This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3621068

AutoMergeNet: AutoML-based M-Source Satellite Data Fusion Evaluated with Atmospheric Case Studies

Julia Wąsala, Joannes D. Maasakkers, Berend J. Schuit, Gijs Leguijt, Ilse Aben, Rochelle Schneider, Holger Hoos, Mitra Baratchi

***Abstract*—Accurate detection of anomalous phenomena in satellite data often requires data layers containing complementary information (e.g., data from different sensors, auxiliary features such as land cover maps, and metadata regarding data quality). However, existing highly specialised approaches to fuse multiple data layers cannot be transferred to other related problems, as they rely on expert-selected features and manual pipeline design. In this work, we propose *AutoMergeNet*, a framework for satellite image data fusion based on Neural Architecture Search (NAS). AutoMergeNet generates neural networks that fuse any number of raster data layers. Consequently, it can address different classification problems based on satellite images without manual pipeline design. We designed the search space of AutoMergeNet by identifying relevant design choices from the image classification and data fusion literature. AutoMergeNet automatically transforms image classification networks into multi-branch networks by optimising critical architectural and training hyperparameters. Since the high dimensionality of multimodal image data poses a challenge for data fusion problems with limited labels, we use an auxiliary unimodal classifier combined with AutoMergeNet. We evaluate AutoMergeNet on a methane plume detection dataset from the literature and our newly created carbon monoxide plume detection dataset. AutoMergeNet performs strongly and consistently on these two multimodal classification problems, outperforming six baseline methods selected from state-of-the-art image classification approaches. Finally, we demonstrate the usability of our framework with a realistic methane plume detection use case, which shows that AutoMergeNet can be used as a highly specialised, state-of-the-art approach.**

***Index Terms*—Neural Architecture Search, Multimodal Image Data Fusion, Earth Observation, Atmospheric Plume Detection**

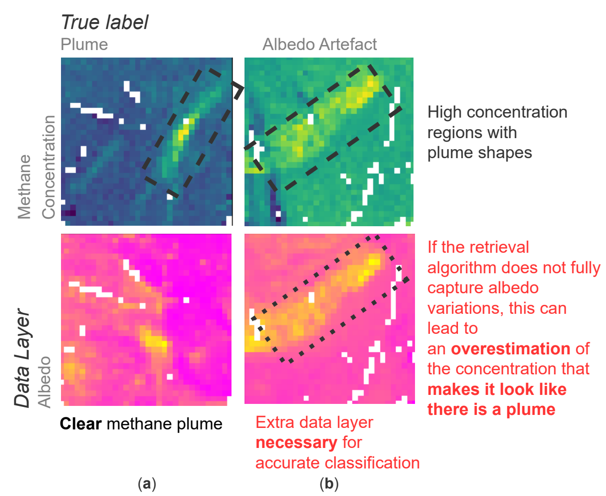


Fig. 1. Accurate methane plume detection requires consideration of multiple modalities. Columns **(a,b)** show the methane concentration and albedo (surface reflectivity) at two locations, as observed by the TROPOMI satellite instrument: a true plume **(a)** and an albedo artefact **(b)**. Both methane concentration fields seem to contain a plume, but closer inspection of the albedo shows that column **(b)** is an artefact caused by high surface reflectivity. Variations in albedo can lead to overestimation of the methane concentration, if the variation is not fully captured by the retrieval algorithm. The albedo data layer is necessary to discover the correlation and correctly classify this observation as "not plume." Similar artefacts can occur in other data layers; therefore, more than two data layers are necessary.

I. Introduction

A common problem in the Earth Observation (EO) domain is the detection of a specific type of anomaly (e.g., oil spills on the surface of the sea or methane plumes in the atmosphere). This problem is often addressed by training a binary classifier, using a dataset where each image is labelled as either positive, showing "the anomaly of interest" or negative, depicting "anything else". This approach tends to misclassify phenomena in the satellite data that appear similar to the anomaly of interest. These phenomena, called "artefacts", lead to false positives through different mechanisms. For instance, algae on water can resemble oil spills in synthetic aperture radar (SAR) data, and variations in surface albedo can be erroneously interpreted as gas enhancements in the retrieval of gas concentrations from radiance spectra and be mistaken for methane plumes [1] (see Figure 1).

Multimodal image data from different sources or measuring different physical quantities can reduce false positives. While both unimodal and multimodal image data can contain multiple channels (e.g., ImageNet's RGB channels [2]), they differ fundamentally. RGB channels are tightly linked, because they are obtained using a single sensor measuring one physical phenomenon—the number of photons collected per pixel. The

Julia Wąsala, Holger Hoos and Mitra Baratchi were with the Leiden Institute for Advanced Computer Science (LIACS), Leiden University. Holger Hoos was additionally with the Chair for AI Methodology (AIM) at RWTH Aachen. Joannes D. Maasakkers, Berend J. Schuit, Gijs Leguijt and Ilse Aben were with the SRON Space Research Organisation Netherlands. Berend J. Schuit was additionally with GHGSat Inc; Gijs Leguijt with the Department of Climate, Air and Sustainability at the Netherlands Organisation for Applied Scientific Research, TNO; Ilse Aben with the Department of Earth Sciences, Vrije Universiteit Amsterdam. Rochelle Schneider was with the Φ-Lab, ESA-ESRIN

Manuscript received June 24, 2025; revised September 18, 2025; accepted September 28, 2025.

channels inherit sensor biases, such as damaged pixels, and record information simultaneously. In contrast, multimodal image data are linked by time and location, but different sensors and physical phenomena provide multiple views of the subject with additional information and independent biases. Existing approaches show that multimodal image data can help reduce false positives by leveraging discerning features extracted from auxiliary data [1], [3], [4]. In methane plume detection, concentrations are inferred from satellite-measured spectra through "*retrieval*". Data used or produced in the retrieval (such as albedo), meteorological data and outputs of physical models are subsequently used to reduce the number of false positives. Similarly, oil spill detection benefits from including data from Geographical Information Systems (GIS) [5]. However, existing approaches are not easily generalisable, as they address artefacts on a case-by-case basis, relying on problem-specific features or models that do not transfer to other datasets, requiring new pipelines for each application.

A more general solution is to fuse original raster data using deep learning rather than relying on manual feature engineering. An effective deep learning framework applicable to a wide range of tasks needs to: (i) model complex relationships between variables; (ii) fuse arbitrary numbers of data layers; (iii) address high dimensionality from additional data layers; and (iv) rely on minimal assumptions for transferability to related problems. These properties are challenging to achieve due to the significant domain and ML knowledge required for specialised architectures. An automated approach to designing such models would significantly increase accessibility for domain experts with limited deep learning experience.

We propose an Automated Machine Learning (AutoML) framework for detecting anomalies in artefact-prone satellite data products. AutoML (see, e.g., [6], [7]) research focuses on automating high-performance ML system design configuration to enhance its accessibility. Neural Architecture Search (NAS) is a subarea of AutoML that addresses the model design problem by focusing on automatic neural network design through optimising architectural hyperparameters (i.e., connections and operations between neurons).

In this paper, we introduce *AutoMergeNet*, a NAS system that automatically designs image data fusion networks for M modalities based on a search space of neural network design choices (addressing properties (i), (ii), and (iv)). We created this search space by identifying relevant design choices from existing image classification and data fusion literature. To address the high dimensionality of the problem (property (iii)) when limited labelled data is available, we propose using a two-step approach, which filters the images on the primary modality only to remove clear negative images and uses our NAS-generated fusion network to discern the true positive images from the false positive ones. Our contributions are as follows:

- We propose *AutoMergeNet*, the first NAS system to automatically create multi-branch data fusion networks for any number of modalities by optimising critical hyperparameters, with an auxiliary classifier using single data layers to reduce the dimensionality problem.[1]
- We evaluated and showed the strong performance of AutoMergeNet on two high-impact applications (i.e., detection of methane and carbon monoxide plumes), significantly outperforming six baselines.
- For evaluation, we collected and labelled a new benchmark dataset for carbon monoxide plume detection, presenting the first ML approach to carbon monoxide plume detection and the detection of plumes of multiple gases.
- We demonstrate the operational performance of the best methane plume detection model found by *AutoMergeNet* on a realistic use case. We further compare with the highly specialised model currently in operation, proposed by Schuit et al. [1].

II. Related work

In this section, we describe related work in the field of multimodal satellite data fusion and AutoML.

Computer vision approaches: Our research is highly related to general image classification. Existing neural networks for image classification have varying architectures, mostly consisting of repeating blocks or cells. These blocks are continually refined using network architectures such as ResNet [8] (e.g., BANet [9], EPSANet [3]), MobileNetv2 [10] and CVT [11] as backbones. Multi-image fusion is another related image task that aims to either improve visual image quality by fusing two images of the same subject with different image qualities (e.g., under- and over-exposed [12]) or to improve downstream performance on tasks such as object detection [13], [14]. These images are single modality but provide additional information, for instance, because they show the subject from different angles or lighting conditions. The networks used in these contexts tend to be customised to the task at hand instead of using common backbones.

Multimodal image fusion approaches for EO: Inspired by computer vision approaches for natural images, the fusion of two image modalities (e.g., optical and synthetic aperture radar (SAR) images) has been shown to improve EO tasks, including land use/land cover mapping (LULC). Unlike natural image fusion, where images are from the same modality, EO data layers (representing distinct modalities) can have different distributions, thereby complicating feature extraction and convergence. Treating multimodal data as a single cube increases the risk of overfitting [15]. Consequently, existing work often uses two-branch networks with separate feature extraction paths [16]–[18]. Data fusion studies have explored optimal fusion strategies and depths. Hong et al. [16] compared concatenation, addition, and using cross-connections at different network depths, while Audebert et al. [17] compared early and late fusion. Both studies confirmed that the optimal choice of fusion depth and strategy is task-dependent.

While most EO image fusion approaches only fuse two modalities, using additional modalities may further improve performance [19]. For instance, Marjani et al. [20] combine 9 channels relevant to methane plume detection using a vision transformer. However, each new modality increases data

[1]The code is available at https://github.com/ADA-research/AutoMergeNet.

dimensionality, creating training challenges when labelled data is limited. Consequently, many approaches that fuse more than two modalities rely on manual feature engineering. For example, Schuit et al. [1] combine Convolutional Neural Network (CNN)-based filtering with Support Vector Machine (SVM) classification using hand-crafted features based on auxiliary data for methane plume detection. We propose a more general framework that is less dependent on domain knowledge by replacing manual feature extraction with automated M-modal neural networks.

NAS for multimodal EO data: NAS frameworks consist of three components: a search space of design choices (i.e., all relevant hyperparameters) of neural network architectures and training algorithms, a search strategy that efficiently explores and samples architectures from the search space, and a performance evaluation procedure for assessing the performance of candidate architectures (see, e.g., [7]). Developing NAS systems tailored to EO data has become increasingly popular to overcome the problem of the time and expertise required to design neural networks. Recent work has extended AutoKeras [21] for satellite image classification [22] and super-resolution [23], but these approaches remain limited to unimodal data.

Existing NAS approaches for multi-image fusion focus on designing two-branch networks for the fusion of two modalities [24], [25]. Li et al. propose a NAS framework for LULC that optimises the micro-architecture of the block that blends the features of the input branches [24]. More recently, Li et al. propose a framework that additionally optimises the architectures of the input branches [25]. Similarly, Feng et al. propose a micro-level NAS framework for fusion of hyperspectral and LiDAR data, optimising the architectures of the input branches and the fusion module [26]. These approaches mainly focus on the micro-level design of the neural network's building blocks and do not optimise other important hyperparameters, such as the fusion depth. Furthermore, they fix the number of fused modalities to two, restricting the application domain of their approaches.

Here, we extend the application domain of previous work by extending the two-branch architecture popular in data fusion to M-modality fusion, and we automatically optimise the fusion depth and fusion strategy. To reduce the dimensionality of the problem when limited labelled data is available, we use a two-step pipeline that (i) filters the images based on the primary modality and (ii) uses deep learning for automated multimodal feature extraction, avoiding problem-specific manually engineered features while still including sufficient information to reduce the number of artefacts.

III. AutoMergeNet

Multimodal image data fusion is the task of learning a joint representation from M co-located source images $I^m \in \mathbb{R}^{W \times H \times C_m}$, where $m \in \{1, \ldots, M\}$, W and H are the spatial dimensions width and height, and C_m is the number of channels in the input image. We define our problem as a binary satellite image classification task that requires multiple data layers to reduce false positives. These data layers can be divided into (i) the primary data layer I^1 showing the class of interest and (ii) the supporting data layers $\{I^2, \ldots, I^M\}$, which include information about the potential artefacts.

We propose a trial-based NAS system for fusion of multimodal image datasets with any number of data layers M. The system automatically designs multi-branch neural networks using a primary data layer (showing the class of interest) and supporting data layers (including information about the potential artefacts). The high dimensionality of the data can reduce generalisation performance. Therefore, we introduce an auxiliary classifier that initially filters the images using only the primary data layer. This combination enables automated solutions to M-modal problems, whereas previous NAS approaches were limited to two modalities; thus, a larger set of problems can be solved using our system. We describe our NAS framework and auxiliary classifier filtering below.

A. Multimodal image data fusion with NAS

A trial-based NAS system iteratively searches for a high-performance architecture. Starting from a random architecture, at each trial, the search strategy samples a new candidate architecture from the search space of relevant hyperparameters (see Section III-A1). The performance evaluation strategy evaluates this architecture (see Section III-A2). At the next trial, the search strategy uses the performance of the previously evaluated candidate architectures to select a new, promising architecture. The final candidate architecture can be selected at any trial when a certain number of trials have been completed or a minimum performance threshold has been met. Our focus in formulating a NAS problem is on designing the search space to ensure effective network architectures can be efficiently generated. The following sections describe our search space and the search and evaluation strategies.

1) Search space: Multi-branch networks for data fusion: We base *AutoMergeNet* on two-branch data fusion networks and extend this concept to multiple modalities, building multi-branch networks with M input branches. Multi-branch networks extract multimodal features from the data in two steps: (i) extracting independent features from each modality, (ii) combining the independent features into a joint feature set for classification. Single-branch networks, however, consider the modalities as a single data cube of dimensions $\sum_{m=0}^{M} H \times W \times C_m$ and extract joint features from the start. This architecture may be suboptimal as the differing distribution of data sources can make it harder to extract meaningful features, possibly hurting convergence. Extracting features independently can allow the model to learn a normalisation for each data layer, making fusing information from data layers with different distributions possible. This ability is crucial for multimodal image problems, where the data layers are of a different nature, and distributions of a single data layer can even vary from image to image. As a result, improper normalisation or fusion of the features can cause loss of information or may over-emphasise confounding features with large value ranges, leading to increased false positive detections. Therefore, multi-branch architectures are an effective approach to reducing

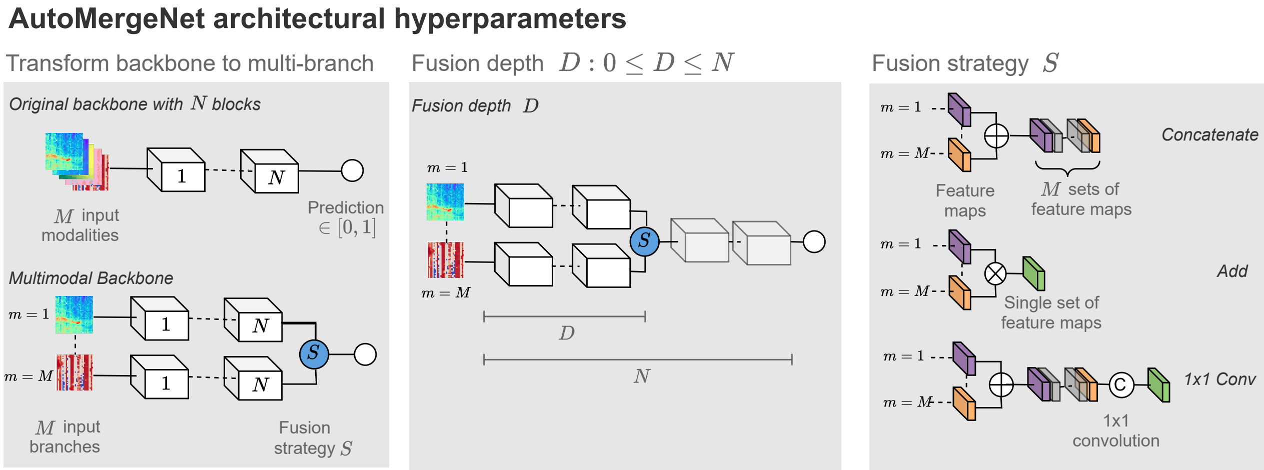


Fig. 2. The architectural hyperparameters optimised by AutoMergeNet. Compared to previous work, our NAS data fusion approach extends the application domain to more than two modalities and automatically tunes the merge depth. (**Left**) The *backbone architecture* consists of N blocks and is transformed into multimodal image networks by creating identical but independent branches for each modality. (**Center**) The *fusion depth* D determines at which stage in the network a fusion strategy S fuses the branches into a single branch. The branches are fused right before the output layer if the fusion depth equals the number of modules in the network. (**Right**) The *fusion strategy* determines how the input branches are fused: concatenating the feature maps of all branches (concatenation); summing the feature maps (addition), or concatenating the feature maps and dividing the number of maps by the number of modalities M using a 1×1 convolution to restore the original number of feature maps (convolution).

false positive detections in multimodal image tasks with many modalities.

We introduce a simple heuristic to transform modular (consisting of repeating, identical blocks) single-branch neural networks into multi-branch networks by making copies of parts of the network, creating a multi-branch network with identical but independent branches to extract independent features from each modality (Figure 2, left). The architecture of these networks is governed by three critical architectural hyperparameters: the choice of backbone, fusion depth and fusion strategy. The choice of backbone determines the architecture of the blocks $\mathbf{B}$ applied on the input I and the total depth of the network N (with $\mathbf{B}_D^m$ representing the architecture of the block for modality m at depth D). The fusion depth D determines where in the network the features are fused (Figure 2, centre). Given inputs I^m, the independent feature sets for each modality F^m at depth D (i.e., F_D^m) are obtained as:

$$F_D^m = \begin{cases} I^m & \text{if } D = 0 \\ \mathbf{B}_D^m(I^m) & \text{if } D = 1 \\ \mathbf{B}_D^m(F_{D-1}^m) & \text{if } D = 2, \ldots, N, \end{cases} \quad (1)$$

where $0 \leq D \leq N$; $D = 0$ corresponds to early fusion and $D = N$ to late fusion. Finally, the fusion strategy determines how $F_D^1, \ldots, F_D^M$ are fused to obtain a joint feature set F'_D (Figure 2, right) by concatenating, adding or convolving them (see, e.g., [16]). Concatenation retains features from individual branches, while scaling the fused branch parameters by a factor of M for the maximum merge depth ($D = N$), and by a smaller factor for smaller merge depths ($D < N$). Due to the parallel branch structure of multi-branch networks, they have more trainable parameters than early fusion networks and are more expressive as a result. Addition directly blends the features and does not increase the number of trainable parameters. Finally, inspired by attention modules, we propose a new fusion strategy which combines the advantages of concatenation and addition: 1×1 convolution. This operation learns a mapping of the branches while only slightly increasing the number of trainable parameters compared to addition. F'_D is then defined as

$$F'_D = \begin{cases} [F_D^0, \ldots, F_D^M] & \text{if } S = \text{concatenate}, \\ \sum_{m=0}^{M} F_D^m & \text{if } S = \text{add}, \\ f([F_D^0, \ldots, F_D^M]) & \text{if } S = 1 \times 1 \text{ convolution}, \end{cases} \quad (2)$$

where f is a learnable 1×1 convolution function. The final representation used for classification is obtained by applying the remaining blocks $\mathbf{B}_{D+1}, \ldots, \mathbf{B}_N$ to F'_D. The network depth N is fixed and not tuned by AutoMergeNet. Our search space also contains the key training hyperparameters (learning rate, dropout, weight decay and batch size), which can significantly impact final network performance. Following this procedure, AutoMergeNet overcomes the issues of single-branch networks for multimodal image data fusion while leveraging state-of-the-art vision architectures.

2) Search and performance evaluation strategy: The types of search strategies used for NAS range from simpler techniques, such as random or grid search, to state-of-the-art approaches, such as Bayesian and gradient-based optimisation [7]. Bayesian optimisation approaches build an internal model of the performance of the architectures evaluated during search. Using such a model, it is possible to iteratively select new architectures based on their estimated performance.

AutoMergeNet uses BOHB [27], a well-known recent search strategy based on Bayesian optimisation, to effectively explore the search space. BOHB uses the validation loss of

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3621068

each configuration to update the internal model the search strategy uses to sample new configurations. The best candidate network is selected based on validation loss at the end of the search process. BOHB combines the strong performance of Bayesian optimisation and the computational efficiency of Hyperband [28]; the latter is a performance evaluation strategy that can be used for estimating the performance of sampled networks with a reduced computational budget by stopping the training of networks with high validation losses early. Furthermore, the performance of BOHB is relatively insensitive to the value of its hyperparameters [27]. In principle, other search and evaluation strategy combinations could be used: the structure of our search space is the core of our approach.

B. Auxiliary unimodal classifier

In many EO anomaly detection problems, a single data layer is the primary source of information (e.g., the methane concentration in methane plume detection), and the other data layers provide supplementary information to reduce false positives. Subsequently, the contribution of different data layers to multimodal anomaly detection problems is imbalanced. A downside of the multi-branch approach is that each data layer is treated equally, and the network can only learn a hierarchy of the input modalities from the training data. However, modelling data layer utilisation is challenging because multimodal problems are inherently high-dimensional and often have relatively few labelled instances. The supporting data layers may disproportionally contribute to the model predictions and overwhelm the features extracted from the primary layer. Paradoxically, this causes the model to misclassify instances where the primary data layer clearly shows the image is of the negative class. A simple solution to this problem is to add an auxiliary classifier to discern clear negative images based on the primary modality only, thereby reducing false positives through a multimodal approach [1]. The final AutoMergeNet predictions combine the predictions of the auxiliary classifier and the multimodal model (see Figure 3). All images classified as negative by the auxiliary classifier are assigned a negative label, and the remaining images are assigned labels according to the predictions of the multimodal model.

IV. Empirical Evaluation Setup

Our empirical evaluation aims to answer the following questions:

- **Q1** Does transforming conventional, single-branch image classification networks into multi-branch networks improve the results of multimodal image data fusion for satellite image classification?
- **Q2** Does enforcing implicit focus on the primary modality (using a unimodal auxiliary classifier) improve the results of multimodal image data fusion for satellite image classification?
- **Q3** How do AutoMergeNet-created models compare to domain-specific methods when applied in an operational scenario?

In the following, we describe the data used to evaluate the performance of AutoMergeNet, our procedure for selecting and configuring baselines, and the setup of our empirical evaluation.

A. Data

To critically assess our methods, we need image classification datasets with (i) the necessary modalities for accurate classification of complex concepts, such as plumes, and (ii) labels for the concept to be detected. We evaluated the performance of AutoMergeNet on two datasets for atmospheric anomaly detection from EO (which are publicly available on Zenodo [2]). We used the images from the methane plume dataset created by Schuit et al. [1] for automated feature extraction and additionally created a new dataset for the detection of carbon monoxide plumes.

All images were obtained from the TROPOspheric Monitoring Instrument (TROPOMI), a spectrometer aboard ESA's Sentinel-5P satellite that provides daily global observations of concentrations of different trace gases at up to $7 \times 5.5\text{km}^2$ resolution for the gases considered here [29]. The TROPOMI Level-2 data products include in- and outputs of the retrieval algorithms used to calculate the atmospheric concentrations of gases in each pixel [30]–[32]. The datasets are multimodal, as each data layer represents a different physical quantity, and some algorithm inputs are obtained from sensors on board of other satellites (e.g., the Moderate Resolution Imaging Spectroradiometer (MODIS) or Visible Infrared Imaging Radiometer Suite (VIIRS)). Domain knowledge is crucial to create these datasets for two reasons: it is necessary to know where plumes are likely to occur, because only a small number of images contain plumes, and the mechanisms leading to false positives are highly problem-specific and require detailed knowledge of the problem. Figure 4 shows the steps for creating both datasets.

1) Methane plumes: We used the methane plume detection datasets from Schuit et al. [1], available on Zenodo, version 1.0.0 [33]. They created and manually verified two separate datasets, one for training each stage of their pipeline. We merge these datasets and remove duplicate images included in both datasets. Each training instance is a datacube consisting of the 11 channels (Table I) used as input for the feature engineering step in their approach. Following Schuit et al., we normalised the methane data layer per image. The normalisation approach proposed by Schuit et al. [1] is custom-made to maintain information about both spatial patterns and plume magnitude. The remaining data layers were standardised feature-wise using the training set mean and standard deviation. We partitioned the image set into 64% training, 16% validation, and 20% testing data. We used the same simple geometric data augmentation as Schuit et al. (rotating by 90° and flipping), but instead of creating a larger dataset consisting of all augmented images, we randomly apply the augmentations at each training epoch. The dataset consists of 3888 images of $32 \times 32 \times 11$ pixels (height $\times$ width $\times$ number of channels). 32% of the images are positives.

[2]Carbon monoxide: https://zenodo.org/records/17226619, methane: https://zenodo.org/records/13903869.

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3621068

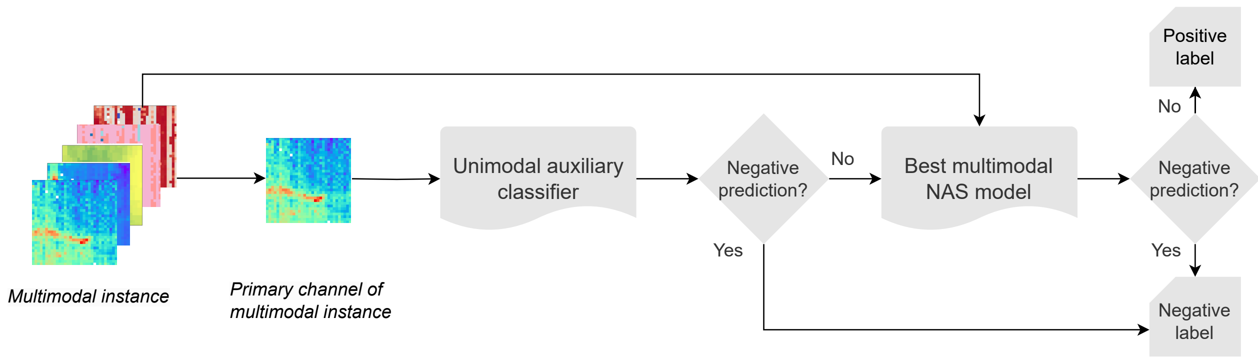


Fig. 3. The AutoMergeNet pipeline obtains final predicted labels by combining the predictions of the multimodal NAS-created model and the unimodal auxiliary classifier that only uses the primary modality. If the auxiliary classifier predicts a negative label, the instance is labelled as negative. Otherwise, the label predicted by the multimodal model is used.

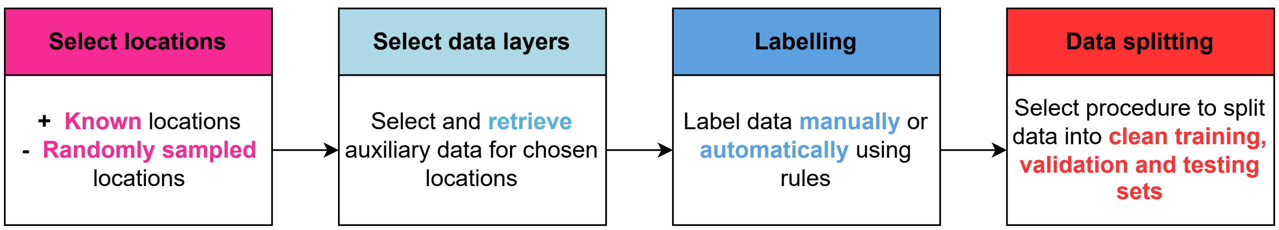


Fig. 4. The four-step workflow for creating a plume detection dataset. (i) Selection of locations: positives (plumes) are selected based on known locations of large emitters, and negatives are sampled randomly. Artefacts may be selected manually. (ii) Identification of relevant data layers. (iii) Image verification: manual labelling (part of) the images (methane dataset) or heuristic-based labelling using manually-defined filters (carbon monoxide dataset). (iv) Split into training, validation, and testing data, assigning overlapping images to the same partition.

TABLE I
THE CHANNELS IN THE DATASET BASED ON SCHUIT ET AL. [1], INCLUDING THE SOURCE OF EACH CHANNEL: TROPOMI IF IT IS AN OUTPUT OF THE RETRIEVAL PROCESS OR A REFERENCE IF IT IS OBTAINED FROM AN EXTERNAL SOURCE.

	Description	Source
1	Methane	TROPOMI
2	Surface pressure	ECMWF, GMTED2010 [34]
3	Albedo (SWIR)	TROPOMI
4	Aerosol optical thickness (SWIR)	TROPOMI
5	Data quality assurance	TROPOMI
6	Cloud fraction (custom)	VIIRS [35]
7	Pixel surface area	TROPOMI
8	χ^2 of the retrieval	TROPOMI
9	Landflag	MODIS [36], [37]
10	Wind v_{10}	ERA5 [38]
11	Wind u_{10}	ERA5 [38]

TABLE II
THE CHANNELS INCLUDED IN THE CARBON MONOXIDE DATASET, INCLUDING THE SOURCE OF EACH CHANNEL: TROPOMI, IF IT WAS AN OUTPUT OF THE RETRIEVAL PROCESS, OR A REFERENCE, IF IT WAS OBTAINED FROM AN EXTERNAL SOURCE.

	Description	Source
1	Carbon Monoxide	TROPOMI
2	Carbon Monoxide Precision	TROPOMI
3	Data quality assurance	TROPOMI
4	Geolocation flag	TROPOMI
5	Surface Pressure	ERA5 [38], GMTED2010 [34]
6	Degrees of Freedom	TROPOMI
7	Height Scattering Layer	TROPOMI
8	Scattering optical thickness (SWIR)	TROPOMI
9	Surface Albedo 2335 nm	TROPOMI
10	Ground pixel	TROPOMI

2) Carbon monoxide plumes: We created a new carbon monoxide plume dataset, selecting ten channels known to be relevant to carbon monoxide detection (Table II). The positive instances were selected starting with plume locations of emissions from African cities by Leguijt et al. [39]. The dataset contains multiple plumes per location, detected on different dates. We randomly selected negative images from the African continent to ensure image diversity and inclusion of artefacts. The number of negative images, almost 4000, is too large to label manually. Thus, we filtered the data to minimise the likelihood of including plumes. The dataset consists of 5207 images of $32 \times 32 \times 10$ pixels (height $\times$ width $\times$ number of channels). 24% of the images are positives. We created location-based testing (see, e.g., [16], [19]) set by including images within a circle that encompasses approximately 20% of the data and approaches a stratified split; the set of remaining images was split into 80% training and 20% validation data. This procedure provides realistic performance estimates of generalisation to unseen regions, as it ensures that unique geographic features, such as rivers, cannot be used by the model to learn the concept of a plume. Finally, we normalised and augmented the data following the procedure from the methane plumes dataset.

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3621068

B. Baseline selection

Selecting baselines for M-modal data fusion is challenging due to the absence of end-to-end deep learning architectures for this task. From a practitioner's perspective, two options exist: (i) extensively modifying dual-modality architectures (e.g., [16]–[18], [24], [40]) for M modalities, as the optimal fusion depth and fusion strategy can change as the number of modalities (and, therefore, dimensions) increases, or (ii) applying standard image recognition networks (e.g., ResNet [8], MobileNetV2 [10], CvT [11]) with early fusion (by treating the M-modal data as a data cube). The first approach requires fundamental architectural redesign that goes beyond baseline evaluation and into method development. The second requires only hyperparameter tuning — a standard practice for fair baseline comparisons. Therefore, we select early fusion as our state-of-practice, representing the only viable approach currently available to practitioners working with M-modal data. As baseline architectures, we selected the state-of-the-art image classification and multimodal image data fusion models that we also used as backbones in the search space of *AutoMergeNet*:

- **Early fusion CNN** [16]: one of the fusion CNNs proposed by Hong et al. [16]. Their approach takes steps towards a unified framework for data fusion through the analysis of different combinations of fusion depth and fusion strategy.
- **ResNet18** [8]: an architecture with a significant role in the development of deep neural networks through its introduction of skip connections. ResNet is often used as a baseline in the evaluation of EO approaches [41]–[43] and is still used as a backbone network for the development of new blocks and modules (see, e.g., [9], [44], introduced below) or as part of the search space in NAS for EO [22].
- **EPSANet** [44]: introduces a new channel attention mechanism using pyramid blocks that can extract multi-scale information, often present in EO imagery.
- **BANet** [9]: introduces another channel attention mechanism, bridging previous features to implement attention with fewer parameters but higher performance than other attention mechanisms.
- **MobileNetV2** [10]: uses depth-wise separable convolutions to reduce memory usage, an important consideration in the design of neural networks for EO [45].
- **CvT** [11]: introduces convolutions to the transformer architecture, which allows combining the strengths of CNNs and transformers.

We used reference implementations by the original authors as much as possible (CvT[3], Mobilenetv2[4], ResNet[5], BANet[6], EPSANet[7]). We reimplemented the original Tensorflow v1 Early fusion CNN [16] in PyTorch for consistency with the other architectures. We manually tuned the model depth of each baseline. When multiple model sizes were available, we chose the smallest network, as preliminary experiments showed that larger networks systematically achieved lower performance while requiring more computational resources. We optimised the training hyperparameters separately for each dataset, as described in the following section. Hyperparameter optimisation can significantly improve the performance of neural networks and is critical when the network is applied to new datasets. Some architectures initially chosen as baselines, such as Swin Transformer [46], were omitted because of poor or unstable performance despite hyperparameter optimisation.

C. NAS and HPO setup

We optimised the training hyperparameters of the baseline models for each dataset, running five independent hyperparameter optimisation (HPO) runs per architecture with 100 trials to optimise the baselines' dropout, weight decay, learning rate, and batch size (Table III). We optimised the validation loss using Ray Tune's random search in combination with the ASHA scheduler [47].

To find an architecture with AutoMergeNet, we ran 30 independent NAS runs with 200 samples, optimising the loss on the validation set. The number of AutoMergeNet runs was calculated to match the number of runs of the baselines: AutoMergeNet's search space contains all six baseline architectures, which were trained and evaluated with five runs each, leading to a total of 30 runs for AutoMergeNet. Table III lists the hyperparameters in the search space. The maximum fusion depth depends on the architecture, as shown in Table III. We used the BOHB implementation offered in the Ray Tune hyperparameter optimisation package [48].

1) Auxiliary classifier setup: We used the CNN proposed by Schuit et al. [1] as our auxiliary classifier. We reimplemented the author's Keras code in PyTorch to ensure consistency with the baseline architectures. We obtained prediction scores for each image in the dataset using the training procedure for stacking ensembles [49]: we split the training and validation sets into parts called A and B. First, we trained the model with training set A and early stopping on validation set A, and we predicted the labels of training set B and validation set B. We repeated this with training on sets B to predict sets A. Splitting the training into two sets is necessary to obtain unbiased labels for the training and validation sets. Finally, we trained the model from scratch, using the entire training and validation set to predict labels for the testing set. We repeated this procedure five times for each dataset. We saved the predictions of the model with the lowest validation loss. The predicted scores are combined with the predictions of the multimodal networks generated by AutoMergeNet to produce final predicted labels as described in Section III and Figure 3.

2) Model evaluation: All experiments were run on the Leiden University GRACE computing cluster, which consists of 26 homogeneous CPU nodes containing 94 GBs of memory and using Intel Xeon E5-2683 v4 CPUs running at 2.10GHz and 9 homogeneous GPU nodes each equipped

[3]https://github.com/microsoft/CvT
[4]https://github.com/pytorch/vision/blob/main/torchvision/models/mobilenetv2.py
[5]https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py
[6]https://github.com/zhaoy376/Bridge-Attention
[7]https://github.com/murufeng/EPSANet

with 2 NVIDIA GeForce GTX 1080Ti. The OS is CentOS-7. We selected best configurations based on the validation loss. We trained each best configuration five times with early stopping on the validation set with a patience of 10 epochs. We trained the models with AdamW [50], using loss weights and setting the weight of the positive class as $\frac{N_-}{N_+}$, where N_- is the number of negative examples, and N_+ is the number of positive examples in the dataset. Directly calculating the mean and standard deviation will likely yield results susceptible to outliers and noise. Therefore, to ensure the robustness and representation of our statistical results, we simulated selecting the single best model based on the validation loss with bootstrapping. We bootstrapped the accuracy, precision, and recall with 1000 samples and saved the test score of the model with the highest validation score in each bootstrap sample. We calculated the mean and standard deviation of the test scores of these models, yielding a robust estimate of the performance of the best model obtained from multiple runs.

V. RESULTS

In this section, we present the results of the application of AutoMergeNet to two M-modal fusion problems and compare with early fusion baselines. Next, we focus on the best AutoMergeNet-generated model based on the methane dataset, and we report results applying this model to a realistic use case involving a week of satellite data and comparing its performance to a model manually designed by a domain expert.

A. Q1: AutoMergeNet-designed multi-branch networks outperform single-branch

We evaluated AutoMergeNet and the baseline models on the methane and carbon monoxide datasets described in Section IV-A. We found that on both datasets, AutoMergeNet outperformed all baselines by significant margins in terms of accuracy, precision, and recall, with and without the auxiliary classifier. Tables IV (methane) & V (carbon monoxide) show the average test scores of the best configurations found by AutoMergeNet.[8] The high performance on both datasets further underscores the value of AutoML in automatically creating ML pipelines for different datasets. AutoMergeNet never selected models with fusion depth 0. These results suggest that multi-branch networks of any depth achieve higher performance for methane and carbon monoxide plume detection than early fusion networks. The baselines achieved low precision but high recall on the methane dataset, sometimes even lower than predicting all negatives. AutoMergeNet uses the same backbone networks, data, training, and evaluation pipeline as the baselines. This rules out implementation issues. Investigating the results of these networks with Grad-CAMs revealed that even clearly empty images (no plume-like shapes) were regularly misclassified as plumes, explaining the low precision. Furthermore, the Grad-CAMs often exhibited attention on almost the full image, even if a plume was present, suggesting the single-branch models may be overfitting to a confounding feature in the data, such as the number of missing pixels. The features in the auxiliary data layers are sparse: artefacts often show plume-like features in only one of the auxiliary data layers at a time. As a result, early fusion may be effective when the number M of modalities is small, but can have increasing difficulty learning salient features as the number of modalities increases, compared with multi-branch fusion. Cross-branch attention mechanisms could potentially guide the model towards learning better features from sparse data; however, our preliminary experiments found no significant improvements, likely due to insufficient selectivity in the computed attention maps.

Our results show consistent evidence that simpler architectures yield better results. The CNN performs better or similarly to more complex baselines on methane with the auxiliary classifier and outperforms the baselines without the auxiliary classifier. On the carbon monoxide dataset, the CNN outperforms the other baselines regardless of using the auxiliary classifier. These results are also reflected in the configurations found by AutoMergeNet (Figure 5). On both datasets, the CNN was chosen most frequently as a backbone by a large margin, and MobileNetV2 and the transformer-based CvT were never chosen, despite recent work by Marjani et al. [20] reporting high performance in methane plume detection with a vision transformer or improvements shown by these networks in ImageNet classification [10], [11]. Therefore, high performance in one vision dataset does not guarantee state-of-the-art performance on other datasets. Furthermore, preliminary experiments suggested that ResNet18 performed better on plume detection than the larger variants (ResNet34, ResNet50 and ResNet101). These findings suggest that simpler models are better suited to plume detection in low-resolution images than complex backbone networks. Plume detection requires identifying relatively simple spatial patterns (the plume shape) rather than the complex hierarchical features that state-of-the-art image classification models are designed to extract. Additionally, given our M-modal input and dataset size, the CNN backbone is less susceptible to overfitting than models with more parameters. The baseline results align with AutoMergeNet's consistent selection of simpler architectures, demonstrating the system's ability to select architectures to match the task's requirements.

B. Q2: Auxiliary classifier significantly improves precision

The auxiliary classifier substantially increased the precision and accuracy of all approaches, especially the baselines. The auxiliary classifier corrects the low precision by discarding detections that are clearly empty based on the methane channel. On both datasets, the recall slightly decreased for some approaches because the auxiliary classifier misclassifies some plumes. Furthermore, our results demonstrate that early fusion baselines can achieve high performance with the help of the auxiliary classifier, but using NAS leads to high-performing models more consistently. This finding is further supported by

[8] One out of 30 runs of AutoMergeNet shows very low accuracy (≈ 0.3), which can occur as a result of the non-determinism inherent to neural networks and BOHB. We also repeated two runs that froze because of unknown problems in execution. In both cases, the repeated run completed without any problems.

TABLE III
HYPERPARAMETER VALUES IN THE SEARCH SPACE OF THE BASELINES HPO AND AUTOMERGENET NAS. VALUES SAMPLED FROM RANGES ARE DESCRIBED VIA A TUPLE [START, STOP, STEPSIZE (OPTIONAL)]. WHEN NO STEP SIZE IS GIVEN, THE HYPERPARAMETER IS TREATED AS CONTINUOUS. THE NUMBERS NEXT TO THE BACKBONE ARCHITECTURES IN THE NAS SEARCH SPACE INDICATE THE MAXIMUM FUSION DEPTH.

Search space	Hyperparameter	Value		Type
NAS & HPO	Batch size	`16,32,64,128,256`		Integer choice
	LR	`[0,0.1]`		Float on logarithmic scale
	Dropout	`[0.0,0.8,0.1]`		Float
	Weight decay	`[0,0.1]`		Float on logarithmic scale
NAS	Fusion depth	`[0,max_fusion_depth]`		Integer
		CNN	7	
		CVT	3	
		ResNet, BANet, EpsaNet	4	
		MobileNetv2	11	
	Fusion strategy	`concat, add, 1x1 conv`		Categorical
	Architecture	`fusion, resnet, epsanet, banet, mobilenet, cvt`		Categorical

TABLE IV
THE MEAN AND STANDARD DEVIATION OF THE TEST RESULTS ON THE **METHANE** DATASET OBTAINED FROM 5 INDEPENDENT HPO RUNS FOR EACH BASELINE AND 30 INDEPENDENT NAS RUNS FOR *AutoMergeNet*. WE TRAINED EACH RESULTING BEST CONFIGURATION FROM SCRATCH FIVE TIMES, LEADING TO 25 RESULTS FOR EACH BASELINE AND 150 FOR *AutoMergeNet*. **BOLD** RESULTS ARE STATISTICALLY SIGNIFICANT ACCORDING TO THE WILCOXON SIGNED RANK TEST WITH A p-VALUE OF 0.05.

	With Auxiliary Classifier				Without Auxiliary Classifier			
Model	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
BANet [9]	0.88 ± 0.02	0.85 ± 0.02	0.79 ± 0.06	0.81 ± 0.03	0.58 ± 0.05	0.43 ± 0.03	0.82 ± 0.07	0.57 ± 0.02
CNN [16]	0.89 ± 0.01	0.86 ± 0.02	0.80 ± 0.04	0.82 ± 0.02	0.62 ± 0.04	0.46 ± 0.03	0.83 ± 0.05	0.59 ± 0.01
CvT [11]	0.88 ± 0.03	0.85 ± 0.04	0.78 ± 0.12	0.81 ± 0.07	0.56 ± 0.08	0.42 ± 0.05	0.81 ± 0.012	0.55 ± 0.04
EpsaNet [44]	0.87 ± 0.04	0.84 ± 0.03	0.77 ± 0.18	0.78 ± 0.14	0.55 ± 0.06	0.42 ± 0.06	0.80 ± 0.18	0.53 ± 0.08
ResNet [8]	0.88 ± 0.03	0.84 ± 0.03	0.80 ± 0.12	0.81 ± 0.09	0.55 ± 0.06	0.42 ± 0.06	0.80 ± 0.18	0.53 ± 0.08
MobileNetV2 [10]	0.89 ± 0.00	0.84 ± 0.02	0.81 ± 0.02	0.83 ± 0.01	0.57 ± 0.03	0.43 ± 0.02	0.84 ± 0.03	0.57 ± 0.01
AutoMergeNet	$\mathbf{0.94} \pm 0.01$	$\mathbf{0.91} \pm 0.03$	$\mathbf{0.91} \pm 0.02$	$\mathbf{0.88} \pm 0.09$	$\mathbf{0.90} \pm 0.13$	$\mathbf{0.83} \pm 0.13$	$\mathbf{0.94} \pm 0.04$	$\mathbf{0.91} \pm 0.04$

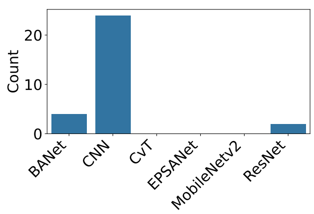
TABLE V
THE MEAN AND STANDARD DEVIATION OF THE TEST RESULTS ON THE **CARBON MONOXIDE** DATASET OBTAINED FROM 5 INDEPENDENT HPO RUNS FOR EACH BASELINE, AND 30 INDEPENDENT NAS RUNS FOR *AutoMergeNet*. WE TRAINED EACH RESULTING BEST CONFIGURATION FROM SCRATCH FIVE TIMES, LEADING TO 25 RESULTS FOR EACH BASELINE AND 150 FOR *AutoMergeNet*. **BOLD** RESULTS ARE STATISTICALLY SIGNIFICANT ACCORDING TO THE WILCOXON SIGNED RANK TEST WITH A p-VALUE OF 0.05.

	With Auxiliary Classifier				Without Auxiliary Classifier			
Model	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
BANet [9]	0.87 ± 0.02	0.91 ± 0.02	0.54 ± 0.09	0.67 ± 0.07	0.81 ± 0.04	0.64 ± 0.07	0.64 ± 0.09	0.64 ± 0.07
CNN [16]	0.91 ± 0.02	0.91 ± 0.02	0.71 ± 0.07	0.80 ± 0.05	0.87 ± 0.02	0.71 ± 0.05	0.83 ± 0.08	0.76 ± 0.05
CvT [11]	0.83 ± 0.02	0.92 ± 0.04	0.36 ± 0.09	0.51 ± 0.09	0.74 ± 0.01	0.54 ± 0.12	0.44 ± 0.11	0.46 ± 0.08
EpsaNet [44]	0.86 ± 0.02	0.91 ± 0.02	0.50 ± 0.09	0.64 ± 0.08	0.81 ± 0.03	0.65 ± 0.07	0.61 ± 0.10	0.62 ± 0.07
ResNet [8]	0.86 ± 0.02	0.93 ± 0.03	0.48 ± 0.11	0.62 ± 0.09	0.81 ± 0.03	0.64 ± 0.07	0.57 ± 0.13	0.6 ± 0.08
MobileNetV2 [10]	0.86 ± 0.02	0.92 ± 0.03	0.50 ± 0.10	0.64 ± 0.08	0.82 ± 0.02	0.66 ± 0.05	0.60 ± 0.11	0.62 ± 0.07
AutoMergeNet	$\mathbf{0.91} \pm 0.04$	$\mathbf{0.93} \pm 0.01$	$\mathbf{0.89} \pm 0.02$	$\mathbf{0.85} \pm 0.03$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.78} \pm 0.06$	$\mathbf{0.89} \pm 0.08$	$\mathbf{0.82} \pm 0.04$

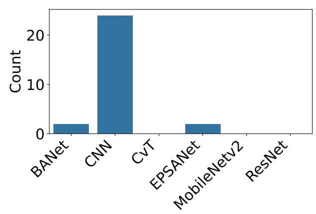
Figure 6, showing the bootstrapped accuracy, precision, and recall (with auxiliary classifier) of the configurations found by each approach. Figure 6 shows that AutoMergeNet most consistently found high-performing models. The early fusion baselines found models with wide ranges of recall on the carbon monoxide dataset, showing that while it is possible to find models with high performance, it may be necessary to run early-fusion HPO multiple times to find these models.

C. Q3: AutoMergeNet-created models applied in practice

In this section, we show how our methods can be applied to the high-impact use case of methane plume detection and compare it to the operational model of Schuit et al. [1]. Methane is of particular interest, as it is a strong greenhouse gas with a global warming potential 28 times that of carbon dioxide on a 100-year timescale [51]. A significant portion of the worldwide anthropogenic methane emissions comes from so-called super-emitters — a limited number of sources with abnormally high emission rates [52], including leaks in the oil and gas industry, large landfills, and coal mines. These emissions can often be reduced at no net cost, making them prime targets for climate change mitigation [53]. However, as methane is an odourless and invisible gas, identifying super-emitters is difficult. Satellite observations are increasingly being used to detect super-emitters. A prime example is the use of the TROPOMI instrument that provides daily global



(a) Methane



(b) Carbon Monoxide

Fig. 5. AutoMergeNet strongly favoured the CNN over the larger models as a backbone for both methane **(a)** and carbon monoxide plume detection **(b)**. These results align with the high baseline performance of the CNN and demonstrate AutoMergeNet's ability to identify architectures matching task requirements.

coverage of atmospheric methane, allowing the detection of large methane plumes. Because TROPOMI takes millions of observations, of which only a few contain super-emitter plumes, it is necessary to efficiently automate the search for these plumes in the data.

We performed methane plume detection with AutoMergeNet on a previously unseen testing set, consisting of a week of TROPOMI methane data over land, acquired 25—31 October 2021. We performed the same strategy as Schuit et al., but using their reported results and labels. We cropped 32×32 pixel images using a shifted window with an offset of 16 pixels and preprocessed these images the same way as the methane plumes dataset, yielding 17 760 images. These images were fed to our single best model with auxiliary classifier, as shown in Tables IV & V. We followed the approach by Schuit et al. [1] to generate binary pixel plume masks of each image labelled as plume and used these masks to remove duplicate, overlapping detections. Two independent labellers, including an atmospheric methane analysis expert, labelled each plume detection (because labelling all images was not possible due to the size of the test set). Labelling the detections requires domain-specific knowledge and must happen carefully to prevent false positives, which can unjustly attribute blame to facilities and states and, consequently, severely harm the trust in the monitoring system that produced these detections. We label detections as plumes, not plumes or inconclusive cases, when we could not assign conclusive labels to images where the data layers suggested contradicting labels. The inconclusive cases also include methane signals deemed real but less "plume-like" and instead show emissions from a wider area, resulting in spread-out elevated methane concentrations in the observation.

Our methods found 73 plumes, 67 non-plumes, and 46 inconclusive cases, compared to 85 plumes, 20 non-plumes, and 48 inconclusive cases found by Schuit et al. These results suggest the model automatically created by *AutoMergeNet* is competitive with the expert-designed pipeline in terms of detecting plumes, although it does produce relatively more false positives (67 for *AutoMergeNet*, 20 for the model by Schuit et al. [1]. The right panel of Figure 7 shows the locations of the false positives detected by our model. The false detections near the edges of the Sahara, in Yemen and Oman, and in the North of China near the Mongolian border suggest that our model is not robust to false positives caused by albedo variation in the desert. Adding more training examples of this terrain type to the training dataset could be a potential solution to this problem. Furthermore, the automatically retrieved data is uncurated and, therefore, may contain more missing pixels than the expert-selected images in the training set. Training the model to be more robust against noise is challenging because many strong augmentations, such as CutMix [54] and RandomErase [55], risk erasing a plume and subsequently invalidating the label. The model by Schuit et al. may be less sensitive to these noise sources, because the physics-derived features summarise each image at the cost of loss of transferability to related problems such as carbon monoxide detection. On the other hand, our methods do not use any domain-specific knowledge except the choice of data layers, showing that an automatically designed, multimodal deep learning pipeline can be used similarly to a state-of-the-art method designed using a great deal of domain knowledge and potentially increase the speed of development of such models. We believe that our results demonstrate the successful application of AutoML to methane detection and that AutoML is a promising method for solving high-impact EO problems previously untackled with ML.

VI. Conclusions

We presented *AutoMergeNet*, a NAS framework for M-modal EO image data fusion. We show that AutoML can successfully be applied to fuse as many as 11 distinct data layers. To design AutoMergeNet, we defined a search space that transforms widely used image classification networks into multi-branch fusion networks. These networks model complex multimodal EO satellite data facilitated by optimising critical multi-branch model hyperparameters. Furthermore, we introduced an auxiliary unimodal classifier to address the high dimensionality of the given data and the relative dearth of labels. We validated our methods on a methane plumes dataset with 11 data layers and a new carbon monoxide dataset with 10 data layers, which we gathered and labelled ourselves.

Our experiments show that AutoMergeNet outperforms all early fusion baselines and consistently achieves high performance on both datasets. The auxiliary classifier greatly improves the precision of our pipeline and the overall performance of the early fusion baselines. We also demonstrated the applicability of our approach in a real-life use case. Our automated and adaptable framework achieves results comparable to a highly specialised state-of-the-art baseline in terms of plume detection. These results show the potential of designing AutoML solutions for tracking many more events impacting our environment.

Compared to previous work in NAS for image data fusion, AutoMergeNet automatically creates and configures networks capable of fusing any number of modalities. This ability, combined with automated feature extraction, makes AutoMergeNet applicable to satellite image datasets with more than two modalities without the need for manually designing new machine-learning pipelines. As a result, AutoMergeNet makes

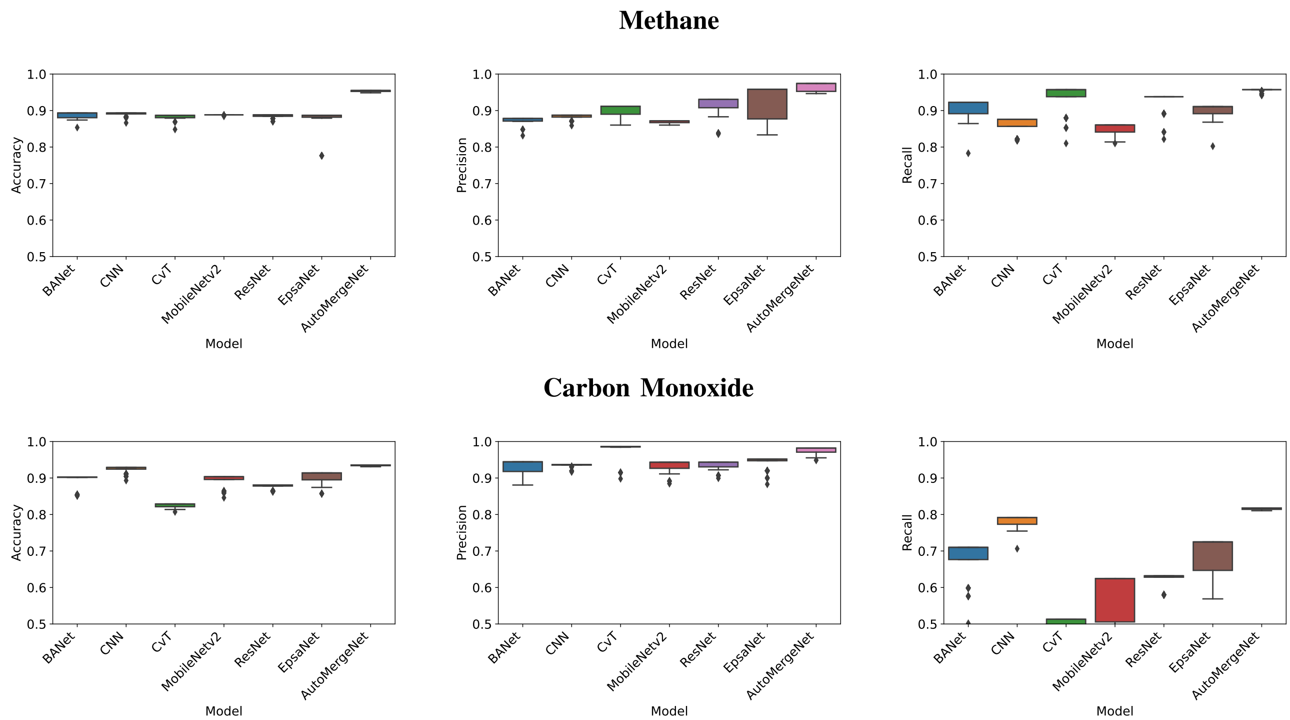


Fig. 6. Boxplots of bootstrapped accuracy, precision, and recall of the baselines and AutoMergeNet (both with auxiliary classifier). AutoMergeNet most consistently finds good models for both datasets.

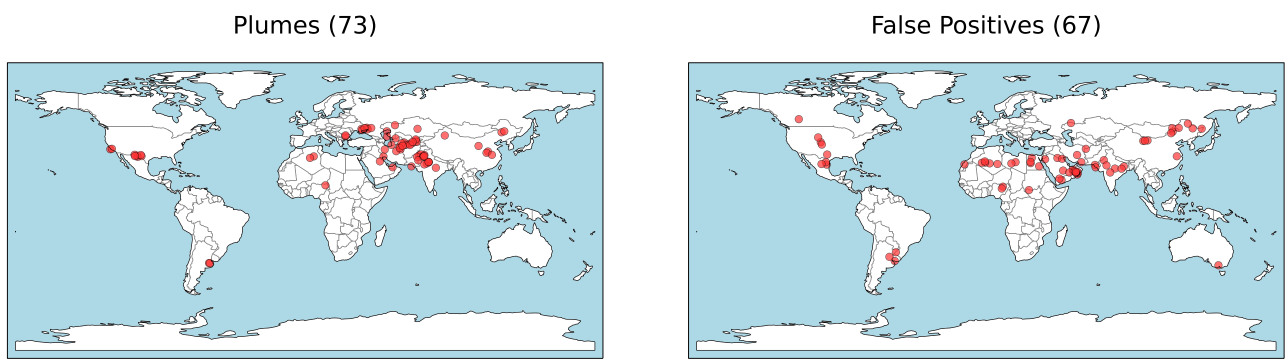


Fig. 7. The methane detections of the single best *AutoMergeNet* model with filtering after detections were labelled by a domain expert: 73 detected plumes (**left**) and 67 detected false positives (**right**). Many false positives occur in sandy regions (edges of the Sahara, Yemen, Oman, North China), indicating our model is not robust to artefacts related to albedo variations.

ML more accessible to domain experts from Earth Observation, providing a user-friendly solution to complex data fusion tasks by allowing to iteratively identify sources of artefacts and improve models by adding additional layers faster. In future work, we aim to identify causes of the generalisation gap between the testing results and the use case application.

ACKNOWLEDGEMENT

This publication is part of the "Physics-aware Spatio-temporal Machine Learning for Earth Observation Data" project (project number OCENW.KLEIN.425) of the Open Competition ENW research programme, which is partly financed by the Dutch Research Council (NWO). Additionally, this project was supported by the Open Space Innovation Platform[9] as a Co-Sponsored Research Agreement and carried out under the Discovery programme of, and funded by, the European Space Agency (contract number 4000136204/21/NL/GLC/my) and by an Alexander von Humboldt Professorship in Artificial Intelligence awarded to Holger Hoos. We thank the team that realized the TROPOMI instrument and its data products, consisting of the partnership between Airbus Defence and Space Netherlands, KNMI, SRON, and TNO and commissioned by NSO and ESA. The Sentinel-5 Precursor is part of the EU Copernicus program, and Copernicus (modified) Sentinel-5P data (2018– 2023) have been used.

[9]https://ideas.esa.int. Last Accessed 19 August 2024.

REFERENCES

[1] B. J. Schuit, J. D. Maasakkers, P. Bijl, G. Mahapatra, A.-W. van den Berg, S. Pandey, A. Lorente, T. Borsdorff, S. Houweling, D. J. Varon, J. McKeever, D. Jervis, M. Girard, I. Irakulis-Loitxate, J. Gorroño, L. Guanter, D. H. Cusworth, and I. Aben, "Automated detection and monitoring of methane super-emitters using satellite data," *Atmospheric Chemistry and Physics*, vol. 23, no. 16, p. 9071–9098, Sep. 2023.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[3] Z. Zheng, A. M. Fiore, D. M. Westervelt, G. P. Milly, J. Goldsmith, A. Karambelas, G. Curci, C. A. Randles, A. R. Paiva, C. Wang, Q. Wu, and S. Dey, "Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed," *Journal of Advances in Modeling Earth Systems*, vol. 15, no. 3, p. e2022MS003099, 2023.

[4] S. Kurchaba, J. van Vliet, F. J. Verbeek, and C. J. Veenman, "Anomalous NO2 emitting ship detection with TROPOMI satellite data and machine learning," *Remote Sensing of Environment*, vol. 297, p. 113761, Nov. 2023.

[5] E. Amri, P. Dardouillet, A. Benoit, H. Courteille, P. Bolon, D. Dubucq, and A. Credoz, "Offshore Oil Slick Detection: From Photo-Interpreter to Explainable Multi-Modal Deep Learning Models Using SAR Images and Contextual Data," *Remote Sensing*, vol. 14, p. 3565, 2022.

[6] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*, ser. The Springer Series on Challenges in Machine Learning. Cham, Switzerland: Springer International Publishing, 2019.

[7] M. Baratchi, C. Wang, S. Limmer, J. N. van Rijn, H. Hoos, T. Bäck, and M. Olhofer, "Automated machine learning: Past, present and future," *Artificial Intelligence Review*, vol. 57, no. 5, p. 122, Apr. 2024.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] Y. Zhao, J. Chen, Z. Zhang, and R. Zhang, "BA-Net: Bridge attention for deep convolutional neural networks," in *Computer Vision – ECCV 2022*. Cham, Switzerland: Springer, 2022, pp. 297–312.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Washington DC, USA: IEEE, Jun. 2018, p. 4510–4520.

[11] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Washington DC, USA: IEEE Computer Society, oct 2021, pp. 22–31.

[12] X. Deng, Y. Zhang, M. Xu, S. Gu, and Y. Duan, "Deep Coupled Feedback Network for Joint Exposure Fusion and Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 3098–3112, 2021.

[13] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington DC, USA: IEEE Computer Society, jun 2023, pp. 13 955–13 965.

[14] X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Information Fusion*, vol. 92, pp. 127–138, 2023.

[15] D. Li, W. Xie, Y. Li, and L. Fang, "FedFusion: Manifold-driven federated learning for multi-satellite and multi-modality fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, p. 1–13, 2024.

[16] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, J. Yao, and B. Zhang, "More Diverse Means Better: Multimodal Deep Learning Meets Remote Sensing Imagery Classification Hyperspectral Multispectral Synthetic Aperture Radar OpenStreetMap Fasle-color Image Classification Map," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[17] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, p. 20–32, Jun. 2018.

[18] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M3Fusion A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, p. 4939–4949, Dec. 2018.

[19] J. Hu, R. Liu, D. Hong, A. Camero, J. Yao, M. Schneider, F. Kurz, K. Segl, and X. X. Zhu, "MDAS: a new multimodal benchmark dataset for remote sensing," *Earth System Science Data*, vol. 15, no. 1, p. 113–131, Jan. 2023.

[20] M. Marjani, M. Mahdianpari, D. J. Varon, and F. Mohammadimanesh, "The integration of vision transformers and SAM for automated methane super-emitter detection using TROPOMI data," *Journal of Environmental Management*, vol. 393, p. 127034, Oct. 2025.

[21] H. Jin, Q. Song, and X. Hu, "Auto-Keras: An Efficient Neural Architecture Search System," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, p. 1946–1956.

[22] N. R. Palacios Salinas, M. Baratchi, J. N. van Rijn, and A. Vollrath, "Automated Machine Learning for Satellite Data: Integrating Remote Sensing Pre-trained Models into AutoML Systems," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, ser. Lecture Notes in Computer Science, vol. 12979. Cham, Switzerland: Springer International Publishing, 2021, p. 447–462.

[23] J. Wąsala, S. Marselis, L. Arp, H. Hoos, N. Longépé, and M. Baratchi, "AutoSR4EO: An AutoML Approach to Super-Resolution for Earth Observation Images," *Remote Sensing*, vol. 16, no. 3, p. 443, Jan. 2024.

[24] X. Li, L. Lei, and G. Kuang, "Multi-Modal Fusion Architecture Search for Land Cover Classification Using Heterogeneous Remote Sensing Images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, p. 5997–6000.

[25] X. Li, L. Lei, C. Zhang, and G. Kuang, "Multimodal Semantic Consistency-Based Fusion Architecture Search for Land Cover Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–14, 2022.

[26] S. Feng, Z. Li, B. Zhang, T. Chen, and B. Wang, "DSF2-NAS: Dual-Stage Feature Fusion via Network Architecture Search for Classification of Multimodal Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–14, 2025.

[27] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Breckenridge, Colorado, USA: PMLR, 10–15 Jul 2018, pp. 1437–1446. [Online]. Available: https://proceedings.mlr.press/v80/falkner18a.html

[28] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: a novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 6765–6816, Jan. 2017.

[29] J. P. Veefkind, I. Aben, K. McMullan, H. Förster, J. de Vries, G. Otter, J. Claas, H. J. Eskes, J. F. de Haan, Q. Kleipool, M. van Weele, O. Hasekamp, R. Hoogeveen, J. Landgraf, R. Snel, P. Tol, P. Ingmann, R. Voors, B. Kruizinga, R. Vink, H. Visser, and P. F. Levelt, "TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications," *Remote Sensing of Environment*, vol. 120, pp. 70–83, May 2012.

[30] O. Hasekamp, A. Lorente, H. Hu, A. Butz, J. Aan de Brugh, and J. Landgraf, "Algorithm Theoretical Baseline Document for Sentinel-5 Precursor methane Retrieva, SRON The Netherlands Institute for Space Research, Leiden, the Netherlands," 2022. [Online]. Available: https://sentinels.copernicus.eu/documents/247904/2476257/Sentinel-5P-TROPOMI-ATBD-Methane-retrieval.pdf/f275eb1d-89a8-464f-b5b8-c7156cda874e?t=1658313508597

[31] J. Landgraf, J. de Brugh, R. Scheepmaker, T. Borsdorff, and O. Houweling, S. Hasekamp, "Algorithm theoretical baseline document for sentinel-5 precursor: Carbon monoxide total column retrieval, Netherlands Institute for Space Research, the Netherlands, SRON-S5P-LEV2-RP-002," 2018. [Online]. Available: https://earth.sron.nl/wp-content/uploads/2023/12/SRON-ESA-S5L2PP-ATBD-002.pdf

[32] A. Lorente, T. Borsdorff, A. Butz, O. Hasekamp, J. Aan De Brugh, A. Schneider, L. Wu, F. Hase, R. Kivi, D. Wunch, D. F. Pollard, K. Shiomi, N. M. Deutscher, V. A. Velazco, C. M. Roehl, P. O. Wennberg, T. Warneke, and J. Landgraf, "Methane retrieved from TROPOMI: Improvement of the data product and validation of the first 2 years of measurements," *Atmos. Meas. Tech*, vol. 14, pp. 665–684, 2021.

[33] B. J. Schuit, J. D. Maasakkers, P. Bijl, and I. Aben, "Training datasets with manually labeled TROPOMI data for Machine Learning models [Schuit et al. 2023: Automated detection and monitoring of methane super-emitters using satellite data]," 2025.

[34] J. J. Danielson and D. B. Gesch, "Global multi-resolution terrain elevation data 2010 (gmted2010)," 2011. [Online]. Available: http://dx.doi.org/10.3133/ofr20111073

[35] R. Siddans, "Algorithm Theoretical Baseline Document for Sentinel-5 Precursor Cloud Processor, SRON The Netherlands Institute for Space Research, Leiden, the Netherlands," 2016. [Online]. Available: https://sentiwiki.copernicus.eu/__attachments/1673595/S5P-NPPC-RAL-ATBD-0001%20-%20Sentinel-5P%20Cloud%20Processor%20ATBD%202016%20-%201.0.pdf?inst-v=ced695cc-1f77-439d-9415-e622981ae935

[36] M. Friedl and D. Sulla-Menashe, "Modis/terra+aqua land cover type yearly l3 global 500m sin grid v061," 2022. [Online]. Available: https://lpdaac.usgs.gov/products/mcd12q1v061/

[37] M. Carroll, C. DiMiceli, J. Townshend, R. Sohlberg, A. Hubbard, M. Wooten, C. Spradlin, R. Gill, S. Strong, A. Burke, M. Blanco-Rojas, and M. Frost, "Modis/terra land water mask derived from modis and srtm l3 global 250m sin grid v061," 2024. [Online]. Available: https://lpdaac.usgs.gov/products/mod44wv061/

[38] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[39] G. Leguijt, J. D. Maasakkers, H. A. C. Denier van der Gon, A. J. Segers, T. Borsdorff, and I. Aben, "Quantification of carbon monoxide emissions from african cities using TROPOMI," *Atmospheric Chemistry and Physics*, vol. 23, no. 15, pp. 8899–8919, 2023.

[40] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, Aug. 2022.

[41] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in Earth Observation: The MiniFrance suite, dataset analysis and multi-task network study," *Machine Learning*, vol. 111, no. 9, pp. 3125–3160, Sep. 2022.

[42] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, L. Hughes, H. Li, Y. Sun, G. Zhang, S. Han, M. Schmitt, and Y. Wang, "So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[43] G. Sumbul, A. de Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, "BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, May 2021.

[44] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part III*. Berlin, Germany: Springer-Verlag, 2023, pp. 541–557.

[45] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Mission Critical – Satellite Data is a Distinct Modality in Machine Learning," Feb. 2024.

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[47] L. Li, K. Jamieson, A. Rostamizadeh, K. Gonina, M. Hardt, B. Recht, and A. Talwalkar, "Massively parallel hyperparameter tuning," 2018. [Online]. Available: https://openreview.net/forum?id=S1Y7OOlRZ

[48] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, and I. Stoica, "Ray: A distributed framework for emerging AI applications," in *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'18. USA: USENIX Association, Oct. 2018, pp. 561–577.

[49] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[50] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[51] Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2023.

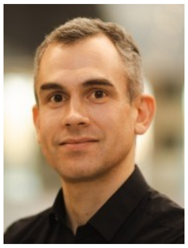
[52] D. Zavala-Araiza, D. Lyon, R. A. Alvarez, V. Palacios, R. Harriss, X. Lan, R. Talbot, and S. P. Hamburg, "Toward a Functional Definition of Methane Super-Emitters: Application to Natural Gas Production Sites," *Environmental Science & Technology*, vol. 49, no. 13, pp. 8167–8174, Jul. 2015.

[53] T. Lauvaux, C. Giron, M. Mazzolini, A. d'Aspremont, R. Duren, D. Cusworth, D. Shindell, and P. Ciais, "Global assessment of oil and gas methane ultra-emitters," *Science*, vol. 375, no. 6580, pp. 557–561, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abj4351

[54] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[55] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

Julia Wąsala is a PhD candidate at the Leiden Institute for Advanced Computer Science (LIACS) at Leiden University and at SRON Space Research Organisation Netherlands. Her research in the field of Automated Machine Learning for Earth Observation focuses on designing new methods and validating them in real-world applications such as atmospheric plume detection.



J. D. (Bram) Maasakkers is a senior scientist at SRON Space Research Organisation Netherlands. He works on the interpretation of satellite observations of methane and carbon monoxide with a focus on how to use those observations to quantify emissions and support their mitigation.

Berend J. Schuit received a B.Sc. degree in Civil Engineering in 2018 and M.Sc. in Aerospace Engineering (Spaceflight, Space Exploration) in 2021 from Delft University of Technology. He is currently working towards a Ph.D. degree (Faculty of Earth Sciences, Vrije Universiteit Amsterdam) at SRON Space Research Organisation Netherlands in cooperation with GHGSat Inc.. Focus areas of his research include satellite remote sensing of methane emissions (e.g. using TROPOMI, Sentinel-2) and machine learning for methane plume detection.

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3621068

Gijs Leguijt Gijs Leguijt obtained his master's degree in Particle Physics at the University of Amsterdam, Netherlands, in 2020. After this, he changed his scope to climate research in order to make a contribution to a sustainable future. From 2021 onwards, he has worked towards a Ph.D. on space-based monitoring of anthropogenic emissions as part of a collaboration between the SRON Space Research Organisation Netherlands and the Netherlands Organisation for Applied Scientific Research. His research is mainly focused on satellite observations and emission quantification techniques, including the modeling of atmospheric transport and comparisons to ground-based methods.

Ilse Aben is a senior scientist at SRON Space Research Organisation Netherlands. She is an expert in satellite remote sensing of greenhouse gases.

Rochelle Schneider received the PhD degree in Geospatial Analytics in 2019 from the University College London (UCL, UK). She is currently working as Copernicus and Destination Earth Ecosystem Operations Engineer at the European Space Agency (ESA), delivering Europe's Climate Digital Ecosystem for Destination Earth Initiative (destination-earth.eu). Previously, Dr Schneider served as AI Applications Lead at ESA's Φ-lab (https://philab.esa.int), deploying machine learning workflows using Earth observation satellites, climate models, and digital twin products.

Holger H. Hoos holds an Alexander von Humboldt professorship in AI at RWTH Aachen University (Germany), where he also leads the AI Center, as well as a professorship in machine learning at Universiteit Leiden (the Netherlands) and an adjunct professorship in computer science at the University of British Columbia (Canada). He is a Fellow of the Association of Computing Machinery (ACM) and the Association for the Advancement of Artificial Intelligence (AAAI), and he currently serves as president of the European AI Association (EurAI); he is also past president of the Canadian Association for Artificial Intelligence and one of initiators of CAIRNE (formerly CLAIRE), an initiative by the European AI community that seeks to strengthen European excellence in AI research and innovation (https://cairne.eu).

Mitra Baratchi is associate professor of artificial intelligence at Leiden University, where she leads the Spatio-temporal data Analysis and Reasoning (STAR) and co-leads of the Automated Design of Algorithms research group. Her research interests lie in spatio-temporal, time-series, and mobility data modelling. She strongly focuses on developing algorithms for wearable sensors data, Earth observations and other open spatio-temporal data sources. Specifically, she explores the design of algorithms that can automatically handle all necessary data processing tasks from the point of data collection to high-level modelling, extraction of information, and effective decision-making from such data. Her research targets applications in a broad range of urban, environmental, and industrial domains, for which she has collaborated, notably with the European Space Agency, Netherlands Institute for Space Research, Honda Research Institute, various municipalities, and researchers in other scientific disciplines.