Reducing False Positives in Operational Remote Sensing: Lessons from Methane Plume Detection

Julia Wąsala[1,2], Joannes D. Maasakkers[2], Holger Hoos[3], and Mitra Baratchi[1]

[1] Leiden Institute for Advanced Computer Science (LIACS), the Netherlands
[2] SRON Space Research Organisation Netherlands, the Netherlands
[3] Chair for AI Methodology (AIM) at RWTH Aachen, Germany.
{j.wasala}@liacs.leidenuniv.nl

Abstract. False positives remain a persistent challenge in operational remote sensing detection tasks, such as greenhouse gas plume monitoring and oil spill detection. Benchmark performance often masks this problem, as operational settings are characterised by severe class imbalance, limited labelled data, and difficult true negatives that are hard to separate from true positives. We study this problem through the lens of atmospheric methane plume detection, where end-to-end models, which are generally applicable and not task-specific, achieve competitive benchmark results but generate substantially more operational false positives than a machine learning pipeline designed by domain experts. We find that the discriminating features of difficult true negatives form a marginalised discriminant, a feature that is informative only within a subpopulation of the training data, making it hard to learn from the full distribution. We investigate multiple model- and data-centric techniques targeting the marginalised discriminant to address false positives in an end-to-end pipeline and analyse the properties of successful approaches. The most effective solution, introduced in related work, most strongly limits what the model needs to learn by decomposing the problem into two steps to reveal the hidden feature, and by increasing data efficiency through pre-computed features using feature engineering. Our findings suggest that successful operational pipelines do not solely rely on smarter learning algorithms, but also on the degree to which the complexity of the problem to be solved is constrained before machine learning is applied.

Keywords: False positives · Machine Learning · Earth Observation

1 Introduction

Operational remote sensing detection tasks, such as greenhouse gas plume monitoring and oil spill detection are often implemented as binary classification by detecting a class of interest against everything else. Such settings are often characterised by severe class imbalance, limited labelled data, and difficult negatives. Difficult negatives are negative scenes whose spatial or spectral features resemble

that of the target class in the sensor data and are difficult to separate from true positives [13,18,16,1]; they are sometimes also referred to as *hard negatives*. These conditions make operational deployment substantially harder than performance on clean, balanced benchmarks suggests, and false positives remain a persistent challenge even when machine learning (ML) models achieve competitive test set performance. Addressing this gap is critical for making automated detection systems operationally viable at the scale usually required when dealing with satellite data.

Practitioners bridge this generalisation gap through compound pipelines that combine learned components with application-specific design choices, such as extensive feature engineering. These pipelines are typically designed and validated empirically in a specific setting (e.g., atmospheric plume detection with a particular sensor dataset); as a result, the conditions under which they generalise are not well understood. End-to-end, data-driven models that learn directly from the data are more generalisable, as they make fewer application-specific assumptions and are not tied to a particular data distribution. Still, in operational remote sensing, compound pipelines remain the norm, and whether this reflects the demands of these settings, limitations of end-to-end approaches, or both, is not well understood. Standard remedies for class imbalance and false positive control (e.g., hard or difficult negative mining, focal loss, cost-sensitive learning [9,2]) are well understood and effective in general settings, but can fall short in operational remote sensing settings, which suggests that class imbalance is not the sole issue. As a step towards a better understanding of when and why knowledge-driven designs are necessary, we study this question through the lens of TROPOMI [4] methane plume detection, a well-documented operational detection problem with a domain-designed compound pipeline that reliably controls false positives in deployment [18].

We find that the false positive problem in this setting is rooted in a distributional property we call the **marginalised discriminant**: a feature that distinguishes a subgroup of hard negatives from true positives, but appears uninformative in the training data as a whole. As a result, models trained on the full joint distribution may not learn to extract this feature. The marginalised discriminant is structurally the inverse of a confounder: a confounder is a spurious feature that appears informative in the training data but is uninformative in the real world; a marginalised discriminant is a feature that is informative in the real world, but appears uninformative in the training data. Through systematic experimentation with model- and data-centric false positive mitigation measures, we investigate whether end-to-end approaches can address the marginalised discriminant, and what the structure of domain-designed compound pipelines reveals about why they succeed. The *lessons learned* from this work are as follows:

- Across a range of architectural and data-centric mitigation strategies applied to end-to-end methane plume detection models, none consistently learn to use the marginalised discriminant in this data-constrained regime;
- The operational success of domain-designed compound pipelines can be explained by how they restructure the learning problem: decomposing classifica-

tion into simpler steps combined with domain-specific feature representations makes the marginalised discriminant easier to learn by reducing how much the model needs to learn. In regimes with a marginalised discriminant and data constraints, this suggests that simplifying the machine learning problem is more effective for false positive control than end-to-end mitigation strategies.

2 Related Work

False positive control in machine learning has been approached from two complementary angles: data-driven methods addressing imbalance and difficult negatives through model, data, and loss design, and knowledge-driven pipelines incorporating domain structure alongside learned components.

False positives and class imbalance. Class imbalanced training data is a well-documented challenge in remote sensing classification [12,2], and addressing it has been the subject of substantial methodological effort. Proposed approaches broadly fall into three categories [2,9]: data-level methods, such as resampling and augmentation; algorithm-level methods, including cost-sensitive learning, transfer learning, and ensembles; and loss-based methods, such as focal loss, which assigns higher weights to frequently misclassified instances. Across these categories, the gains achieved by various methods can be inconsistent or setting-dependent. Ghosh et al. [3] systematically investigated the effect of CNN depth, regularisation, and dataset size on model performance under class imbalance, finding that increasing model capacity does not always improve robustness to class imbalance, and can even have negative effects. A persistent issue in operational settings is the mismatch between balanced training distributions and highly imbalanced deployment distributions, which causes benchmark performance to underestimate operational false positive rates [12]. Here, we consider the problem of class imbalance in a concrete operational setting, examining which mitigation measures can and cannot address false positives arising from a specific difficult negative subpopulation.

End-to-end *vs* Compound Plume Detection Pipelines. Several compound, knowledge-driven approaches have been proposed in the methane plume detection literature, each addressing the problem differently. Schuit et al. [18] proposed a two-step pipeline in which a CNN model pre-filters TROPOMI images based solely on patterns in the methane data [4] to remove scenes clearly devoid of plumes ("empty"), after which an SVM uses physics-based features to discriminate true plumes from artefacts, reducing complexity at each stage through domain-specific feature design. Kurchaba et al. [7] built on this pipeline by replacing the SVM with a second neural network operating on the raw data cube, made feasible by a substantially larger dataset and guided by a pre-computed plume mask from the first-step CNN. Wąsala et al. [20] took a different approach, proposing a neural architecture search approach to automatically design multimodal fusion networks from TROPOMI data, but still found that an auxiliary methane-only classifier is necessary to reduce operational false positives. In multispectral and hyperspectral detection, ML models are typically applied to derived methane

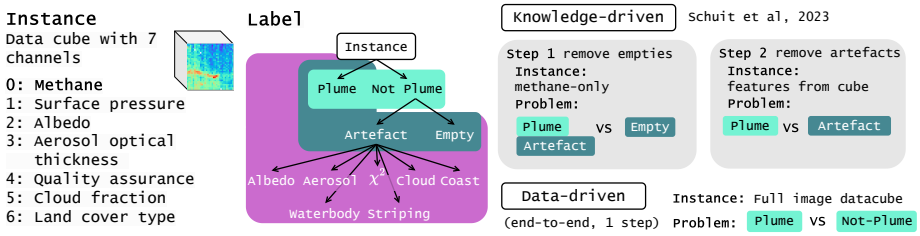


Fig. 1. Left: Methane plume detection data representation (the training instance and annotations). Right: Knowledge-driven domain pipeline [18], using different parts of the data cube and images of different classes at each step, where an end-to-end pipeline aims to solve the problem in a single step with all available data.

enhancements rather than the raw spectral data cube, reducing the complexity of the learning problem without explicit feature engineering at the classification stage. Růžička et al. [16] followed this approach and further combined multiple models in a random seed ensemble to reduce false positives in methane plume segmentation. Across these works, end-to-end approaches consistently were found to require additional interventions to reduce operational false positive rates. Here, we investigate what this reveals about when and why purely data-driven approaches are insufficient, with the aim of deriving lessons that extend beyond a specific sensor and application.

3 Methane Plume Detection Data

The TROPOMI satellite methane data product [4] provides daily global observations of atmospheric methane concentrations, harbouring many localised methane emission plumes. We use the methane plume detection dataset from Wąsala et al. [19], adapted from Schuit et al. [18] (Zenodo, version 1.0.0 [17]). The dataset and its use in knowledge-driven and data-driven pipelines is summarised in Figure 1. The dataset comprises 7694 images, each consisting of 32×32 pixel TROPOMI observations (Operational data v02.04.00 [4]). Each training instance is a datacube of seven channels, consisting of a primary methane concentration channel and six auxiliary channels from the TROPOMI methane data product, labelled as either "plume" or "not plume." We apply the same normalisation, data augmentation, and data partitioning as Wąsala et al. [19]. We fill missing values with the per-image median, and apply the resampling scheme proposed by Wąsala et al. [19] to mitigate the representation bias caused by missing pixels.

The dataset presents two properties common to operational EO detection tasks: a hierarchical class structure (Figure 1, right) with a conditional discriminant, and a mismatch between training and deployment distributions. The "not plume" class consists of two subpopulations: clearly empty images (30%), and artefacts (14%) that exhibit plume-like features in the primary methane channel arising from imperfections in the retrieval process. The artefact class was defined

by Schuit et al. [18] after discovering a pattern in the false positives produced by their methane-only CNN; this motivated the development of a follow-up SVM incorporating features calculated using auxiliary channels to discriminate artefacts that are indistinguishable from true plumes in the primary channel alone (Figure 1, right). This introduces a conditional structure: the auxiliary data layers only provide discriminative information if a plume-like feature is already present in the primary channel.

Artefacts. Multiple mechanisms can lead to artefacts. We categorise them into seven sub-types based on the auxiliary channels (also shown in Figure 1): albedo, aerosol, χ^2 (goodness-of-fit of the retrieval), cloud, coast, waterbody (including rivers and lakes), and striping. A scene can carry multiple subtype annotations, and the leading cause is not always identifiable. This reflects a fundamental difficulty: the artefact class is for a large part defined by what it is not (not a real plume), rather than by shared positive characteristics. Intra-class variability among artefacts is therefore high, and the class boundary is inherently fuzzy. Because auxiliary channels are only informative for artefact discrimination, which is a minority subpopulation, their signal is sparse across the full training distribution. For most training instances, the auxiliary channels carry no discriminative information, which risks creating a vanishing gradient problem, in which the model receives too little signal to reliably learn auxiliary channel features end-to-end.

Operational scenario. Plumes occur relatively infrequently in operational scenarios [18]. Our training set is balanced, which introduces a mismatch between training and deployment distributions. As a result, test performance may not be representative of operational performance [20,6]. To address this, we evaluate on a previously unseen operational dataset, following Wąsala et al. [19], obtained from one week of TROPOMI observations over land (14–20 March 2022, consisting of 20 965 images). Images were extracted using a moving window of 16 pixels over each TROPOMI scene. We counted images flagged as containing a plume at a CNN score threshold of 0.5 and assessed geographical distribution as a qualitative indicator of generalisation. We did not label flagged images; instead, we compared against the 260 evaluated detections[4] produced by the Schuit et al. [18] for the same week: they first deduplicated flagged images by removing spatially overlapping flags based on plume mask overlap, then two methane analysis experts labelled the remaining scenes, retaining only scenes labelled as plumes. Though the false negative rate is not known, we do not expect the true number of plumes to be substantially higher, as methane plumes large enough to be detectable by TROPOMI are rare. Consequently, substantially higher flag counts suggest increasing false positives.

4 Uncovering the Marginalised Discriminant

Following Wąsala et al. [20], we define the methane plume detection problem as an end-to-end binary classification task over multimodal satellite images, where

[4] http://earth.sron.nl/methane-emissions. Last accessed 27 May 2026

each instance is a data cube consisting of a primary data layer I^1 showing the methane concentration field and $M-1$ supporting data layers $\{I^2, \ldots, I^M\}$ from the TROPOMI methane data product [4]. Due to the conditional structure, the auxiliary data layers $\{I^2, \ldots, I^M\}$ have limited discriminative power in the marginal distribution, but are substantially more informative in the conditional distribution $p(y \mid I^2, \ldots, I^M \in \mathcal{P})$, where y is the class label and $\mathcal{P}$ denotes scenes exhibiting a plume-like signal in the primary layer. We refer to this conditional signal as a **marginalised discriminant**. Models trained on the joint distribution are more likely to underexploit this conditional signal, converging instead to solutions that use the supporting layers inconsistently and without the conditional structure the problem requires. Neural networks can inherently learn similar conditional patterns. The problem arises in the current regime, where training instances are relatively sparse and limited labelled data is available.

We investigate two complementary groups of approaches designed to enable end-to-end models to learn the marginalised discriminant, which is necessary for artefact discrimination in methane plume detection. First, we take a *model-centric* approach, using domain knowledge to increase the inductive bias through architectural re-design and the introduction of an attention mechanism. Second, considering that the marginalised discriminant arises from data limitations, we take a *data-centric* approach, applying data cleaning to address label-level issues and incorporating auxiliary information into the labels to guide the training.

4.1 Model-Centric Mitigation Strategies: Architecture Design

The first likely explanation for why end-to-end models fail to discover the marginalised discriminant is that their architecture does not reflect the hierarchy of the problem: in a standard multi-branch network, all input branches are treated equally, and the prioritised position of the methane channel is not enforced, but must be learned. We investigate two architectural mitigation strategies designed to make this hierarchy explicit in end-to-end binary classification (shown in Figure 2 along with the baseline architecture):

- **Hierarchical architecture**: We transform the standard multi-branch architecture, where each channel has an independent branch, into a two-branch structure with a dedicated methane branch and a single shared branch for all auxiliary channels (Figure 2b), which we call a hierarchical model. This architecture explicitly gives more weight to the methane channel.
- **Cross-branch attention**: We introduce a *cross-branch* attention mechanism to guide attention in auxiliary channels towards where the model is looking in the primary channel (Figure 2c). Using the spatial attention module from CBAM [22], we compute attention maps from the methane branch and use them to gate feature extraction in the auxiliary branches. This is conceptually related to the masking procedure by Schuit et al. [18], which computes SVM features only within a plume mask derived from Grad-CAM activations of the methane channel alone, though unlike that approach, the attention mechanism operates on all scenes, including non-plume, and any plume-artefact hierarchy is not enforced explicitly, but must instead be learned.

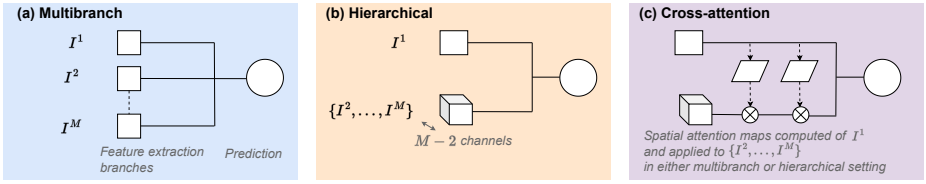


Fig. 2. Three architectural variants. **(a)** Multi-branch: independent feature extraction branches per input channel [20,19]. **(b)** Hierarchical: a dedicated branch for the primary methane channel and a single shared branch for all auxiliary channels. **(c)** Cross-attention: spatial attention maps computed from the primary channel guide auxiliary feature extraction; applicable in either the multi-branch or hierarchical setting.

4.2 Data-Centric Mitigation Strategies: Data Pruning

A second potential explanation is that label noise prevents end-to-end models from learning the conditional artefact discriminant. The dataset contains two such sources of noise. First, since labels were originally produced for the pipeline by Schuit et al. [18], in which the SVM evaluates a detected plume (mask) rather than the full image, scenes labelled as "not plume" may contain one or more plumes outside the mask region (*label scope mismatch*). Second, some detections were validated using external data sources, such as optical imagery, that are unavailable at training time, which means that some scenes may be indistinguishable from the opposite class without this context (*label context mismatch*). Both sources disproportionately affect artefact scenes, where the distinction between true and false positives is most context-dependent. We investigate two data-centric mitigation strategies designed to address these sources of noise:

- **Automated data pruning**: Since it is infeasible to identify affected scenes post hoc, we investigate automated data pruning using confident learning (CL) [15], which identifies potentially mislabelled or noisy examples by finding instances confidently misclassified by the model, using out-of-fold predictions from cross-validation. We evaluate two pruning methods by Northcutt et al. [15]: Confident Joint (CJ), which flags off-diagonal entries of a thresholded per-class confusion matrix, and Prune by Class and Noise Rate (CNR), which combines two thresholding criteria by an elementwise and-operator.
- **Formulating a multi-label classification problem**: We hypothesise that fine-grained artefact subtype labels can help CL more accurately target label context mismatch, by providing richer supervision that gives more information on the conditional structure of the problem. We therefore evaluate both pruning methods under two label configurations: standard binary labels, and an expanded multi-label setup incorporating artefact subtype annotations.

5 Setup of Experiments

The following describes the shared framework underlying our empirical evaluation. A compound pipeline with domain-specific features is already known to achieve

robust operational performance [18], so the central question is whether end-to-end models can match this without problem decomposition or feature engineering. We investigate this through the following questions:

Q1 Can standard end-to-end models reliably discriminate artefacts from plumes, and does failure to do so account for elevated false positive rates?
Q2 Does encoding the feature hierarchy into the model architecture provide sufficient inductive bias for end-to-end models to learn conditional artefact discrimination without explicit enforcement?
Q3 Does automated data pruning of potentially mislabelled or ambiguous instances help end-to-end models learn the conditional artefact discriminant?

Model training. Our base architecture is a multi-branch data fusion CNN with one input branch for each data source (based on Wąsala et al. [20] and shown in Figure 2, left); we selected it over higher-capacity backbones (ResNet [5] and ConvNeXt [21]), which showed greater susceptibility to overfitting and higher variance across random seeds. We trained each model for 50 epochs, keeping the weights at the epoch with the lowest validation loss. We used the AdamW optimiser [11] and the cosine learning rate scheduler [10]. Binary classification uses a single sigmoid output node with binary cross-entropy loss; multi-label classification uses $K = 10$ output nodes (corresponding to plume, empty, artefact and seven artefact subtype classes, where the majority of artefacts have an additional subtype class label) with independent sigmoid activations and averaged binary cross-entropy. Whenever the architecture or training data changed substantially, we tuned the batch size, learning rate, weight decay and dropout by using Ray Tune's random search [14] in combination with the ASHA scheduler [8] (with 50 trials). We trained and evaluated each optimised configuration five times with different pseudo-random number seeds. All experiments were run on a compute node equipped with an NVIDIA L4 GPU and 32 GB of RAM and an Intel Xeon Gold 6126 2.6GHz 12-core CPU per neural network configuration running CentOS-7. This work was performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

Data cleaning. We applied CJ and CNR to binary and multi-label classification (the latter through a one-vs-rest scheme, as implemented in the `cleanlab` introduced in Northcutt et al. [15]). Out-of-fold predictions were produced with 5-fold cross-validation on both training and validation sets, the latter to avoid data leakage from the validation-loss-based learning rate scheduler. We then pruned using each cleaning method and re-trained the models from scratch on each cleaned dataset version.

6 Analysis

In this section, we investigate the marginalised discriminant in operational methane plume detection, and we specifically examine what the structure of the domain pipeline reveals about the reasons underlying the difficulty of learning the discriminant from the joint distribution.

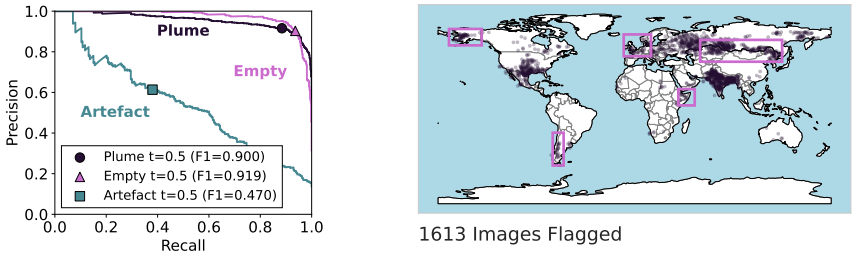


Fig. 3. Per-class precision-recall curves (left) and geographic distribution of flagged images (right) for a multi-label late fusion CNN. **Left**: markers indicate performance at threshold 0.5; while plume and empty classes are well-fitted across a wide range of thresholds, no threshold yields both high precision and recall for the overall artefact class. **Right**: flagged images include detections in regions unlikely to contain concentrated methane emissions based on historic detections (marked by pink boxes), as well as implausibly high detection densities in others (India, U.S., eastern Russia), suggesting a significant portion are false positives.

6.1 Q1: Characterising the marginalised discriminant

To characterise how the marginalised discriminant occurs in end-to-end models, we train a multimodal late fusion CNN [19] in a multi-label setting, with label vectors for plume, empty, artefact and the seven artefact subtypes (described in Section 3). This allows us to assess artefact discrimination independently of aggregate performance through per-class precision-recall curves. Figure 3 (left) shows the results of a single training run. While the model learns the plume and empty classes well (i.e., with high precision and recall achievable across a wide range of thresholds), no classification threshold yields both high precision and recall for the artefact class, and the best F1-score achieved in our experiments remains around 0.55. Reweighting the loss based on class balance and focal loss did not resolve the issue, and in some configurations destabilised training. This pattern is largely invisible in aggregate metrics: the same configuration achieves a balanced binary accuracy of 0.895 ± 0.002, precision of 0.905 ± 0.009, and recall of 0.905 ± 0.016 on the plume/not plume task, illustrating how standard evaluation can mask the artefact problem in the operational binary task. For comparison, the confusion matrix for the SVM step reported by Schuit et al. [18] yields an artefact F1-score of 0.894 (compared to 0.470 for the multi-label end-to-end model here) with few misclassifications between artefacts and confirmed plumes, achieved with a smaller dataset containing fewer artefact examples than used here. This result shows that strong artefact discrimination is achievable in this data-constrained regime.

The operational evaluation makes the gap concrete. Direct comparison of classification metrics with Schuit et al. [18] is not straightforward: their pipeline operates on a different task (the second-stage SVM classifies image regions pre-selected by a CNN, rather than full images), and our dataset merges and extends their two original training sets with additional instances, redistributed across

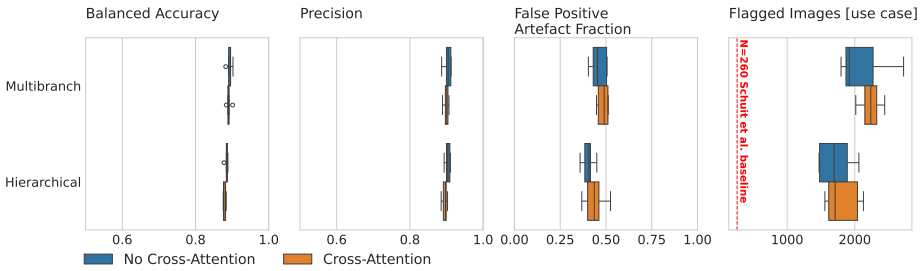


Fig. 4. Architectural mitigation results: balanced accuracy, precision, false positive artefact fraction (artefact share of all false positives), and operational flagged image count, for multi-branch and hierarchical architectures with and without cross-branch attention. Schuit et al.'s [18] baseline of 260 detections is shown for reference. Hierarchical models flag fewer images, but not to a deployable level, with detection densities remaining implausibly high in several regions (Figure 3). Cross-branch attention has no significant effect (according to the Wilcoxon signed-rank test).

train and test splits for class balance and spatial decorrelation. We therefore use operational flag counts as the most comparable proxy. The model flags 1710 ± 277 images as plume over the week of observation data considered in our use case, far exceeding the 260 detections from Schuit et al. [18].

While a direct numerical comparison is complicated by deduplication (Schuit et al. [18] merge overlapping mask-based detections; we did not deduplicate our overlapping image-level windows) and the lack of ground truth labels for false negatives, we rely on geographic plausibility as a proxy for false positive identification. The geographic distribution of flagged images (Figure 3, right) suggests an elevated false positive rate: detections appear in regions unlikely to contain concentrated methane emissions based on historic emission locations (e.g., Alaska, Chile, western and northern Europe),[5] and at implausibly high densities in others (e.g., India). Together, these results indicate that the artefact discrimination gap observed on the labelled test set translates into an operational false positive problem. The following analyses investigate how the same could be achieved in end-to-end models.

6.2 Q2: Architectural changes do not enable conditional learning

We tuned and trained the multi-branch and hierarchical architectures on the binary classification task. In addition to standard binary balanced accuracy and precision, we tracked the fraction of false positive artefacts, which is the proportion of false positive (plume) predictions attributable to artefact scenes. A high fraction indicates that false positives are driven by artefact misclassification rather than clear empties. Figure 4 shows that hierarchical models achieve a

[5] https://atmosphere.copernicus.eu/ghg-services/cams-methane-hotspot-explorer. Last visited 27 May 2026

slightly lower fraction of false positive artefacts, with no significant changes in binary accuracy and precision, suggesting that the classification boundary is slightly shifting, but the hierarchical models are still not effectively extracting the marginalised discriminant. This shift in boundary does lead to a decrease in the number of flagged images, though the count still far exceeds the baseline of 260 validated detections by Schuit et al. [18], rendering the architectural change alone insufficient for operational use.

To better understand what the models are learning, we visualise attention patterns using Grad-CAMS (Figure 5). Both architectures consistently identify plume-like shapes in the methane channel. However, attention patterns in the auxiliary channels are inconsistent: multi-branch models attend to incorrect spatial regions (instead of extracting features from the same region where the plume was identified), and hierarchical models sometimes attend to almost the full image, not discerning specific regions. These patterns confirm that the models are not learning to use the auxiliary channels conditionally: they recognise the primary signal, but do not extract discriminative features from auxiliary channels in the regions where they would be informative. The hierarchical architecture (bottom row) treats the auxiliary channels as a single data cube, but this does not produce more focused or meaningful attention patterns.

When introducing the cross-attention mechanism, we find no significant improvement (Figure 4, orange). A likely explanation is that the mechanism does not help when methane attention is not sufficiently focussed (e.g., for empty scenes and difficult artefacts).

6.3 Q3: Over-pruning of artefacts leads to loss of generalisation

Automated data cleaning reduces the number of flagged images for binary models (Figure 6), although the total number of flagged images is still elevated compared to the results by Schuit et al. [18]. For multi-label models, cleaning has the opposite effect: flagged image counts increase, driven by disproportionate pruning of artefacts. Unlike binary models where artefacts and empties are grouped, the multi-label models expose artefact performance explicitly. Out-of-fold multi-label predictions on the training set, which are the predictions used by the cleaning algorithms, yield an artefact precision of 0.694 ± 0.018 and recall of 0.493 ± 0.022, substantially below the plume and empty classes (plume: precision of 0.900 ± 0.006, recall of 0.909 ± 0.007; empty: precision of 0.880 ± 0.10, recall of 0.899 ± 0.005). These lower artefact scores cause confident learning to identify artefacts as potentially mislabelled and remove them, triggering a negative feedback loop: poor artefact performance leads to artefact removal, which further degrades artefact discrimination and increases false positives. The fact that removing too many artefacts worsens performance confirms that the model does extract some signal from them, just not enough to model artefacts reliably.

Training with extended label vectors in a multi-label setting can improve generalisation compared to binary models (Figure 6), but the over-pruning result reveals its limits; even with richer supervision, the predicted scores for many artefacts are of low enough quality for the cleaning algorithms to identify

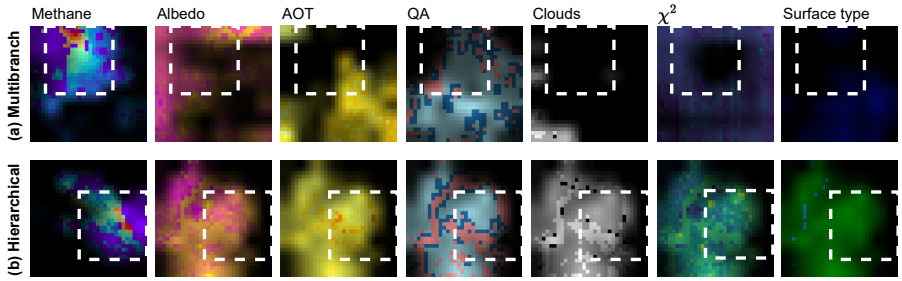


Fig. 5. Grad-CAM attention maps for two artefacts, overlaid on each input channel. The white box marks the plume-like region in the methane channel where effective attention maps should focus. The multi-branch model **(a)** produces independent attention maps per channel; the hierarchical model **(b)** produces one shared map for the auxiliary channels. Black indicates no attention, fully transparent indicates maximum attention. Both architectures attend consistently to plume-like shapes in the methane channel, but fail to do so in the auxiliary channels, confirming that auxiliary features are not extracted conditionally on the primary signal.

them as potentially mislabelled, underscoring that artefact discrimination is still suboptimal. In this regime, the cleaning algorithm has no way to tell the difference between a hard example and a noisy or mislabelled one, which disproportionately affects artefacts as the hardest subclass.

6.4 Discussion: The role of knowledge-driven pipeline design

The false positive mitigation measures evaluated in **Q2** and **Q3** sometimes improved aggregate performance measures, but did not sufficiently help the end-to-end models learn the marginalised discriminant, unlike the pipeline by Schuit et al. [18], whose critical design choices (decomposition into two steps and feature engineering) are not learned, but decided by the authors, reducing what the model needs to learn from limited data. Both our architectural and data-centric interventions still depend on the model learning the right patterns, which may explain their limited effectiveness.

Concurrent work by Kurchaba et al. [7] suggests a more data-driven solution becomes viable with substantially more artefact data, although notably, their approach still retains a non-learned structural simplification in the form of a pre-computed plume mask concatenated to the raw data cube. This approach is consistent with our finding that problem simplification remains important even as data availability increases. More broadly, our analyses suggest knowledge-driven design plays a crucial role in data-limited ML for EO pipelines. This is reflected in practice, as the ML4EO community continues to rely on classical models and compound pipelines, not only due to limitations of end-to-end approaches, but also because physics-based design is often readily available, and individual domain scientists rarely need pipelines that generalise beyond their target application. Yet, the tradeoff between data-driven and knowledge-driven design has received

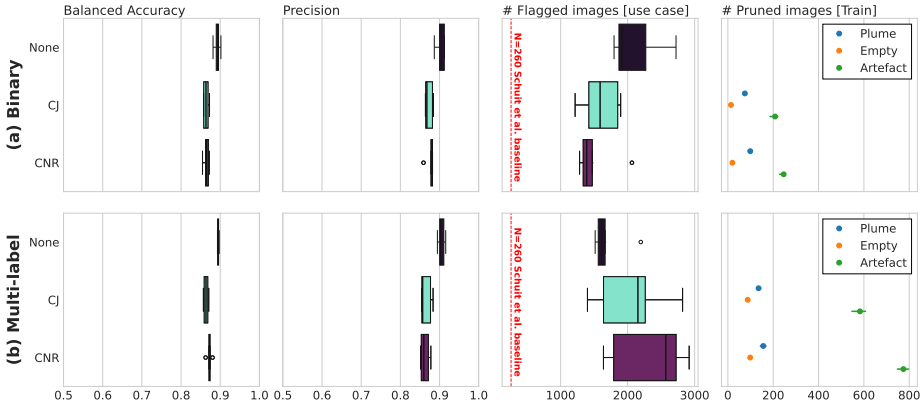


Fig. 6. Results of data-centric mitigation strategies. Balanced accuracy, precision, flagged image counts in the use case evaluation, and number of pruned training images per class, for binary **(a)** and multi-label **(b)** models with no cleaning, CJ, and CNR pruning strategies. The Schuit et al. [18] baseline of 260 detections is shown for reference. Data cleaning reduces flagged images for binary models but increases them for multi-label models, driven by disproportionate pruning of artefacts (rightmost column).

little systematic study; a structured overview of ML for Earth observation through this lens would be of practical value to practitioners and researchers alike.

7 Conclusion

In this work, we investigated false positives in methane plume detection in TROPOMI observations using end-to-end ML pipelines. The false positives are driven by hard negative examples, which are difficult to model because their discriminating feature appears conditionally in a subpopulation of the dataset. We dub this type of feature as a marginalised discriminant, i.e. a feature whose discriminative signal is only present conditionally in a subpopulation of the training data, and therefore does not reliably appear during learning. In this particular application domain, we investigated multiple model- and data-centric mitigation measures to learn the marginalised discriminant: (i) increasing the inductive bias through architectural changes based on properties of the problem, and (ii) leveraging auxiliary domain knowledge to automatically prune training instances with potentially noisy labels. Additionally, we interpreted the properties of a domain pipeline that successfully addresses the false positive problem. We found that successful mitigation measures increase data efficiency by substantially limiting the complexity of the machine learning problem, suggesting that operational success depends not just on sophisticated learning algorithms but also on simplifying the problem as much as possible to match the available data. More broadly, this principle may extend to automated pipeline design for other data-limited problems: in such settings, automating the search over well-

designed knowledge-driven components may be more tractable than end-to-end data-driven automation.

Acknowledgments. We thank Tobias de Jong, Berend Schuit, and Thomas Newman for fruitful discussions. This research has been partly supported by: the "Physics-aware Spatio-temporal Machine Learning for Earth Observation Data" project financed by the Dutch Research Council (OCENW.KLEIN.425); the Open Space Innovation Platform (https://ideas.esa.int) as a co-sponsored research funded by the European Space Agency (4000136204/21/NL/GLC); and an Alexander von Humboldt Professorship in Artificial Intelligence awarded to Holger Hoos. Sentinel-5 Precursor is part of the EU Copernicus program, and Copernicus (modified) Sentinel-5P data (2018—2023) have been used.

References

1. Arnob, A.K.B., Mridha, M.F., Alfarhood, M., Che, D., Safran, M.: A convolutional neural network-based segmentation approach for oil spill detection in satellite imagery with explainable AI. Earth Science Informatics **18**(3), 499 (2025). https://doi.org/10.1007/s12145-025-01988-6
2. Chen, P., Ren, Y., Zhang, B., Zhao, Y.: Class Imbalance in the Automatic Interpretation of Remote Sensing Images: A Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 9483–9508 (2025). https://doi.org/10.1109/JSTARS.2025.3555567
3. Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., Japkowicz, N.: The class imbalance problem in deep learning. Machine Learning **113**(7), 4845–4901 (2024). https://doi.org/10.1007/s10994-022-06268-8
4. Hasekamp, O., Lorente, A., Hu, H., Butz, A., Aan de Brugh, J., Landgraf, J.: Algorithm Theoretical Baseline Document for Sentinel-5 Precursor methane Retrieval, SRON The Netherlands Institute for Space Research, Leiden, the Netherlands (2022), https://sentinels.copernicus.eu/documents/247904/2476257/Sentinel-5P-TROPOMI-ATBD-Methane-retrieval.pdf
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
6. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9), 100804 (2023). https://doi.org/10.1016/j.patter.2023.100804
7. Kurchaba, S., Maasakkers, J.D., Schuit, B.J., Aben, I.: Explainable Comparison of Feature-Based and Deep Learning Models for TROPOMI Methane Plume Screening. arXiv preprint arXiv:2605.27236 (2026)
8. Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Ben-Tzur, J., Hardt, M., Recht, B., Talwalkar, A.: A System for Massively Parallel Hyperparameter Tuning. In: Proceedings of Machine Learning and Systems. vol. 2, pp. 230–246 (2020)
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
10. Loshchilov, I., Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts. In: International Conference on Learning Representations. vol. 3, pp. 1769–1784 (2017)

11. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations. vol. 6, pp. 4061–4078 (2019)
12. McGovern, A., Bostrom, A., McGraw, M., Chase, R.J., Gagne, D.J., Ebert-Uphoff, I., Musgrave, K.D., Schumacher, A.: Identifying and Categorizing Bias in AI/ML for Earth Sciences. Bulletin of the American Meteorological Society **105**(3), E567–E583 (2024). https://doi.org/10.1175/BAMS-D-23-0196.1
13. Mifdal, J., Longépé, N., Rußwurm, M.: Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **3**, 285–293 (2021)
14. Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M.I., Stoica, I.: Ray: A distributed framework for emerging AI applications. In: Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation. pp. 561–577. OSDI'18, USENIX Association, USA (2018)
15. Northcutt, C., Jiang, L., Chuang, I.: Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021). https://doi.org/10.1613/jair.1.12125
16. Růžička, V., Mateo-García, G., Irakulis-Loitxate, I., Johnson, J.E., Martín, M.M.S., Allen, A., Raunak, A., Castaneda, C., Guanter, L., Thompson, D.R.: Operational machine learning for remote spectroscopic detection of CH_4 point sources. arXiv preprint arXiv:2511.07719 (2025)
17. Schuit, B.J., Maasakkers, J.D., Bijl, P., Aben, I.: Training datasets with manually labeled TROPOMI data for Machine Learning models (2025). https://doi.org/10.5281/zenodo.13903868
18. Schuit, B.J., Maasakkers, J.D., Bijl, P., Mahapatra, G., van den Berg, A.W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D.J., McKeever, J., Jervis, D., Girard, M., Irakulis-Loitxate, I., Gorroño, J., Guanter, L., Cusworth, D.H., Aben, I.: Automated detection and monitoring of methane super-emitters using satellite data. Atmospheric Chemistry and Physics **23**(16), 9071–9098 (2023). https://doi.org/10.5194/acp-23-9071-2023
19. Wąsala, J., Maasakkers, J.D., Aben, I., Schneider, R., Hoos, H., Baratchi, M.: Mitigating representation bias caused by missing pixels in methane plume detection. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 325–334. Springer Nature Switzerland, Cham (2026)
20. Wąsala, J., Maasakkers, J.D., Schuit, B.J., Leguijt, G., Aben, I., Schneider, R., Hoos, H., Baratchi, M.: AutoMergeNet: AutoML-based M-Source Satellite Data Fusion Evaluated with Atmospheric Case Studies. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025). https://doi.org/10.1109/JSTARS.2025.3621068
21. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01548
22. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, vol. 11211, pp. 3–19. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1