. UNet with the MobileNetV2 encoder under-predicts the plume body, while SegFormer with the ConvUp decoder produces only fragmented activations that fail to capture the elongated plume shape. FLAME yields a focused score that is largely confined to the plume region, and the corresponding mask preserves the elongated morphology visible in the ground truth.

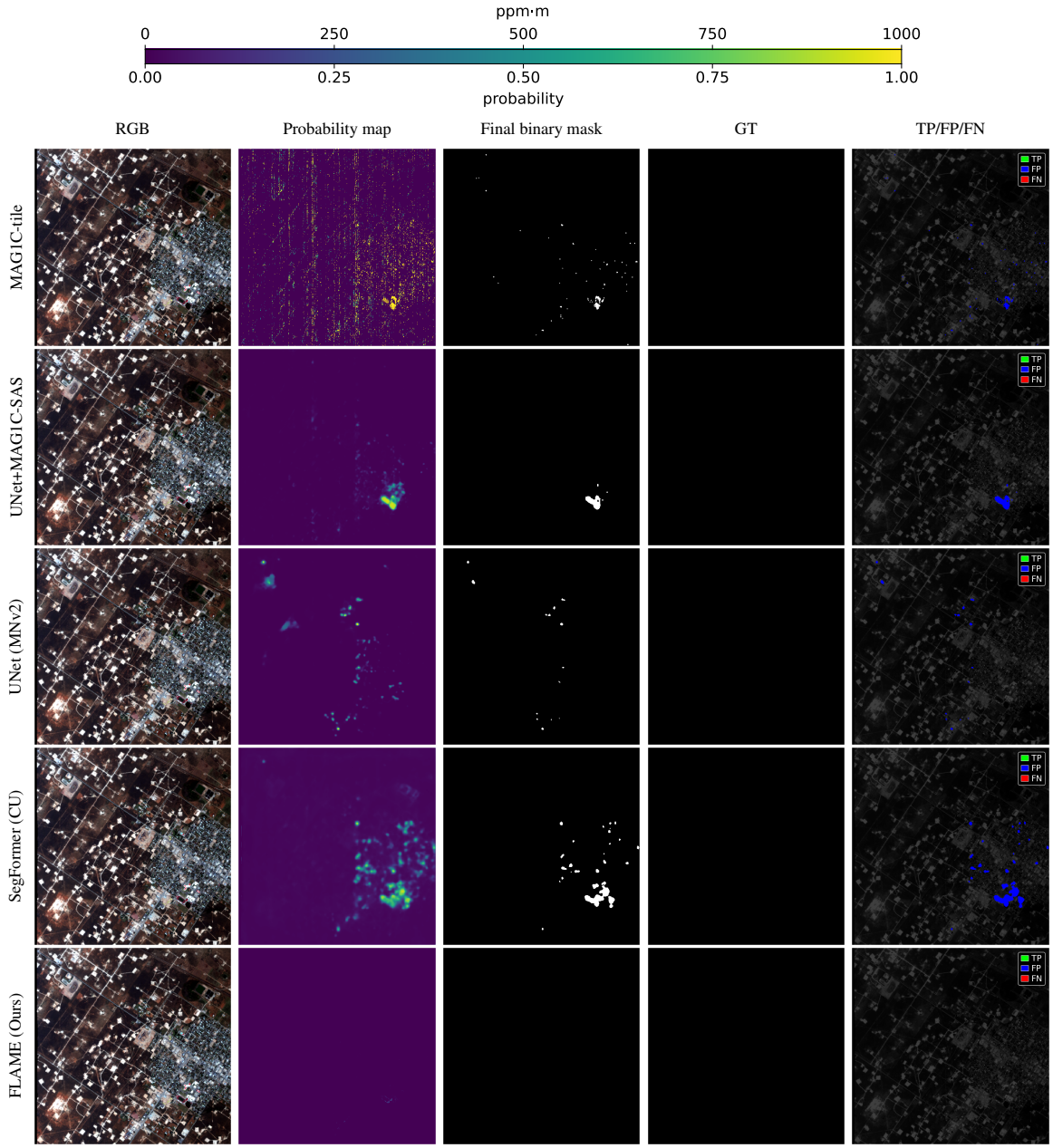


*Figure 13.* Per-pixel outputs on a plume-free test tile. Row and column layouts follow Figure 11. The ground-truth column contains no plume pixels, and the TP/FP/FN overlay shows false positives in blue on a darkened RGB image.

**Detailed Plume-Free Case** Figure 13 shows the same comparison on a tile that contains no annotated plume but a high density of urban surface clutter. MAG1C-tile produces strong directional responses along streets and rooftops, and these responses propagate into the binary mask of UNet+MAG1C-SAS. UNet with the MobileNetV2 encoder and SegFormer with the ConvUp decoder both generate scattered detections in regions where surface materials share spectral cues with methane absorption. The FLAME score remains close to zero across the entire tile, and its binary mask is empty, which is consistent with the lower pixel false positive rate reported in Section 5.2.