FAINT: False-Alarm-aware Input-Representation Trade-offs for Methane Plume Detection on AVIRIS-NG

Sophia Peng
Green Level High School

Learned segmentation is the state of the art for detecting methane point-source plumes in imaging spectroscopy, yet the choice of input representation, raw radiance bands, a physics-based enhancement product such as mag1c, or their combination, is usually justified by aggregate metrics that say little about faint plumes near the sensitivity floor, where detection actually fails. We contribute a controlled measurement of this choice, FAINT (False-Alarm-aware INput-representation Testing). On the STARCOP AVIRIS-NG dataset (Permian Basin flightlines; held-out test set of 342 tiles spanning roughly 100–7,000 kg/hr) we train one fixed U-Net under six input representations forming a mag1c/RGB/SWIR ablation grid, three seeds each, holding architecture, loss, training budget, normalization, and a flightline-grouped train/validation split fixed. Every detection measurement is paired with its false-alarm cost, resolved by emission-rate bin, and summarized by q_{50}, the emission rate at 50% detection probability, adapting probability-of-detection methodology from the controlled-release literature.

Three findings stand out. First, the physics-based product carries the signal: every mag1c-bearing representation beats a raw matched-filter baseline by 2.6–3.2× in pixel AUC-PR (0.44–0.55 vs. 0.17), while RGB-only and SWIR-only models are near-useless, and the baseline's high tile detection rate (0.95) conceals a diffuse false-alarm burden (false-positive pixels in 97% of clean tiles, against 7.6–18.7% for the learned mag1c models). Second, false-alarm-aware calibration changes conclusions. At operating points selected on validation clean tiles under the operational four-pixel connected-component alert rule, locked, and applied to the test set once, mag1c+RGB+SWIR detects the most plumes at every tested false-alarm budget (5–20%), both pre-specified contrasts are positive with flightline-clustered bootstrap intervals excluding zero, and an apparent budget-dependent crossover from an earlier maximum-confidence proxy analysis is overturned; the raw matched filter, though operable, detects far fewer plumes (0.08 vs. 0.39–0.52 at a 10% budget). Third, stability checks bound the claims: a three-seed ensemble gives the strongest calibrated detector (0.52 detection at a 10% budget, 0.68 at 20%), but the aggregate SWIR advantage changes sign across both model seeds and train/validation splits, while the RGB gain stays positive on all three tested splits and mag1c+RGB pairs the lowest detection floor ($q_{50} \approx 681$ kg/hr) with the best across-split mean AUC-PR. Exploratory zero-shot transfer to 20 EMIT granules detects roughly half the plume scenes but fails to reject EMIT's mag1c confounders, so cross-sensor deployment is unsupported without sensor-specific adaptation.

The contribution is a reusable false-alarm-aware protocol and an honest account of which conclusions survive seeds, splits, semi-synthetic plume injection, and cross-sensor transfer.