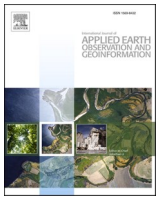
International Journal of Applied Earth Observation and Geoinformation 144 (2025) 104876

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Dynamic fusion of medium-resolution optical and SAR imagery for methane source infrastructure classification

Yanglangxing He [a], Xueliang Zhang [a,*], Pengfeng Xiao [a], Zhenshi Li [a], Dilxat Muhtar [a], Feng Gu [a], Binxiao Liu [b], Pengming Feng [b]

[a] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing, Jiangsu 210023, China
[b] State Key Laboratory of Space Information System and Integrated Application, Beijing Institute of Satellite Information Engineering, Beijing 100086, China

ARTICLE INFO

Keywords:
Earth observation
Methane source infrastructure
Multimodal data fusion
Dynamic neural networks
Deep learning

ABSTRACT

Accurate classification of methane source infrastructure across sectors is critical for building comprehensive emission inventories and tracing emission sources. Existing approaches predominantly rely on high-resolution remote sensing imagery to capture discriminative features, but their scalability is limited by high costs and restricted availability. In contrast, medium-resolution imagery offers scalable alternatives with enhanced spectral signatures, while its lower spatial resolution challenges precise characterization and facility differentiation. To address this issue, we propose a multimodal fusion method on Sentinel-2 and Sentinel-1 data, with the aim of exploiting the complementary characteristics of optical, infrared, and SAR imagery to improve classification accuracy. We present a multimodal dynamic fusion network (DMFNet), which incorporates a gating module and multimodal attention fusion modules (MAFM) to adaptively address sample variability and multimodal heterogeneity. Additionally, DMFNet enables tracking and interpreting the fusion process by analyzing data-driven weights, providing deep insights into modality combinations and fusion strategies for specific facility. Experiments on the METER-ML dataset demonstrate that the proposed model achieves a precision of 0.740 and a recall of 0.757, outperforming existing single-modal and static fusion methods. Transferability experiments further confirm the practical applicability of the proposed method and its complementarity with existing open-source data in improving methane emission inventories.

1. Introduction

Methane is the second-largest greenhouse gas, responsible for approximately 30 % of global temperature rise since the Industrial Revolution (Masson-Delmotte et al., 2021; Saunois et al., 2016). Anthropogenic activities, including agriculture, energy production, and waste management (Jackson et al., 2020; Lee et al., 2023), are the primary drivers of increasing atmospheric methane concentrations, accounting for around 60 % of annual global emissions (Kirschke et al., 2013). Effective mitigation of these emissions requires accurate estimation and attribution of methane sources. Currently, methane emissions are commonly assessed through two complementary approaches: bottom-up methods aggregate emission data from individual sources to generate regional estimates (Saunois et al., 2016; Vaughn et al., 2018), whereas top-down methods utilize direct measurements of atmospheric methane concentrations (Zhang et al., 2023). For bottom-up approaches, accurate source types and quantities are essential for selecting accurate emission factors and minimizing uncertainties arising from misclassification or omitted fugitive emissions (Chen et al., 2022). For top-down approaches, precise facility geographic locations are necessary to effectively link observed methane concentrations with specific emission activities (Lauvaux et al., 2022; Schuit et al., 2023).

The Intergovernmental Panel on Climate Change (IPCC) classifies anthropogenic greenhouse gas emission sources into six sectors: energy, industrial processes, solvent and other product use, agriculture, land-use change and forestry, and waste. In 2023 (IEA, 2024), agriculture was the largest source of global methane emissions (145 million tonnes), followed by energy (130 million tonnes) and waste (71 million tonnes). Several publicly accessible databases currently document methane emission sources, including oil and gas facilities and wastewater treatment plants (Ehalt Macedo et al., 2022; Maus et al., 2022; Sabbatino, 2018). These databases typically integrate national, regional, and

* Corresponding author.
E-mail address: zxl@nju.edu.cn (X. Zhang).

https://doi.org/10.1016/j.jag.2025.104876
Received 27 March 2025; Received in revised form 9 September 2025; Accepted 19 September 2025
Available online 28 September 2025

corporate datasets, supplemented by published literature, government reports, and industry publications. However, limitations in data availability and delays in reporting often lead to outdated or inaccurate spatial locations. Additionally, facilities like landfills, animal feeding operations, and mines frequently bypass licensing requirements or remain underregulated, especially in developing countries, exacerbating data incompleteness and inaccuracies (Handan-Nader and Ho, 2019; Sun et al., 2023).

To address these gaps, large-scale and timely identification of methane-emitting infrastructure is urgently needed to improve emission inventories and support targeted mitigation efforts. Remote sensing has become an essential tool for monitoring energy systems and infrastructure in this context (Ren et al., 2022). Methane-emitting facilities vary widely in type and spatial distribution, ranging from industrial sites with concentrated layouts and distinct structural patterns (e.g., oil refineries, processing plants) to agricultural and waste facilities characterized by dispersed forms and less prominent features (e.g., CAFOs (Concentrated Animal Feeding Operation), landfills). These variations in structural composition and scene characteristics influence their detectability at different spatial resolutions and determine the choice of data sources and identification methods. Existing studies have leveraged multi-dimensional features, including geometric structure, spectral properties, and spatial configuration, to develop facility identification models across multiple scales and analytical strategies. Based on recent research advances, remote sensing-based methane facility identification approaches can be broadly classified into three categories: feature-based facility screening, scene-based facility recognition, and component-based fine-scale identification.

Feature-based facility screening primarily targets large-scale, cost-effective preliminary mapping tasks. These approaches exploit distinctive spectral signatures, thermal-related information, or land use characteristics of target facilities and their surroundings to enable rapid localization and identification. They typically rely on imagery from multispectral satellites such as Sentinel-2 and Landsat-8, analyzing features such as high-temperature anomalies (HTAs) (Liu et al., 2021b; Wu et al., 2024), land surface temperature (LST) (Beaumont et al., 2014; Gill et al., 2019; Yan et al., 2014), and land use/land cover (LULC) classifications to distinguish potential facilities (Li et al., 2022a).

For facility types with stable spatial layouts and coherent scene structures, scene-based facility recognition methods are widely adopted. These approaches utilize image classification or object detection models to extract and learn discriminative scene-level features that capture macro-scale distributions and geometric forms (Handan-Nader and Ho, 2019; Niu et al., 2023; Sheng et al., 2020). However, such methods often face limitations in representing the detailed shapes, sizes, and complex internal spatial relationships of facility components.

Component-based identification focuses on detecting key structural elements within methane-emitting facilities, such as storage tanks (Zalpour et al., 2020; Zhang et al., 2015), pond (Li et al., 2025), and livestock barns (Robinson et al., 2022). These targets often exhibit distinct geometric shapes and strong semantic features. By leveraging the fine spatial detail available in high-resolution remote sensing imagery, combined with object detection and semantic segmentation techniques, these methods can achieve precise recognition and localization of internal facility structures.

Regarding data sources, while high-resolution imagery provides rich texture and structural information, its high acquisition cost and limited transferability pose considerable challenges for large-scale applications. In contrast, medium-resolution remote sensing imagery offers broader accessibility and higher spectral resolution, providing diverse modalities such as visible (RGB), infrared (IR), and synthetic aperture radar (SAR). Previous studies have effectively utilized spectral reflectance, thermal radiation, and radar scattering characteristics to identify various methane emission facilities. Optical imagery, particularly in the visible spectrum, conveys detailed information on color, shape, and texture, thus serving as a primary modality for infrastructure detection. Infrared imagery, a passive sensing approach, captures thermal characteristics of facilities and their surrounding environments, offering advantages such as day-and-night imaging and reduced susceptibility to interference (Jiang et al., 2024). Specifically, shortwave infrared (SWIR, 1.4-3.0μm) is sensitive to high-temperature phenomena, facilitating detection of thermal anomalies associated with fossil fuel combustion in methane-emitting facilities (Liu et al., 2021b; Wu et al., 2024). Conversely, SAR imagery, as an active sensing technology, mitigates optical remote sensing limitations like cloud and weather interference (Liu et al., 2021a). By leveraging scattering, phase, and amplitude information, SAR has proven effective in applications such as monitoring ground subsidence in mining areas and identifying detailed mining structures (Guo et al., 2024a; Moon and Lee, 2021).

However, relying on a single modality often provides insufficient information for accurately identifying diverse methane-emitting facilities, particularly at medium spatial resolutions (Bansal and Tripathi, 2024). The combination of medium-resolution Sentinel-2 MSI (Multispectral Instrument) and Sentinel-1 SAR data offers global coverage and strong scalability, providing complementary spectral and geometric information. By fusing multi-modal data, we can effectively mitigate limitations of spatial resolution by integrating these complementary features, thereby enhancing detection accuracy. However, the performance of fusion models largely depends on their ability to integrate task-specific information across modalities (Li et al., 2022b). Traditional fusion approaches extract modality-specific features and merge them at fixed network layers (input, intermediate, or decision layers) according to predefined algorithms or rules (Wu et al., 2022; Xu et al., 2018). These static fusion methods assume constant data quality, making them less adaptable to input variability or changing task demands. While computationally efficient and straightforward to implement, static fusion methods lack flexibility and struggle to handle redundancy in multi-modal data.

Unlike static fusion, dynamic fusion methods offer greater flexibility by adaptively adjusting fusion strategies during inference, effectively leveraging variations in input features and scene complexity (Cai et al., 2021; Li et al., 2020).Attention mechanisms are frequently employed in dynamic fusion to emphasize the contributions of individual modalities through learned weighting (Han et al., 2022; Jin et al., 2022). Additionally, gating mechanisms and mixture-of-experts approaches have been explored, enabling dynamic adjustments of network structures to reduce redundancy and improve computational efficiency (Mena et al., 2025; Wei et al., 2024). However, applying dynamic fusion to methane emission facility identification requires addressing both multimodal feature heterogeneity and inter-sample variations. A single dynamic fusion approach often struggles to simultaneously resolve these issues, necessitating a more flexible and adaptive architecture that optimizes fusion strategies for diverse facility types.

Hence, despite its potential, the practical application of medium-resolution remote sensing imagery for methane infrastructure identification still encounters significant hurdles. First, limited spatial resolution restricts models from capturing fine-grained features, hindering the discrimination of structurally similar facilities and reducing detection accuracy. Second, efficiently utilizing multimodal data—maximizing complementarity while minimizing redundancy and noise—remains a critical challenge. Finally, exploiting the flexibility and interpretability of dynamic fusion methods to quantify individual modality contributions, clarify modality interactions, and tailor fusion strategies for distinct facility types requires further exploration.

To address these challenges, we propose a multimodal dynamic fusion method integrating optical, infrared, and SAR features, aiming to improve the accuracy and robustness of methane-emitting facility identification from medium-resolution imagery. Specifically, we introduce a dynamic multimodal fusion network (DMFNet) that leverages gating mechanisms and multimodal attention to adaptively generate data-driven fusion strategies and network paths, enabling progressive feature integration across modalities. Extensive comparisons with

single-modal and static fusion approaches on public datasets demonstrate that our method achieves superior accuracy. Additionally, by analyzing the learned fusion weights, we quantitatively assess the contributions of different modalities in the overall task and specific facility identification, enhancing model reliability and interpretability.

The main contributions of this study are as follows:

1. We propose a multimodal approach utilizing Sentinel-2 and Sentinel-1 imagery to classify methane-emitting facilities. By effectively integrating optical, near-infrared, and SAR features, this method mitigates the limitations of medium-resolution imagery and improves classification accuracy
2. We develop DMFNet tailored to classifying methane emission facilities. This model employs a gating mechanism to dynamically adjust fusion pathways according to input characteristics, while MAFM selectively emphasize complementary information and suppress redundancy. We also provide a quantitative analysis of modality-specific contributions, clarifying the cooperative roles of each modality in classifying distinct facility types
3. We validate the effectiveness of the proposed method using both public and custom datasets. Results on public datasets confirm its robustness, while its application in real-world scenarios highlights its value in improving methane emission inventory completeness and complementing existing methodologies

2. Dataset

We focus on key methane-emitting infrastructures within IPCC major sectors, including concentrated animal feeding operations (CAFOs) in agriculture; coal mines, natural gas processing plants (Proc Plants), and refineries and oil terminals (R&Ts) in energy; and landfills and wastewater treatment plants (WWTPs) in waste management. These facilities are critical contributors to global and regional methane budgets due to their high emission intensities or potential emission risks, and they represent priority targets for mitigation strategies and inventory development.

Our selection encompasses both known methane "super-emitters" and facilities with lower individual emission rates but substantial cumulative emissions. Beyond their emission significance, these facilities typically exhibit distinctive geometric patterns, spectral characteristics, or structural features in medium-resolution remote sensing imagery. Such features make them comparatively easier to detect and distinguish from surrounding land cover types or infrastructure classes, thereby supporting reliable identification through image-based classification approaches.

2.1. METER-ML dataset for model training and validation

For robust development and validation of our identification model, we employed the METER-ML dataset, a multi-sensor dataset specifically designed for automated methane source mapping (Zhu et al., 2022). It consists of 86,599 georeferenced images, encompassing six types of methane-emitting facility scenes as well as negative samples from various landscapes and infrastructure. The dataset is divided into a training set of 85,066 images, a validation set of 515 images, and a test set of 1018 images. All validation and test images were manually verified by experts, ensuring high-quality reference labels for reliable evaluation.

METER-ML includes 19 spectral bands from three image sources: Sentinel-2, Sentinel-1, and NAIP (National Agriculture Imagery Program). For this study, we selected six bands from the Sentinel-2 imagery, including three 10-m resolution visible bands (Bands 2-4), one 20-m resolution narrow near-infrared (NIR) band (Band 8A, centered at 865 nm), and two 20-m resolution SWIR bands (Bands 11 and 12). Additionally, we used the 10-m resolution VV and VH bands from Sentinel-1 imagery. Aerial imagery from NAIP was not used in this study.

During preprocessing, problematic samples were removed from the dataset, and additional negative samples were added to the test set to enhance evaluation robustness. We adopted the original split configuration of the METER-ML dataset, in which the training set is substantially larger than the test set. This split strategy has been adopted and validated in recent studies (Berg et al., 2023; Berg et al., 2024; Irvin et al., 2023), as it ensures comprehensive model learning while maintaining rigorous evaluation, supported by a high-quality, expert-validated test set. Table 1 summarizes the sample distribution across different facility types in the METER-ML dataset after preprocessing.

2.2. Data for transportability experiments

We assessed the generalization and transferability of the proposed model by applying the best-performing model trained on the METER-ML dataset to identify methane-emitting facilities in Los Angeles County. Sentinel-2 and Sentinel-1 imagery, covering the administrative boundary of Los Angeles County with less than 5 % cloud coverage, was obtained from Google Earth Engine (GEE) for the period of May to September 2023. The Sentinel-2 imagery used corresponds to atmospherically and geometrically corrected Level-2A products, with cloud-contaminated pixels removed using the QA60 quality mask. To reduce temporal inconsistency, Sentinel-1 and Sentinel-2 scenes with the closest possible acquisition dates were paired. Across the study area, the median time difference between paired scenes was 1.267 days, with all pairs within 10 days. Although exact synchronization was not always feasible due to acquisition schedules and cloud constraints, such short time differences are unlikely to significantly affect facility observability, as most facility types maintain stable structural and radiometric characteristics over these timescales.

Methane point source emission data for Los Angeles was obtained from Carbon Mapper for comparison with the model's identification results. Carbon Mapper's large-scale detection system utilizes hyperspectral sensors, including the Planet Tanager constellation and NASA's EMIT sensor, to monitor point-source methane emissions at high temporal frequencies. The Level 4B Source Emissions dataset, available through Carbon Mapper's open data platform, provides methane emission records from January 2016 to December 2024, including source locations and sector classifications within Los Angeles County.

2.3. Characteristics of methane source infrastructure in RGB, IR and SAR

This section analyzes distinctive features captured by RGB, IR, and SAR imagery (Fig. 1), highlighting their unique contributions to identifying various methane-emitting facilities.

RGB imagery effectively represents surface reflectance characteristics, capturing clear object shapes, boundaries, and reflectance contrasts. Facilities with well-defined geometries and clear boundaries are easily distinguishable in RGB images. For example, barns in CAFOs typically appear as rectangular or elongated high-reflection features, sharply contrasting with surrounding farmland or pastures (Fig. 1a1). Similarly, storage tanks and structural elements in Proc Plants, R&Ts,

Table 1
Sample distribution of the METER-ML dataset across categories and data splits (training, validation, and test sets).

Category	Train	Valid	Test	Total
CAFOs	24,957	47	92	25,096
Landfills	3918	43	111	4072
Coal Mines	1768	40	72	1880
Proc Plants	1836	38	107	1981
R&Ts	3891	58	108	4057
WWTPs	14,501	40	129	14,670
Negatives	33,974	249	647	34,870
Total	84,845	515	1266	86,626

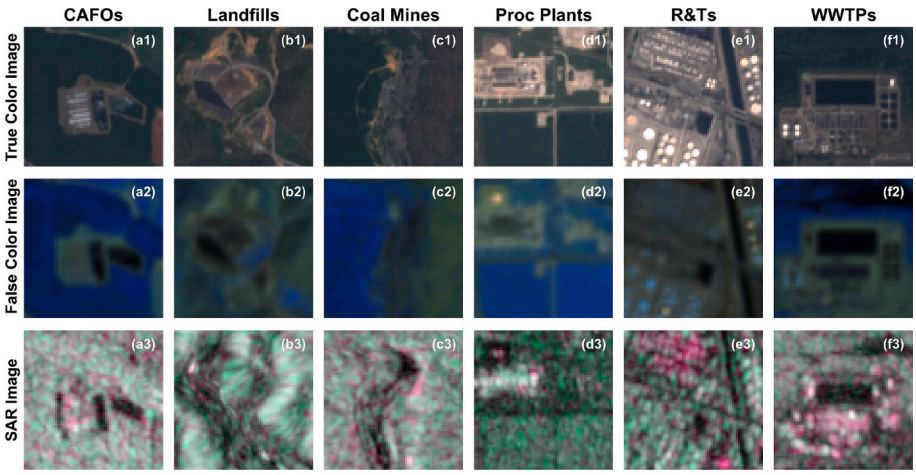


Fig. 1. Representative Sentinel-2 and Sentinel-1 imagery features of methane-emitting facilities. (Top row: Sentinel-2 true color imagery [RGB: Bands 4/3/2]; Middle row: Sentinel-2 false color imagery [RGB: Bands 12/11/8A]; Bottom row: Sentinel-1 SAR imagery [RGB: VV/VH/VV-VH ratio]).

and WWTPs are distinguishable from adjacent bare soils or buildings due to their higher reflectivity (Fig. 1d1–f1). However, RGB imagery is less effective for identifying facilities with complex terrain or minimal reflectance differences, such as coal mine pits that blend into the background environment (Fig. 1b1).

IR imagery is sensitive to high-temperature thermal emissions, as well as surface moisture and material composition, making it particularly effective for identifying facilities that exhibit distinct thermal or spectral anomalies. For example, flare burners in Proc Plants often appear as circular or elliptical high-intensity spots in the SWIR bands, representing high-temperature anomalies (HTAs) caused by industrial flaring, and are typically located at facility edges or within isolated bare areas surrounded by safety buffer zones (Fig. 1d2). In contrast, for facilities such as CAFOs and wastewater treatment plants, IR imagery captures spectral reflectance differences associated with liquid manure ponds or wastewater basins, which show clear contrasts with surrounding land cover (Fig. 1a2, f2). However, facilities with minimal thermal activity or subtle spectral differences, such as landfills or coal mines, often lack distinctive signatures in these bands, making their identification more challenging.

SAR imagery reflects surface roughness and structural characteristics, making it ideal for identifying facilities with strong scattering features. For instance, the edges of coal mine pits and mining traces appear as distinct linear or striped patterns, clearly delineating excavation boundaries and structures (Fig. 1c3). Metal storage tanks at refineries and oil terminals also produce strong backscatter signals, resulting in prominent high-scattering features, particularly noticeable with spherical tanks (Fig. 1e3). However, SAR imagery is highly sensitive to surface roughness; facilities with smoother surfaces, like certain WWTP structures, generate weaker backscatter signals and are therefore more challenging to distinguish.

Each modality offers unique advantages and inherent limitations, and thus no single modality alone fully addresses the identification requirements across all facility types. Integrating the spatial detail of RGB, the thermal sensitivity of d, and the structural insights of SAR is therefore essential for improving the accuracy and robustness of methane source identification.

3. Methodology

3.1. Construction of dynamic multimodal fusion model

3.1.1. Overall structure of DMFNet

The DMFNet architecture is guided by the need for an adaptive and effective fusion framework that optimally integrates multimodal features for methane source infrastructure classification. As illustrated in Fig. 2, DMFNet consists of three parallel ResNet-50 branches, independently extracting modality-specific features to avoid cross-modal interference and preserve distinct characteristics. To accommodate varying input complexities, we introduce a dynamic fusion mechanism based on gating. The convolution gating module flexibly adjusts the fusion decision based on the specific characteristics of each input sample and selects one of five fusion paths for them. Furthermore, stacked MAFMs progressively enhance cross-modal interactions by selectively emphasizing complementary features and suppressing redundancy. Thus, DMFNet provides an adaptive, multi-level fusion framework that leverages modality-specific strengths, dynamically optimizing feature integration according to the distinct characteristics of facilities.

3.1.2. Convolution gating module

The gating module dynamically adjusts the fusion strategy and path by processing the global view of the multi-modal features of the input samples. For “simple” samples, the gating mechanism prioritizes low-level features for quick prediction, skipping the complex fusion process to avoid unnecessary computation. On the other hand, for “complex” samples, the gating mechanism gradually merges features from different modalities in multiple stages to extract deeper information.

Specifically, as shown in Fig. 3, feature maps from the RGB, IR, and SAR modalities are first concatenated along the channel dimension to form a global view, x. This global view is then fed into the gating module $G(x)$ for processing. Through convolution operations, the gating module extracts local information from the input feature maps and learns low-level features between different modalities. The resulting feature maps are then passed through a global average pooling layer for dimensionality reduction, capturing global semantic information. This process not only reduces computational complexity but also preserves important contextual information. Finally, the extracted features are processed by a linear layer and mapped to a sparse decision space, generating a decision vector g that provides effective input for the subsequent fusion decisions.

As a hard gating mechanism, the output g of the gating module is represented as a five-dimensional sparse one-hot vector:

$$g = (g_1, g_2, g_3, g_4, g_5),\ g_i \in \{0, 1\},\ \sum_{i=1}^{5} g_i = 1 \tag{1}$$

where each dimension g_i corresponds to a MAFM of each stage and determines the fusion strategy. Specifically: when $g_i = 1$, fusion is performed up to and including the i-th MAFM, after which further fusion is halted. The fused representation from the selected stage is retained for

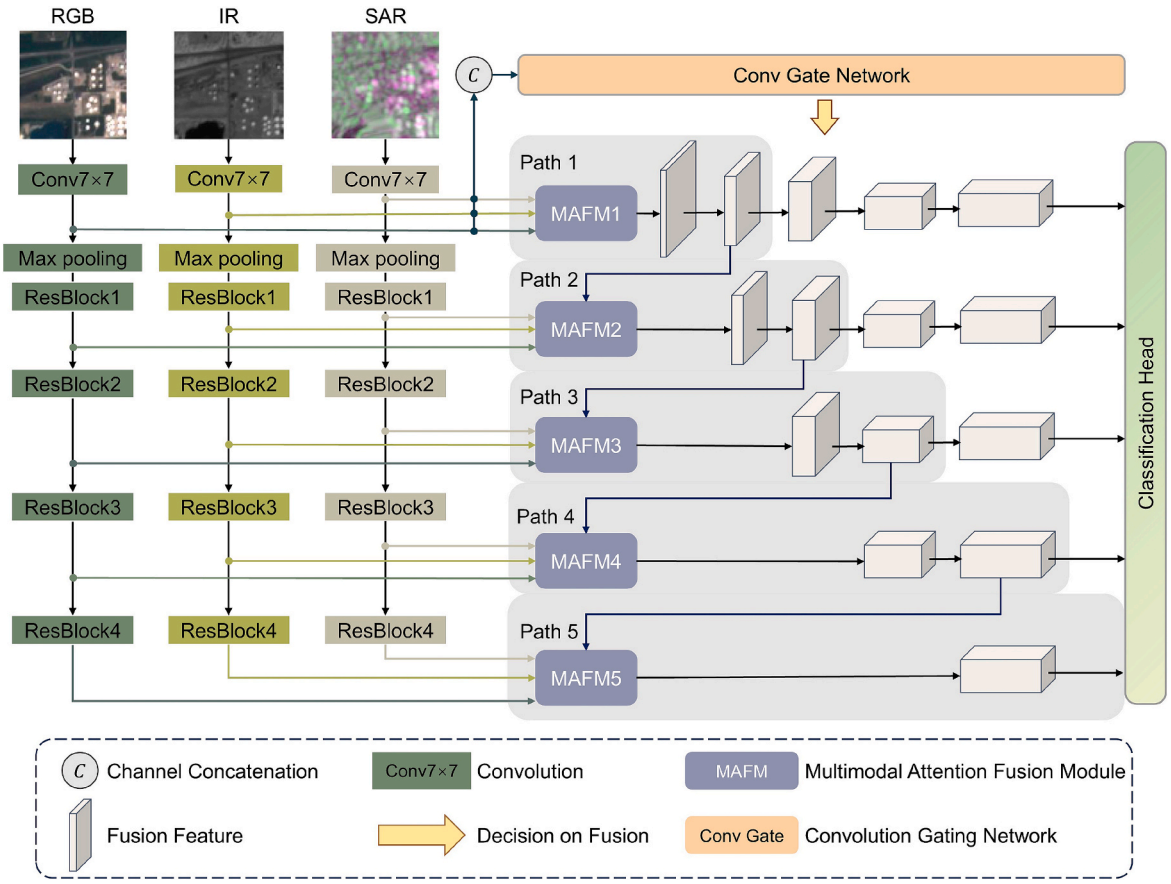


Fig. 2. Architecture of the proposed DMFNet. (Pre-trained ResNet-50 branches extract RGB, IR, and SAR features; the convolutional gating module generates adaptive fusion decisions; MAFM dynamically integrates multimodal features.)

subsequent feature processing and classification.

It is important to note that the gating vector g is generated through a learning and discretization process during the second-stage fine-tuning, where the gating module is jointly optimized with the main network. Specifically, the gating module first outputs a five-dimensional continuous vector, where each value is learned by minimizing the overall loss function to represent the relative importance of each fusion stage for a given input sample. This continuous vector is then processed through a one-hot encoding operation, converting it into a sparse binary vector, where g_i takes a value of 1 at the selected fusion stage and 0 elsewhere. This mechanism defines five distinct dynamic fusion paths, allowing the model to adaptively adjust the fusion process based on input characteristics.

The fusion process follows the gating vector configuration, dictating where multimodal feature integration occurs. The fused representation Y, serving as the input to the final classifier, is computed as:

$$Y = g_1 \bullet FB_{1-4}\left(X_1^{fused}\right) + g_2 \bullet FB_{2-4}\left(X_2^{fused}\right) + g_3 \bullet FB_{3-4}\left(X_3^{fused}\right) + g_4 \bullet FB_4\left(X_4^{fused}\right) + g_5 \bullet X_5^{fused} \quad (2)$$

$$X_1^{fused} = MAFM_1\left(X_0^{RGB}, X_0^{IR}, X_0^{SAR}\right) \quad (3)$$

$$X_i^{fused} = MAFM_i\left(X_{i-1}^{RGB}, X_{i-1}^{IR}, X_{i-1}^{SAR}, \widetilde{X}_{i-1}^{fused}\right), i \in \{2,3,4,5\} \quad (4)$$

$$\widetilde{X}_{i-1}^{fused} = FB_{i-1}\left(X_{i-1}^{fused}\right) \quad (5)$$

where i denotes the stage index of the MAFM (ranging from 1 to 5). $FB_{i-4}(\bullet)$ denotes the fusion block spanning from stage i to stage 4. X_i^{fused} and X_{i-1}^{fused} are the fused feature generated at the i^{th} and $(i-1)^{th}$ MAFMs, respectively. $\widetilde{X}_{i-1}^{fused}$ represents the fused output from the previous stage after being processed by the corresponding fusion block $FB_{i-1}(\bullet)$, which ensures spatial and channel alignment with the modality-specific features. Specifically, for stage $i = 1$, there is no prior fusion input, so $MAFM_1$ directly fuses the initial features from the three modalities. For stages $i > 1$, each $MAFM_i$ integrates both the current modality-specific features and the aligned fused representation $\widetilde{X}_{i-1}^{fused}$.

For example, when $g = (0,1,0,0,0)$, fusion occurs in the first two MAFM stages, allowing both early-stage and intermediate-level feature integration. Unimodal features are first fused in the first MAFM to obtain the initial fused representation. These features, along with the unimodal outputs processed by their respective blocks, are then further fused in the second MAFM. The fused features in the second stage are processed by fusion block and classifier to get the result. This process can be expressed as follows:

$$X_1^{fused} = MAFM_1\left(X_0^{RGB}, X_0^{IR}, X_0^{SAR}\right) \quad (6)$$

$$\widetilde{X}_1^{fused} = FB_1\left(X_1^{fused}\right) \quad (7)$$

$$X_2^{fused} = MAFM_2\left(X_1^{RGB}, X_1^{IR}, X_1^{SAR}, \widetilde{X}_1^{fused}\right) \quad (8)$$

b $Y = FB_{2-4}\left(X_2^{fused}\right)$ (9)

where: $X_0^{RGB}, X_0^{IR}, X_0^{SAR}$ represent the input features from the three modalities. $X_1^{RGB}, X_1^{IR}, X_1^{SAR}$ are the modality-specific features extracted after processing through Block1. X_1^{fused}, X_2^{fused} are the fusion feature

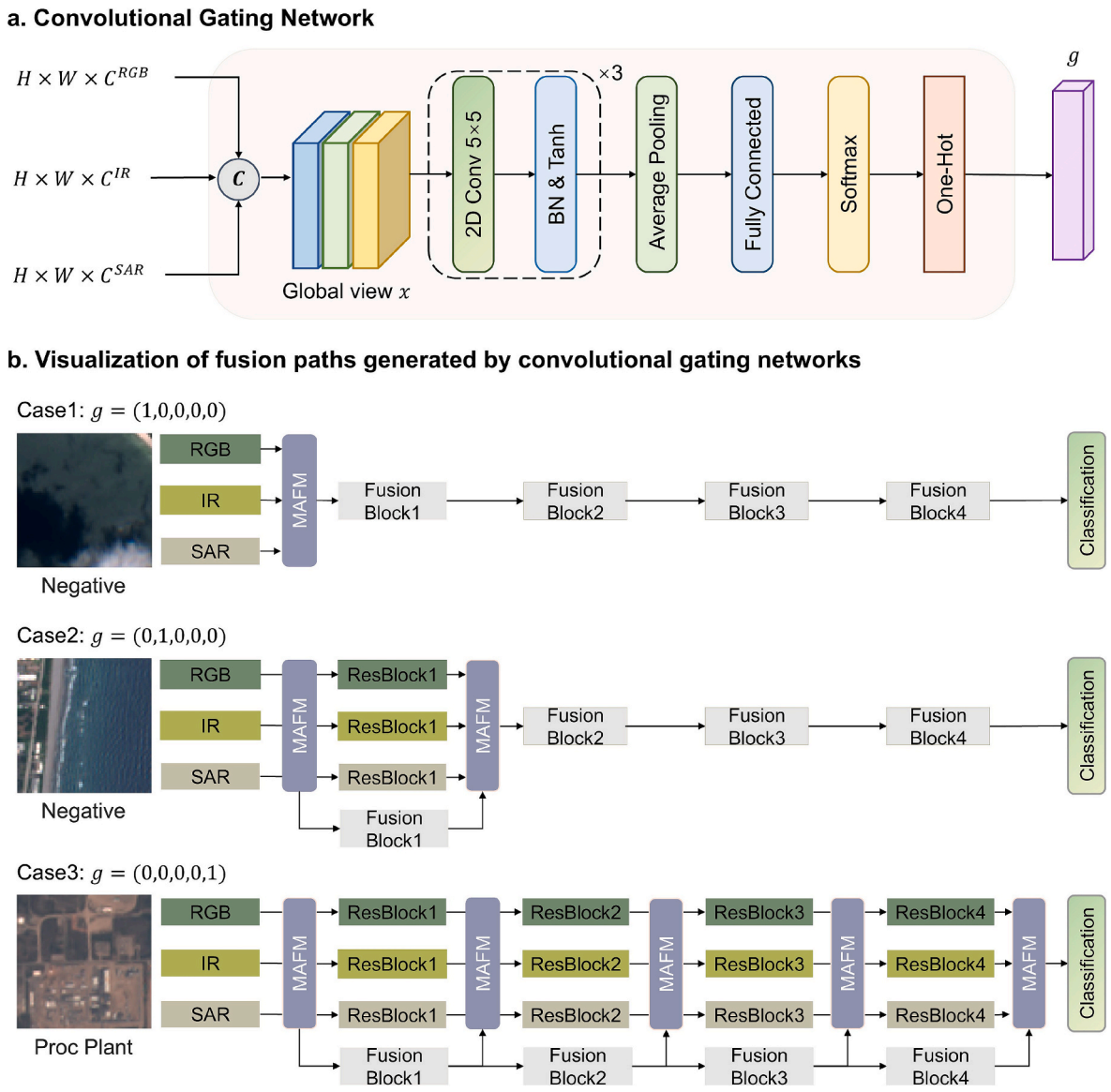


Fig. 3. Convolutional gating module and dynamic fusion path examples. (a) Structure of the convolutional gating module. (b) Sample-specific fusion paths: negative samples ("simple") utilize early fusion, while complex samples (e.g., Proc Plant) require progressive, multi-stage fusion.)

representations obtained at the first and second MAFM stages, respectively.

When $g = (0,0,0,0,1)$, fusion is performed at all five MAFM stages, meaning that unimodal information is continuously integrated throughout the network. This results in the most comprehensive fusion representation, maximizing the utilization of multimodal information:

$$X_i^{fused} = MAFM_i\left(X_{i-1}^{RGB}, X_{i-1}^{IR}, X_{i-1}^{SAR}, \widetilde{X}_{i-1}^{fused}\right), \mathrm{i} = 2,3,4,5 \tag{10}$$

$$\widetilde{X}_{i-1}^{fused} = FB_{i-1}\left(X_{i-1}^{fused}\right) \tag{11}$$

$$Y = X_5^{fused} \tag{12}$$

3.1.3. Multimodal attention fusion module(MAFM)

Effective multimodal fusion must balance complementarity and redundancy while adapting to hierarchical feature extraction in dynamic networks. Traditional fusion methods typically operate at fixed network stages, limiting their ability to progressively refine multimodal features or dynamically adjust modality contributions. Simple concatenation or summation may amplify redundancy, whereas rigid fusion strategies overlook variations in modality relevance at different depths.

MAFM addresses these challenges by progressively refining and integrating multimodal representations through three sequential stages: channel attention, modality attention, and spatial attention. Unlike CBAM (Woo et al., 2018), which applies channel and spatial attention sequentially within a single feature map, MAFM extends this paradigm to the multimodal setting by explicitly modeling cross-modal relationships and incorporating fused representations across stages (Fig. 4). Formally, let

$$X = \left\{X^{RGB}, X^{IR}, X^{SAR}, X^{fused}\right\} \in \mathbb{R}^{C\times H\times W} \tag{13}$$

where X^{RGB}, X^{IR}, X^{SAR} are modality-specific features, and X^{fused} is the intermediate fused representation.

We first apply a channel attention mechanism to adaptively reweight informative channels across modalities:

$$M_c(X) = \sigma(W_2\delta(W_1 GAP(X))) \tag{14}$$

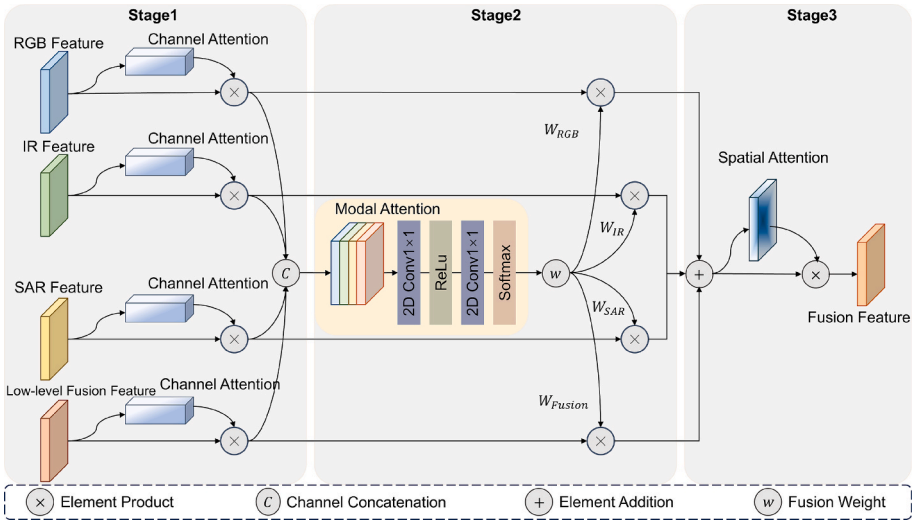


Fig. 4. Internal structure of the MAFM. (Three-stage adaptive fusion: channel attention enhancement, modality attention fusion, and spatial attention refinement.)

$$\widetilde{X}_c = M_c(X) \odot X \tag{15}$$

where $GAP(\bullet)$ denotes global average pooling, W_1, W_2 are learnable parameters, δ is ReLU, and σ is the sigmoid activation, and $\odot$ denotes element-wise multiplication. The output $\widetilde{X}_c$ represents the channel-refined feature.

Next, in the modality attention fusion stage, to balance the contributions of different modalities, a modality attention vector is introduced:

$$\widehat{X} = Concat_c\left(\widetilde{X}_c^{RGB}, \widetilde{X}_c^{IR}, \widetilde{X}_c^{SAR}, \widetilde{X}_c^{fused}\right) \tag{16}$$

$$\alpha = softmax(Conv_{1*1}(\delta(Conv_{1*1}(\widehat{X})))) = \left[\alpha_{RGB}, \alpha_{IR}, \alpha_{SAR}, \alpha_{fused}\right] \tag{17}$$

$$\widetilde{X}_m = \sum_{k\in\{RGB,IR,SAR,fused\}} \alpha_k \bullet \widetilde{X}_c^k \tag{18}$$

where α is the modality attention vector that adaptively weights the contribution of each modality.

Finally, spatial attention is applied to emphasize informative regions in the fused feature:

$$M_s\left(\widetilde{X}_m\right) = \sigma\left(f^{7\times7}\left(\left[AvgPool\left(\widetilde{X}_m\right); MaxPool\left(\widetilde{X}_m\right)\right]\right)\right) \tag{19}$$

$$X_{MAFM} = M_s\left(\widetilde{X}_m\right) \odot \widetilde{X}_m \tag{20}$$

where $f^{7\times7}$ is a convolution with a 7×7 kernel.

In summary, CBAM enhances single-modality features via channel and spatial attention, whereas MAFM introduces an additional modality attention stage to explicitly weight and integrate multiple modalities (RGB, IR, SAR) and fused representations from previous stages. This design not only refines modality-specific information but also adaptively balances cross-modal contributions, enabling more effective multimodal classification.

3.2. Analysis of fusion routes and contributions based on multimodal data

Due to the diversity among methane emission facilities, optimal multimodal data combinations and fusion strategies vary significantly by facility type. To address this issue, we propose an adaptive strategy within the multimodal dynamic fusion framework that outputs fusion paths and corresponding fusion weights for each sample during the post-hoc analysis stage. Specifically, the gating module outputs a five-dimensional one-hot encoded vector representing the adaptive fusion path chosen for each input sample, enhancing traceability and interpretability of the fusion process. Additionally, the MAFMs provide modality-specific fusion weights via a SoftMax layer, allowing quantitative assessment of modality contributions. By systematically visualizing fusion paths and analyzing these learned fusion weights, we explicitly track the contributions and interactions of individual modalities in the identification process.

3.3. Locating facility in transferability experiments

Applying the trained model to real-world facility location involves additional considerations, including appropriate image patch selection, determining facility locations, and validation procedures. This section details the implementation steps for transferability experiments.

Selecting an appropriate patch size is crucial since overly large patches introduce unnecessary noise, while excessively small patches risk fragmenting facilities. Considering the optimal balance between spatial context and resolution, we adopted a patch size of 720 m × 720 m, consistent with prior studies demonstrating optimal model performance at this scale (Zhu et al., 2022). All image bands were resampled to a unified 10-meter resolution via bilinear interpolation, resulting in 29,382 image patches covering the Los Angeles area. Although bilinear interpolation does not guarantee complete semantic consistency across modalities, it preserves spatial correspondence and retains sufficient spectral and structural information for classification. The processed patches are then fed into DMFNet to obtain predicted probabilities. Class labels are assigned using a probability threshold optimized for the highest AUPRC (Area Under the Precision-Recall Curve) value (0.778).

Since the predicted positive patches do not always align precisely with actual facility locations, additional steps are required to refine the mapping. Although prior research (Robinson et al., 2022) proposed combining Class Activation Maps (CAM) and K-means clustering to determine facility coordinates within patches. However, at a 10-meter resolution, CAM results exhibit high variability across facility types, limiting the reliability of this automated approach. Therefore, we manually reviewed all positive predictions, cross-referencing historical imagery from Google Maps and Google Earth to exclude false positives. Adjacent patches classified as the same facility type were merged, and the centroid of the merged area was taken as the estimated facility location.

For validation, identified facilities were compared against methane point source data from Carbon Mapper. Table 2 outlines how our facility types correspond to the IPCC-defined sectors used by Carbon Mapper. A

Table 2
Methane source infrastructure categories and corresponding IPCC sectors.

Methane source infrastructure categories	Corresponding IPCC sectors
Concentrated Animal Feeding Operations (CAFOs)	Enteric Fermentation (4A), Manure Management (4B)
Landfills	Solid Waste Disposal on Land (6A)
Coal Mines	A Coal Mining (1B1)
Natural Gas Processing Plants (Proc Plants)	Oil and Natural Gas (1B2)
Refineries and Oil Terminals (R&Ts)	Petroleum Refining (1A1b)
Wastewater Treatment Plants (WWTPs)	Waste Water Handling (6B)

predicted facility was confirmed as correctly identified if a corresponding Carbon Mapper methane emission source fell within the boundary of the positively detected patch.

3.4. Implementation and training detail

We employ *Gumbel SoftMax* with a two-stage training strategy, consisting of pretraining and fine-tuning, to jointly optimize the dynamic network and gating module (Jang et al., 2016; Xue and Marculescu, 2023). During pretraining, the gating module adopts random path selection to ensure sufficient training of all dynamic paths prior to active gating decisions. In the fine-tuning stage, the gating module dynamically generates fusion paths, enabling end-to-end optimization of the entire framework and achieving optimal fusion strategies.

All experiments are implemented in Python 3.8 using PyTorch and accelerated by an NVIDIA GeForce RTX 3080 10G GPU. We initialize the model with ImageNet pretrained weights for faster convergence. A batch size of 64 and an initial learning rate of 0.001 are used. During pretraining, the model is trained for 100 epochs. In the fine-tuning stage, the *Gumbel SoftMax* temperature parameter (τ) decays exponentially from 1 to 0.0001 over 100 epochs, while other hyperparameters remain unchanged. We adopt a weighted cross-entropy loss to address class imbalance in the 7-class classification task. The loss function is defined as:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{7} w_c \bullet y_{i,c}\log\left(\hat{y}_{i,c}\right) \tag{21}$$

where N is the number of training samples, $c \in \{1, 2, \cdots, 7\}$ indexes the 7 classes, $y_{i,c} \in \{0, 1\}$ denotes the one-hot encoded ground-truth label for sample i, $\hat{y}_{i,c} \in [0, 1]$ is the predicted probability for class c, weight w_c assigned to each class is computed based on the inverse of its relative frequency in the training set:

$$w_c = \frac{1/f_c}{\sum_{j=1}^{7}\left(1/f_j\right)} \tag{22}$$

with f_c is the proportion of class c samples in the training data. To further stabilize training and prevent overfitting, we apply weight decay regularization with a coefficient of 1e-4, and employ a cosine annealing learning rate schedule throughout the training process.

4. Results

4.1. Accuracies of the proposed DMFNet model

4.1.1. Performance comparison of models on METER-ML test set

We conducted experiments on the METER-ML dataset to evaluate the necessity of multimodal data fusion and verify the effectiveness of the proposed dynamic fusion approach. Specifically, we compared single-modality models with various static and advanced multimodal fusion methods. Models 1–3 used ResNet-50 as the backbone architecture, taking RGB, IR, or SAR images as single-modal inputs, respectively.

For Models 4–9, we evaluated six static multimodal fusion methods, all of which incorporated three modalities (RGB, IR, SAR) as inputs: Model 4: Pixel-level fusion, where images from different modalities were stacked along the channel dimension and fed directly into ResNet-50. Model 5: Feature-level fusion (Liang et al., 2021), which extracted features from each modality-specific ResNet-50 stream and fused them before classification. Model 6: MI-Matrix fusion (Jayakumar et al., 2020), which fused modality features through mutual information maximization to enhance complementary information integration. Model 7: Dense Fusion (Hong et al., 2021), employing a densely connected fusion module to hierarchically aggregate multimodal features. Model 8: An MHSA-enabled CNN architecture (Bansal and Tripathi, 2024), incorporating multiple dual-scale convolutional and multi-head self-attention blocks. Model 9: CGINet (Zhao et al., 2024), a fusion model that combines global context and part-level discriminative features. To further evaluate the effectiveness of dynamic fusion, we tested Models 10–13, which explored different modality combinations using the proposed dynamic fusion strategy. All models were trained under consistent environments and parameter settings to ensure fair comparisons.

Table 3 summarizes performance on the METER-ML test set. Among unimodal models, the RGB-based model (Model 1) performs best (F1-score: 0.648; AUPRC: 0.685), indicating that RGB imagery provides richer spatial–textural cues for single-modality classification, whereas the SAR-based model (Model 2) performs the worst among unimodal approaches.

For static multimodal fusion (Models 4–9), most methods outperform the unimodal baselines, confirming the benefit of complementary modalities. Early fusion (Model 4) attains high recall (0.750) but low precision (0.444), suggesting that naïve channel stacking introduces redundancy and noise. The MHSA-enabled CNN (Model 8) and CGINet (Model 9) further improve results, with Model 8 reaching an F1-score of 0.734 and AUPRC of 0.761.

For dynamic fusion (Models 10–13), the proposed DMFNet (Model 13) achieves the best overall performance across recall, macro-F1, and AUPRC. Relative to static method (Model 8), DMFNet increases precision by 1.1 percentage points, recall by 1.9 points, macro-F1 by 1.1 points, and AUPRC by 1.7 points, indicating more effective suppression of redundancy and exploitation of cross-modal complementarity. Although CGINet (Model 9) has slightly higher precision (0.742 vs. 0.740), DMFNet attains higher recall (0.757 vs. 0.719), macro-F1 (0.745 vs. 0.730), and AUPRC (0.778 vs. 0.758). Given the class imbalance and our workflow with expert verification of model candidates, we prioritize recall and macro-F1/AUPRC to minimize missed facilities (false positives can be screened during review, whereas false negatives are typically unrecoverable). Accordingly, DMFNet is preferred overall.

Table 3
Macro-averaged model performance on the METER-ML dataset. (Models 1–3: Unimodal; Models 4–9: Static multimodal fusion; Models 10–13: Dynamic multimodal fusion; Model 13: Proposed DMFNet.)

No	Method	Modal	Precision	Recall	F1	AUPRC
1	Unimodal	RGB	0.654	0.662	0.648	0.685
2	Unimodal	SAR (VV, VH)	0.510	0.489	0.475	0.505
3	Unimodal	IR (NIR, SWIR1-2)	0.687	0.615	0.622	0.679
4	Early Fusion	RGB, SAR, IR	0.444	0.750	0.558	0.641
5	Late Fusion		0.697	0.676	0.662	0.714
6	MI-Matrix		0.662	0.667	0.651	0.701
7	Dense Fusion		0.723	0.721	0.713	0.722
8	MHSA-enabled		0.729	0.738	0.734	0.761
9	CGINet		**0.742**	0.719	0.730	0.758
10	DMFM	RGB, SAR	0.722	0.724	0.721	0.751
11	DMFM	RGB, IR	0.732	0.749	0.737	0.768
12	DMFM	SAR, IR	0.691	0.663	0.663	0.729
13	DMFM	RGB, SAR, IR	0.740	**0.757**	**0.745**	**0.778**

Moreover, compared with dynamic models using only two modalities (Models 10–12), the three-modality dynamic fusion of DMFNet (Model 13) consistently performs best, reinforcing the value of integrating all modalities for facility identification.

Fig. 5 presents the confusion matrix, highlighting key misclassification patterns. Overall, the model achieved high accuracy across most categories; however, notable confusions occurred between facility types with similar features. Specifically, Proc Plants and R&Ts were frequently misclassified as each other, reflecting their overlapping industrial structures (Fig. 5 c and d), while Landfills and Mines also showed considerable confusion due to similar exposed surface materials and geometric layouts (Fig. 5e). Additionally, some negative samples were misclassified as facility types, such as bare soil classified as Landfills and port areas identified as R&Ts (Fig. 5 e and f).

4.1.2. Ablation experiment

We conducted an ablation study on the METER-ML dataset to assess the individual contributions and combined effectiveness of the Gating module and the Multimodal Attention Fusion Module (MAFM). The initial baseline model (BASE) comprises three modality-specific ResNet-50 branches, with the extracted features directly concatenated after global average pooling and passed to the classifier—without any cross-modal fusion or interaction mechanism. To independently evaluate the Gating module, we removed the MAFM and retained the gating mechanism to dynamically select one of the predefined fusion paths. Within each selected path, features from different modalities were merged using element-wise addition to maintain feature dimensionality and ensure compatibility with downstream layers. Conversely, to isolate the effect of the MAFM, we removed the Gating module and adopted a random path selection strategy in its place. In addition, we tested each of the five predefined fusion paths (Path 1 to Path 5) individually to examine performance sensitivity to specific fusion configurations.

Table 4 summarizes the ablation results. Incorporating the gating module alone improved precision by 1.2 %, recall by 1.4 %, and F1 by 1.3 % compared to the baseline. Replacing addition with MAFM and using random path selection (No. 3) further enhanced performance, achieving an F1 of 0.729. Among the fixed fusion paths (No. 4–8), Path 3 (MAFM1–3) yielded the highest performance (F1: 0.732), suggesting that deeper fusion generally benefits classification. However, the proposed dynamic gating strategy (No. 9) achieved the best overall results, with precision, recall, and F1 improvements of 2.7 %, 4.8 %, and 4.3 %, respectively, relative to the baseline. These results demonstrate that the dynamic gating strategy outperforms the best fixed path configuration, highlighting the gating module's ability to adaptively select optimal fusion paths based on input characteristics rather than relying on a single pre-defined path. This adaptability enables the model to fully exploit modality complementarity under varying data conditions. Overall, these findings confirm the complementary roles of the gating module and MAFM, where the gating module dynamically determines effective fusion routes, while the MAFM enhances cross-modal feature interactions.

4.2. Explanation of multimodal dynamic fusion mechanisms

4.2.1. Analysis of multimodal data contribution

The contribution of multimodal data in identifying different methane-emitting facilities was analyzed based on the MAFM. As shown in Fig. 6, MAFM assigns higher weights to the optical modality (RGB) in the early fusion stage (Stage 1), indicating that visible light information remains dominant in facility identification. As the network depth increases, the weight assigned to the fused features progressively increases, while the weights of the three unimodal inputs decrease. This trend suggests that multimodal data synergy becomes more pronounced in deeper network layers, with fused features contributing most to the final decision, while unimodal features serve as auxiliary information to support recognition.

4.2.2. Dynamic fusion path visualization and analysis

Data-driven fusion weights highlight modality-specific contributions for different facility types. For R&Ts and Proc Plants, the model assigns higher weights to RGB and IR, as IR effectively captures thermal radiation from flares and heating units, while RGB provides clear spatial structure information. Conversely, landfills exhibit lower IR weights due to minimal thermal signatures. Coal mines primarily rely on RGB and SAR modalities, as SAR effectively captures structural patterns and surface roughness, while RGB provides clear visual context. WWTPs

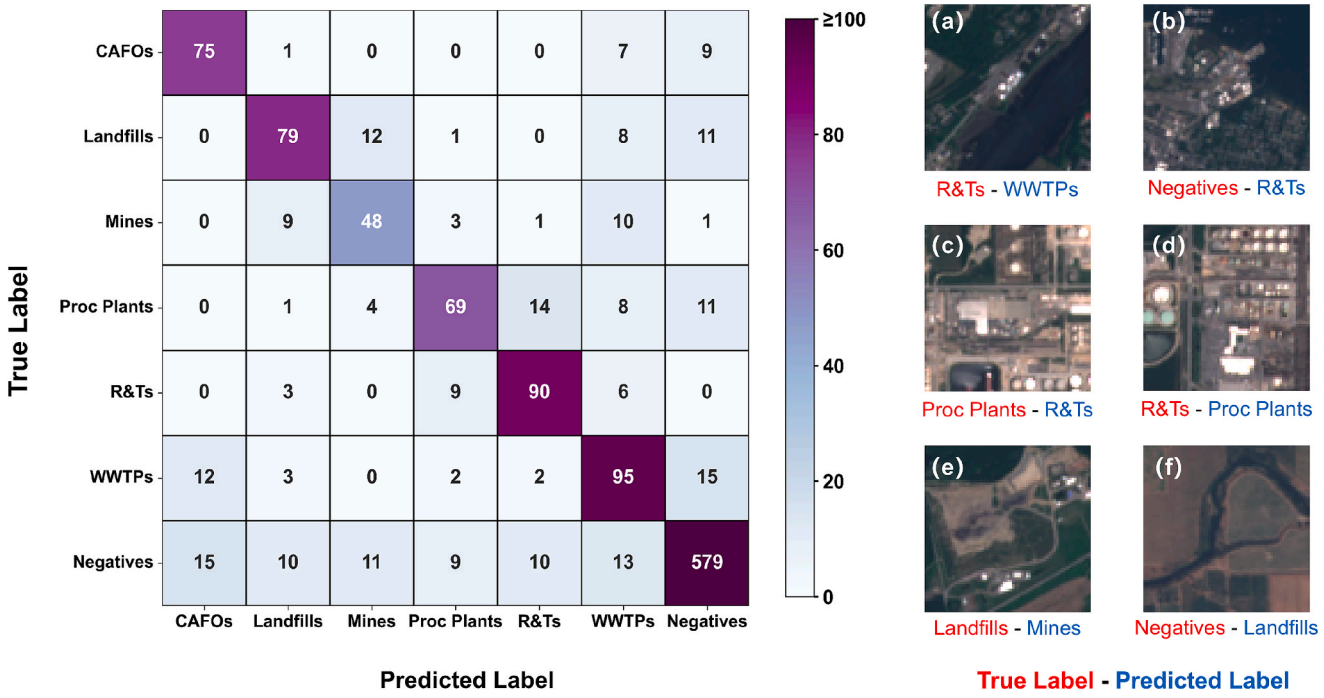


Fig. 5. Confusion matrix showing classification performance and examples of common misclassifications between similar facility types and negative samples.

Table 4
Ablation study results of DMFNet with different fusion strategies and paths.

No	Baseline	Gate module	Fusion method	Fusion path	Precision	Recall	F1	AUPRC
1	√		/	/	0.686	0.665	0.675	0.712
2	√	√	Addition	Dynamic	0.725	0.723	0.715	0.759
3	√		MAFM	Random	0.733	0.734	0.729	0.764
4	√		MAFM	Path 1	0.682	0.690	0.678	0.725
5	√		MAFM	Path 2	0.711	0.725	0.713	0.747
6	√		MAFM	Path 3	0.735	0.739	0.732	0.758
7	√		MAFM	Path 4	0.736	0.737	0.731	0.757
8	√		MAFM	Path 5	0.720	0.735	0.722	0.755
9	√	√	MAFM	Dynamic	**0.740**	**0.757**	**0.745**	**0.778**

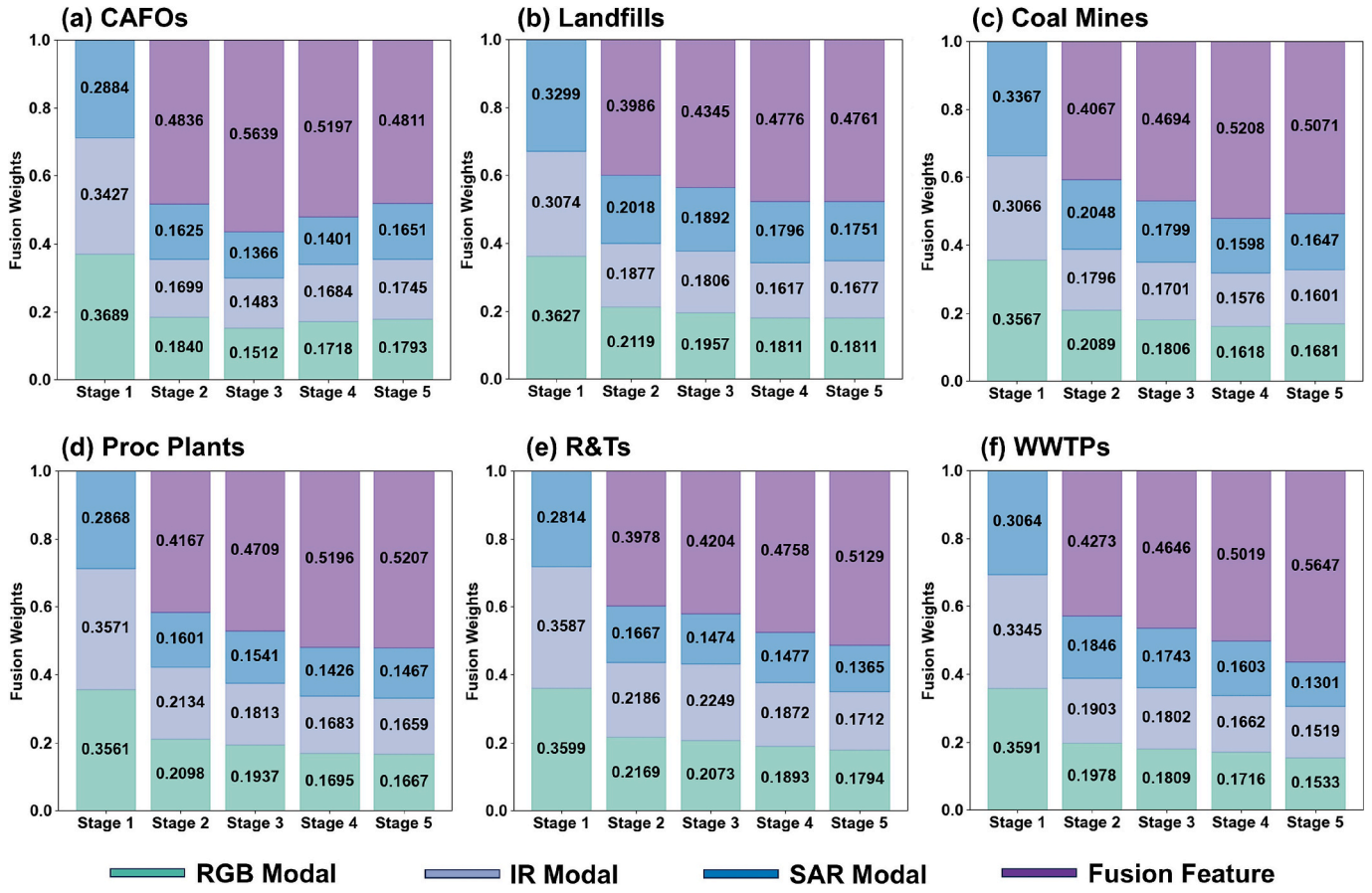


Fig. 6. Multimodal contribution analysis across facility types. (Stacked bars show proportional modality contributions at each fusion stage; lines illustrate modality weight variations across fusion stages.)

exhibit a more balanced reliance on RGB and IR, likely because they contain both highly reflective infrastructure and water bodies, as well as aeration and anaerobic digestion tanks with noticeable temperature variations.

In DMFNet, Paths 1 to 5 represent dynamic fusion pathways adaptively determined by the gating module. As the path number increases, the frequency of multimodal feature interactions grows, enhancing the model's complexity and representational capacity. Shallow fusion paths (Path 1–3) capture primarily low-level texture and shape features, while deep fusion paths (Path 4–5) extract high-level semantic information. Statistical analysis (Fig. 7) shows notable differences in fusion path selection among facility types. CAFOs predominantly utilize Path 3 (81.52 %), suggesting mid-level feature fusion sufficiently captures their structural characteristics. Landfills and Coal Mines mainly adopt Path 2 and Path 3, yet Coal Mines exhibit a higher frequency of deeper fusion (Path 4: 25.11 %) compared to Landfills (14.42 %), highlighting their greater structural complexity. Proc Plants, R&Ts, and WWTPs, characterized by complex layouts and functional diversity, heavily depend on deeper fusion paths (Path 4 and Path 5) to effectively distinguish semantically similar components.

4.3. *Detection results of model deployment*

In the transferability experiment, the model initially identified 228 positive image tiles across Los Angeles. Following manual verification using Google imagery, 171 tiles were confirmed as true positives, yielding an overall precision of 0.75, while 57 tiles were discarded as false positives. After merging adjacent positive tiles, we identified 116 methane-emitting facilities (Table 5), comprising 51 R&Ts, 45 landfills, 15 WWTPs, and 5 Proc Plants (Fig. 8). R&Ts constitute the largest facility group (43.9 %), primarily concentrated in industrial and port zones in southern Los Angeles. Due to the absence of a comprehensive reference inventory, per-category recall and F1-scores could not be calculated for the detection results.

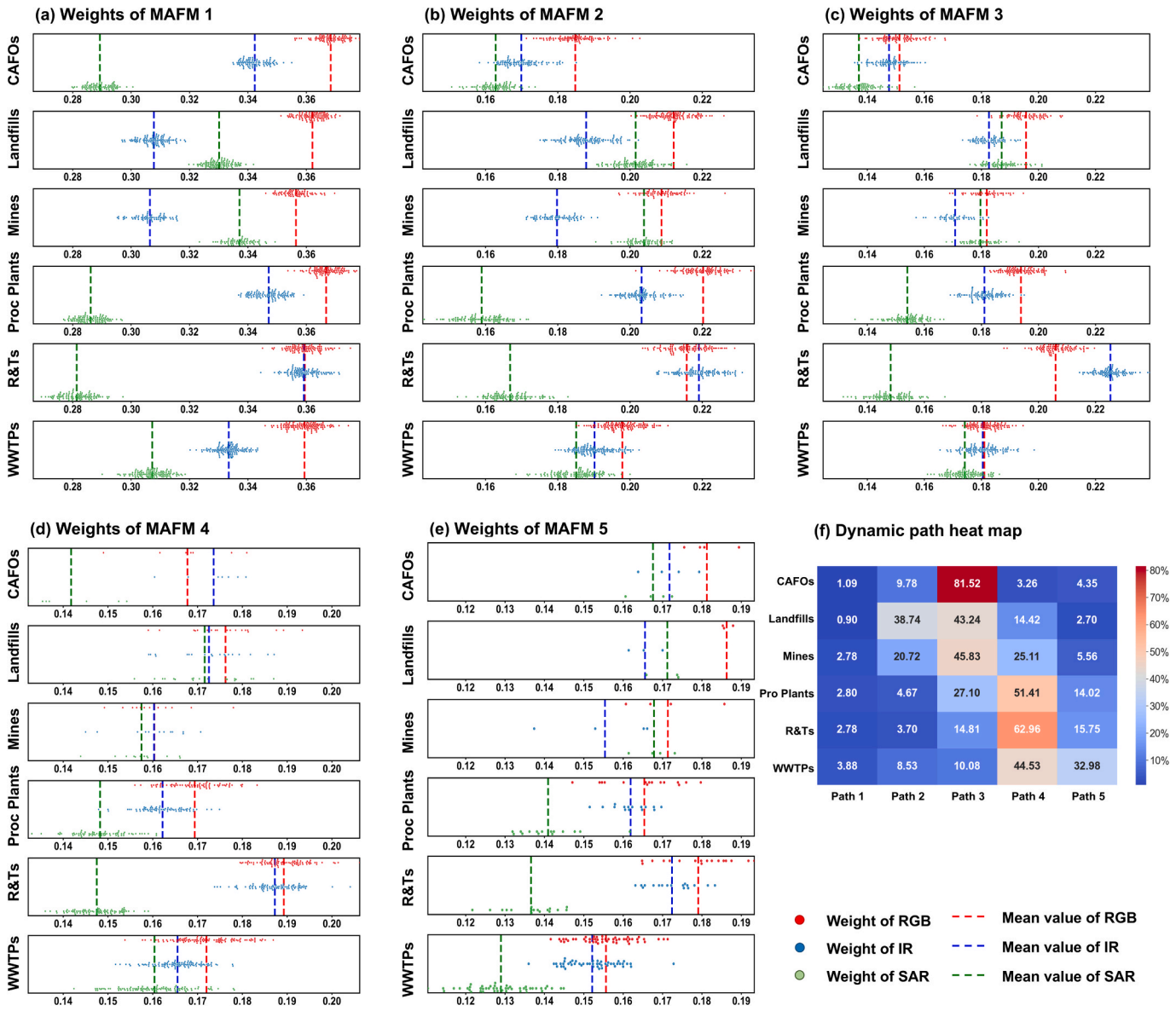


Fig. 7. Dynamic fusion path selection and modality weight distributions. (a–e) Modality weight distributions (RGB, IR, SAR) across fusion stages; (f) Heatmap of dynamic fusion path selection (Paths 1–5 indicate progressively deeper fusion).

Table 5
Transferability experiment results: facility-wise initial detections, validated detections, precision, and merged facility counts in the target region.

Category	Initial detections	False positives	Validated	Precision	Facility count
CAFOs	2	2	0	/	0
Landfills	82	20	62	0.756	45
Coal Mines	6	6	0	/	0
Proc Plants	8	2	6	0.750	5
R&Ts	98	16	82	0.837	51
WWTPs	32	11	21	0.656	15
Total	228	57	171	0.750	116

We validated our detection results against the Carbon Mapper Level 4B dataset, which includes 65 methane point sources within Los Angeles, comprising 8 solid waste, 24 petroleum refining, 26 oil and natural gas, 4 wastewater treatment, and 3 unspecified sources. Our method successfully detected 100 % of wastewater treatment sources, 95.8 % of petroleum refining sources, 62.5 % of solid waste sources, and 15.4 % of oil and natural gas sources (Table 6). Additionally, one unspecified source coincided with a landfill identified by our model. As shown in Fig. 9, missed detections in the oil and gas sector were primarily associated with methane releases from upstream and downstream components such as wellheads, pipelines, and urban gas distribution networks. These emissions often originate from fugitive leaks or venting in oilfields and gas fields, including associated gas flaring or venting at well sites, as well as aging pipeline infrastructure and faulty equipment in residential and industrial end-user systems. However, these sources are typically very small in size and lack distinct geometric or spectral characteristics at 10-meter resolution, making them difficult to detect using a facility-oriented classification approach that relies on visible infrastructure patterns in remote sensing imagery. Three missed solid waste facilities were small, irregularly shaped urban landfills, difficult to detect in medium-resolution imagery. One refinery source was also missed due to the absence of characteristic storage tanks, leading to negative classification. Notably, our method identified 80 potential methane-emitting facilities not listed by Carbon Mapper, including 40 landfills, 28 R&Ts, 11 WWTPs, and 1 Proc Plant.

Fig. 8. Detected methane-emitting facilities in Los Angeles. (Left: Overall facility distribution; Right: Enlarged views of facility types. Background imagery: Sentinel-2 true-color composite [RGB: Bands 4/3/2]).

Table 6
Statistics on methane source infrastructure detected by DMFNet and comparison with Carbon Mapper.

	Solid waste	Oil & gas	Petroleum refining	Waste water	Other
Total Detections	45	5	51	15	0
Coverage of Carbon Mapper	62.5 % (5/8)	15.4 % (4/26)	95.8 % (23/24)	100 % (4/4)	33.3 % (1/3)
Outside Carbon Mapper	40	1	28	11	0

5. Discussion

5.1. Advantages and limitations of medium-resolution imagery for identifying methane source infrastructure

Recent research on methane emissions has primarily focused on regional emission quantification (Shen et al., 2023; Vanselow et al., 2024), methane plume segmentation (Jahan et al., 2024; Rouet-Leduc and Hulbert, 2024), and leak detection at individual sources (Pandey et al., 2023; Watine-Guiu et al., 2023). Although these approaches significantly advance methane emission monitoring, they rely heavily on direct methane concentration measurements, which often introduce uncertainties in source attribution (Zhang et al., 2023). Complementing concentration-based methods, our approach emphasizes identifying methane emission infrastructure directly from remote sensing imagery, providing essential information such as facility types, quantities, and spatial distributions. This infrastructure-level perspective is critical for building more comprehensive and reliable emission inventories, ultimately supporting targeted policy interventions and mitigation strategies.

High-resolution imagery has conventionally been the preferred choice due to its superior spatial detail, enabling precise facility identification and plume attribution at the facility scale. However, its practical application remains limited by high acquisition costs, computational burdens, and constrained spatial coverage, particularly when applied over large geographic extents. In this context, medium-resolution imagery offers a practical balance between resolution and scalability, capturing diverse features including visible, thermal, and radar scattering information. Leveraging multimodal medium-resolution imagery, our study demonstrates that accurate facility recognition is achievable even without the cost and resource burdens of high-resolution data, exemplified by notably high AUPRCs for CAFOs (0.926) and R&Ts (0.881). This highlights the potential of medium-resolution imagery to facilitate large-scale methane infrastructure classification, addressing scalability constraints identified in previous studies.

Nonetheless, limitations remain, particularly in identifying coal mines and natural gas processing plants, where the 10 m spatial resolution limits boundary delineation and increases confusion with visually similar surfaces such as urban bare soils and reflective rooftops. Moreover, while the model performed well within the training region, its transferability to other geographic areas may be limited due to differences in background environments, facility sizes, structural layouts, and construction materials. Adapting the model with region-specific training data and fine-tuning will likely be necessary to maintain detection accuracy in new regions. Future improvements could focus on expanding the geographic diversity of training datasets to enhance generalization across regions. Additionally, integrating supplementary information to refine facility categorization (Li et al., 2022a) and leveraging multi-temporal imagery for time-series observations could further enhance the model's ability to accurately detect and dynamically monitor methane-emitting facilities.

5.2. Effectiveness of multimodal dynamic fusion strategies

The proposed DMFNet introduces a novel multimodal dynamic fusion strategy by combining gating and attention mechanisms, adaptively adjusting fusion pathways based on input characteristics. This adaptive approach allows the model to selectively utilize discriminative modality features tailored to each methane-emitting facility type, effectively addressing inter-sample variations and mitigating

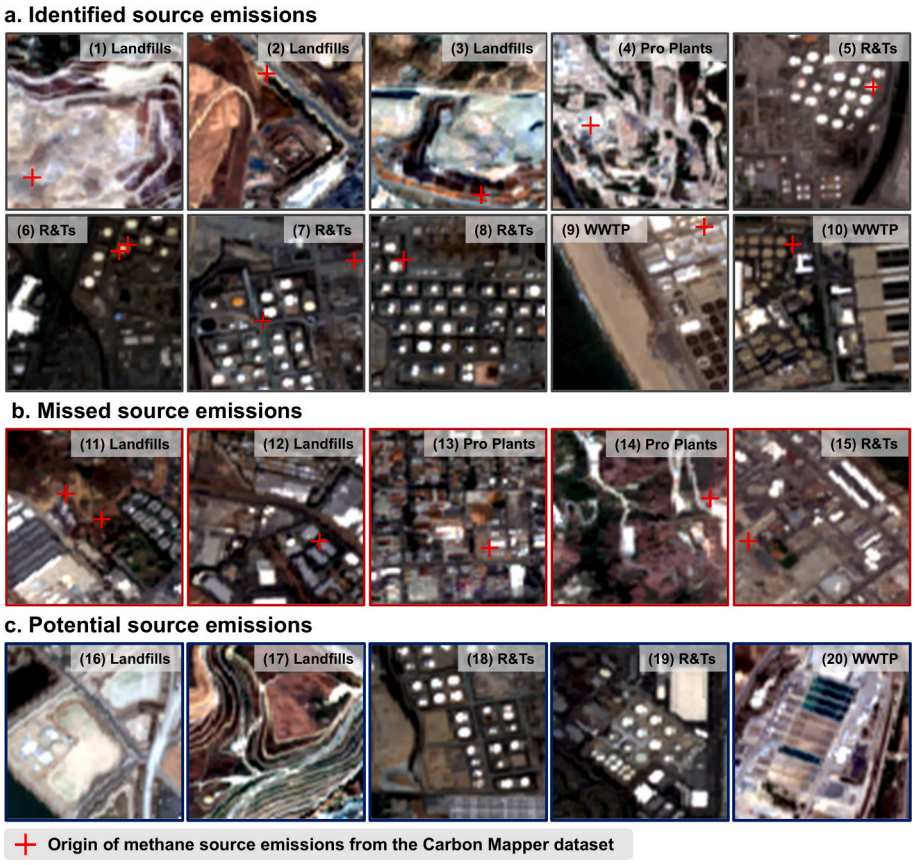


Fig. 9. Comparison between detected facilities and methane point sources. (a) Successful identifications covering point sources; (b) Missed point sources; (c) Potential methane-emitting facilities not listed as point sources.

redundancy or critical feature loss common in traditional fixed fusion methods. The quantitative analysis enabled by learned fusion weights further reveals distinct modality contributions for different facility categories, providing valuable insights into the multimodal synergy. For example, while optical imagery consistently dominates across facility types (Ren et al., 2022), IR imagery is specifically beneficial for detecting combustion-related facilities (e.g., R&Ts, Proc Plants), aligning well with recent findings (Liu et al., 2023). Despite its demonstrated effectiveness, DMFNet remains relatively complex regarding network architecture and training strategy. Future research could pay attention to developing more lightweight and efficient network designs to further optimize model performance in methane-emitting facility identification. Additionally, enhancing model interpretability through techniques such as SHAP (Chu et al., 2024; Lundberg and Lee, 2017), prototype network (Guo et al., 2024b), and attention-based explainability methods (Mena et al., 2025) could provide deeper insights into the decision-making process, improving the model's transparency and reliability.

6. Conclusion

In this study, we proposed DMFNet, a novel multimodal dynamic fusion model that adaptively integrates optical (RGB), infrared, and SAR data from medium-resolution Sentinel-2 and Sentinel-1 imagery for classifying methane emission facility. By leveraging complementary spectral, thermal, and radar scattering features, the proposed dynamic fusion mechanism significantly improves classification accuracy, effectively addressing limitations of single-modality methods and traditional static fusion approaches. Additionally, quantitative analyses of learned fusion weights reveal modality-specific contributions, clarifying how multimodal features collaboratively enhance facility-type recognition. Real-world transferability experiments further demonstrate our model's effectiveness and complementarity with existing methane emission databases, supporting improved emission inventory construction. Future research could explore the integration of multi-platform remote sensing sensors to further enhance detection fidelity, as well as incorporate temporal modeling techniques to enable dynamic, near-real-time facility monitoring.

CRediT authorship contribution statement

Yanglangxing He: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Xueliang Zhang:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Pengfeng Xiao:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Zhenshi Li:** Writing – review & editing, Methodology, Data curation. **Dilxat Muhtar:** Software, Methodology, Formal analysis. **Feng Gu:** Visualization, Investigation, Data curation. **Binxiao Liu:** Resources, Investigation, Formal analysis. **Pengming Feng:** Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 42522112, 42471410) and the AI and AI for

Science Project of Nanjing University (Grant No. 020914380141) We also extend our gratitude to the High-Performance Computing Center of Nanjing University for providing computational resources that facilitated this research.

Data availability

Data will be made available on request.

References

Bansal, K., Tripathi, A.K., 2024. An explainable MHSA enabled deep architecture with dual-scale convolutions for methane source classification using remote sensing. Environ. Modell. Softw. 181, 106178.

Beaumont, B., Radoux, J., Defourny, P., 2014. Assessment of airborne and spaceborne thermal infrared remote sensing for detecting and characterizing landfills. Waste Manag. 180, 237–248.

Berg, P., Pham, M.T., Courty, N., 2023. Joint Multi-Modal Self-Supervised Pre-Training in Remote Sensing: Application to Methane Source Classification, IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, pp. 6624–6627.

Berg, P., Uzun, B., Pham, M.T., Courty, N., 2024. Multimodal supervised contrastive learning in remote sensing downstream tasks. IEEE Geosci. Remote Sens. Lett. 21, 1–5.

Cai, S., Shu, Y., Wang, W., 2021. Dynamic routing networks, proceedings of the IEEE/CVF winter conference on applications of computer vision. IEEE 3588–3597.

Chen, Y., Sherwin, E.D., Berman, E.S., Jones, B.B., Gordon, M.P., Wetherley, E.B., Kort, E. A., Brandt, A.R., 2022. Quantifying regional methane emissions in the New Mexico Permian Basin with a comprehensive aerial survey. Environ. Sci. Technol. 56, 4317–4323.

Chu, W., Zhang, C., Li, H., Zhang, L., Shen, D., Li, R., 2024. SHAP-powered insights into spatiotemporal effects: Unlocking explainable Bayesian-neural-network urban flood forecasting. Int. J. Appl. Earth Obs. Geoinf. 131, 103972.

Ehalt Macedo, H., Lehner, B., Nicell, J., Grill, G., Li, J., Limtong, A., Shakya, R., 2022. Distribution and characteristics of wastewater treatment plants within the global river network. Earth Syst. Sci. Data 14, 559–577.

Gill, J., Faisal, K., Shaker, A., Yan, W.Y., 2019. Detection of waste dumping locations in landfill using multi-temporal Landsat thermal images. Waste Manag. Res. 37, 386–393.

Guo, J., Zhang, Z., Wang, M., Ma, P., Gao, W., Liu, X., 2024a. Automatic detection of subsidence funnels in large-scale SAR interferograms based on an improved-YOLOv8 Model. IEEE Trans. Geosci. Remote Sensing 62, 1–17.

Guo, Z., Hou, B., Guo, X., Wu, Z., Yang, C., Ren, B., Jiao, L., 2024b. MSRIP-Net: Addressing interpretability and accuracy challenges in aircraft fine-grained recognition of remote sensing images. IEEE Trans. Geosci. Remote Sensing 62, 1–17.

Handan-Nader, C., Ho, D.E., 2019. Deep learning to map concentrated animal feeding operations. Nat. Sustain. 2, 298–306.

Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y., 2022. Dynamic neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44, 7436–7456.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sensing 59, 4340–4354.

IEA (2024), Global Methane Tracker 2024, IEA, Paris https://www.iea.org/reports/global-methane-tracker-2024, Licence: CC BY 4.0.

Irvin, J., Tao, L., Zhou, J., Ma, Y., Nashold, L., Liu, B., Ng, A.Y., 2023. USat: A unified self-supervised encoder for multi-sensor satellite imagery. arXiv preprint arXiv: 2312.02199.

Jackson, R.B., Saunois, M., Bousquet, P., Canadell, J.G., Poulter, B., Stavert, A.R., Bergamaschi, P., Niwa, Y., Segers, A., Tsuruta, A., 2020. Increasing anthropogenic methane emissions arise equally from agricultural and fossil fuel sources. Environ. Res. Lett. 15, 071002.

Jahan, I., Mehana, M., Matheou, G., Viswanathan, H., 2024. Deep learning-based quantifications of methane emissions with field applications. Int. J. Appl. Earth Obs. Geoinf. 132, 104018.

Jang, E., Gu, S., Poole, B., 2016. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.

Jayakumar, S.M., Czarnecki, W.M., Menick, J., Schwarz, J., Rae, J., Osindero, S., Teh, Y. W., Harley, T., Pascanu, R., 2020. Multiplicative interactions and where to find them, International conference on learning representations.

Jiang, C., Ren, H., Yang, H., Huo, H., Zhu, P., Yao, Z., Li, J., Sun, M., Yang, S., 2024. M2FNet: Multi-modal fusion network for object detection from visible and thermal infrared images. Int. J. Appl. Earth Obs. Geoinf. 130, 103918.

Jin, J., Zhou, W., Ye, L., Lei, J., Yu, L., Qian, X., Luo, T., 2022. DASFNet: Dense-Attention–Similarity-Fusion Network for scene classification of dual-modal remote-sensing images. Int. J. Appl. Earth Obs. Geoinf. 115, 103087.

Kirschke, S., Bousquet, P., Ciais, P., Saunois, M., Canadell, J.G., Dlugokencky, E.J., Bergamaschi, P., Bergmann, D., Blake, D.R., Bruhwiler, L., 2013. Three decades of global methane sources and sinks. Nat. Geosci. 6, 813–823.

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

Lee, H., Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P., Trisos, C., Romero, J., Aldunce, P., Barret, K., 2023. IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.

Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., 2021. Multibench: Multiscale benchmarks for multimodal representation learning. Adv. Neural Inf. Proces. Syst. 2021, 1.

Y. Li L. Song Y. Chen Z. Li X. Zhang X. Wang J. Sun Li, Y., Song, L., Chen, Y., Li, Z., Zhang, X., Wang, X., Sun, J., 2020. Learning dynamic routing for semantic segmentation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8553-8562.

Li, H., Zech, J., Hong, D., Ghamisi, P., Schultz, M., Zipf, A., 2022a. Leveraging OpenStreetMap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection. Int. J. Appl. Earth Obs. Geoinf. 110, 102804.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022b. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926.

Li, Y., Zhang, Y., Randhawa, S., Yang, C., Zipf, A., 2025. STVAE: Skip connection driven Two-stream property fusion Variational AutoEncoder for cross-region wastewater treatment plant semantic segmentation. Inf. Fusion 118, 102960.

Liu, X., Hong, D., Chanussot, J., Zhao, B., Ghamisi, P., 2021a. Modality translation in remote sensing time series. IEEE Trans. Geosci. Remote Sensing 60, 1–14.

Liu, Y., Zhi, W., Xu, B., Xu, W., Wu, W., 2021b. Detecting high-temperature anomalies from Sentinel-2 MSI images. ISPRS-J. Photogramm. Remote Sens. 177, 174–193.

Liu, Y., Pu, Y., Hu, X., Dong, Y., Wu, W., Hu, C., Zhang, Y., Wang, S., 2023. Global declines of offshore gas flaring inadequate to meet the 2030 goal. Nat. Sustain. 6, 1095–1102.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 4768–4777.

Masson-Delmotte, V., Zhai, P., Pirani, S., Connors, C., Péan, S., Berger, N., Caud, Y., Chen, L., Goldfarb, M., Scheel Monteiro, P.M., 2021. Ipcc, 2021: Summary for policymakers. in: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change.

Maus, V., Giljum, S., da Silva, D.M., Gutschlhofer, J., da Rosa, R.P., Luckeneder, S., Gass, S.L., Lieber, M., McCallum, I., 2022. An update on global mining land use. Sci. Data 9, 433.

Mena, F., Pathak, D., Najjar, H., Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Miranda, M., Siddamsetty, J., Nuske, M., Charfuelan, M., Arenas, D., Vollmer, M., Dengel, A., 2025. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. Remote Sens. Environ. 318, 114547.

Moon, J., Lee, H., 2021. Analysis of activity in an open-pit mine by using InSAR coherence-based normalized difference activity index. Remote Sens. 13, 1861.

Niu, B., Quanlong, F., Jianyu, Y., Boan, C., Bingbo, G., Jiantao, L., Yi, L., Gong, J., 2023. Solid waste mapping based on very high resolution remote sensing imagery and a novel deep learning approach. Geocarto Int. 38, 2164361.

Pandey, S., van Nistelrooij, M., Maasakkers, J.D., Sutar, P., Houweling, S., Varon, D.J., Tol, P., Gains, D., Worden, J., Aben, I., 2023. Daily detection and quantification of methane leaks using Sentinel-3: A tiered satellite observation approach with Sentinel-2 and Sentinel-5p. Remote Sens. Environ. 296, 113716.

Ren, S., Hu, W., Bradbury, K., Harrison-Atlas, D., Malaguzzi Valeri, L., Murray, B., Malof, J.M., 2022. Automated extraction of energy systems information from remotely sensed data: A review and analysis. Appl. Energy 326, 119876.

Robinson, C., Chugg, B., Anderson, B., Ferres, J.M.L., Ho, D.E., 2022. Mapping industrial poultry operations at scale with deep learning and aerial imagery. IEEE J Sel. Top. Appl. Earth Observ. Remote Sens. 15, 7458–7471.

Rouet-Leduc, B., Hulbert, C., 2024. Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer. Nat. Commun. 15, 3801.

Sabbatino, M., 2018. Global oil & gas features database. National Energy Technology Laboratory (NETL), Pittsburgh, PA, Morgantown, WV.

Saunois, M., Bousquet, P., Poulter, B., Peregon, A., Ciais, P., Canadell, J.G., Dlugokencky, E.J., Etiope, G., Bastviken, D., Houweling, S., 2016. The global methane budget: 2000–2012. Earth Syst. Sci. Data Discuss. 2016, 1–79.

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., 2023. Automated detection and monitoring of methane super-emitters using satellite data. Atmos. Chem. Phys. Discuss. 23, 9071–9098.

Shen, L., Jacob, D.J., Gautam, R., Omara, M., Scarpelli, T.R., Lorente, A., Zavala-Araiza, D., Lu, X., Chen, Z., Lin, J., 2023. National quantifications of methane emissions from fuel exploitation using high resolution inversions of satellite observations. Nat. Commun. 14, 4948.

Sheng, H., Irvin, J., Munukutla, S., Zhang, S., Cross, C., Story, K., Rustowicz, R., Elsworth, C., Yang, Z., Omara, M., 2020. Ognet: Towards a global oil and gas infrastructure database using deep learning on remotely sensed imagery. arXiv preprint arXiv:2011.07227.

Sun, X., Yin, D., Qin, F., Yu, H., Lu, W., Yao, F., He, Q., Huang, X., Yan, Z., Wang, P., Deng, C., Liu, N., Yang, Y., Liang, W., Wang, R., Wang, C., Yokoya, N., Hänsch, R., Fu, K., 2023. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. Nat. Commun. 14, 1444.

Vanselow, S., Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., Boesch, H., Burrows, J.P., 2024. Automated detection of regions with persistently enhanced methane concentrations using Sentinel-5 Precursor satellite data. Atmos. Chem. Phys. 24, 10441–10473.

Vaughn, T.L., Bell, C.S., Pickering, C.K., Schwietzke, S., Heath, G.A., Pétron, G., Zimmerle, D.J., Schnell, R.C., Nummedal, D., 2018. Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region. Proc. Natl. Acad. Sci. 115, 11712–11717.

Watine-Guiu, M., Varon, D.J., Irakulis-Loitxate, I., Balasus, N., Jacob, D.J., 2023. Geostationary satellite observations of extreme and transient methane emissions from oil and gas infrastructure. Proc. Natl. Acad. Sci. 120, e2310797120.

Wei, K., Dai, J., Hong, D., Ye, Y., 2024. MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 135, 104241.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Wu, X., Hong, D., Chanussot, J., 2022. Convolutional neural networks for multimodal remote sensing data classification. IEEE Trans. Geosci. Remote Sensing 60, 1–10.

Wu, H., Liu, Y., Pu, Y., Liu, P., Zhao, W., Guo, X., 2024. National-scale nighttime high-temperature anomalies from Landsat-8 OLI images. ISPRS-J. Photogramm. Remote Sens. 212, 212–229.

Xue, Z., Marculescu, R., 2023. Dynamic Multimodal Fusion, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2575–2584.

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., Zhang, B., 2018. Multisource remote sensing data classification based on convolutional neural network. IEEE Trans. Geosci. Remote Sensing 56, 937–949.

Yan, W.Y., Mahendrarajah, P., Shaker, A., Faisal, K., Luong, R., Al-Ahmad, M., 2014. Analysis of multi-temporal landsat satellite images for monitoring land surface temperature of municipal solid waste disposal sites. Environ. Monit. Assess. 186, 8161–8173.

Zalpour, M., Gholamreza, A., Alaei-Sheini, N., 2020. A new approach for oil tank detection using deep learning features with control false alarm rate in high-resolution satellite imagery. Int. J. Remote Sens. 41, 2239–2262.

Zhang, L., Shi, Z., Wu, J., 2015. A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 8, 4895–4909.

Zhang, S., Ma, J., Zhang, X., Guo, C., 2023. Atmospheric remote sensing for anthropogenic methane emissions: Applications and research opportunities. Sci. Total Environ. 893, 164701.

Zhao, Y., Chen, Y., Xiong, S., Lu, X., Zhu, X.X., Mou, L., 2024. Co-enhanced global-part integration for remote-sensing scene classification. IEEE Trans. Geosci. Remote Sensing 62, 1–14.

Zhu, B., Lui, N., Irvin, J.A., Le, J., Tadwalkar, S., Wang, C., Ouyang, Z., Liu, F.Y., Ng, A. Y., Jackson, R.B., 2022. METER-ML: A Multi-Sensor Earth Observation Benchmark for Automated Methane Source Mapping, CDCEO@IJCAI.