METER-ML: A Multi-Sensor Earth Observation Benchmark for Automated Methane Source Mapping

Bryan Zhu[1,*,†], Nicholas Lui[2,†], Jeremy Irvin[1,†], Jimmy Le[1], Sahil Tadwalkar[3], Chenghao Wang[4], Zutao Ouyang[4], Frankie Y. Liu[4], Andrew Y. Ng[1] and Robert B. Jackson[4,5]

[1]Department of Computer Science, Stanford University

[2]Department of Statistics, Stanford University

[3]Department of Civil and Environmental Engineering, Stanford University

[4]Department of Earth System Science, Stanford University

[5]Woods Institute for the Environment and Precourt Institute for Energy, Stanford University

Abstract

Reducing methane emissions is essential for mitigating global warming. To attribute methane emissions to their sources, a comprehensive dataset of methane source infrastructure is necessary. Recent advancements with deep learning on remotely sensed imagery have the potential to identify the locations and characteristics of methane sources, but there is a substantial lack of publicly available data to enable machine learning researchers and practitioners to build automated mapping approaches. To help fill this gap, we construct a multi-sensor dataset called METER-ML containing 86,599 georeferenced NAIP, Sentinel-1, and Sentinel-2 images in the U.S. labeled for the presence or absence of methane source facilities including concentrated animal feeding operations, coal mines, landfills, natural gas processing plants, oil refineries and petroleum terminals, and wastewater treatment plants. We experiment with a variety of models that leverage different spatial resolutions, spatial footprints, image products, and spectral bands. We find that our best model achieves an area under the precision recall curve of 0.915 for identifying concentrated animal feeding operations and 0.821 for oil refineries and petroleum terminals on an expert-labeled test set, suggesting the potential for large-scale mapping. We make METER-ML freely available at this link to support future work on automated methane source mapping.

Keywords
Earth observation, remote sensing, machine learning, deep learning, dataset, climate change, methane

1. Introduction

Anthropogenic methane emissions are the main contributor to the rise of atmospheric methane [1], and mitigating methane emissions is widely recognized as crucial for slowing global warming and achieving the goals of the Paris Agreement [2]. Multiple satellites are in orbit or launching soon which will measure methane emissions from the surface using top-down approaches, but in order to attribute these emissions to specific sources on the ground, a comprehensive database of methane emitting infrastructure is necessary [3]. Although several public databases of this infrastructure exist, the data available globally is incomplete, erroneous, and unaggregated.

AI approaches on Earth observation data have the po-

Figure 1: METER-ML is a multi-sensor dataset containing 86,599 examples of NAIP aerial imagery, Sentinel-2 satellite imagery, and Sentinel-1 satellite imagery. We include 19 spectral bands across these three products, with the RGB and VH&VV bands shown here. Each example is labeled with the presence or absence of six different methane source facilities and is georeferenced. A small amount of examples are labeled to contain facilities from more than one category and 34,870 examples contain no facilities from the six categories.
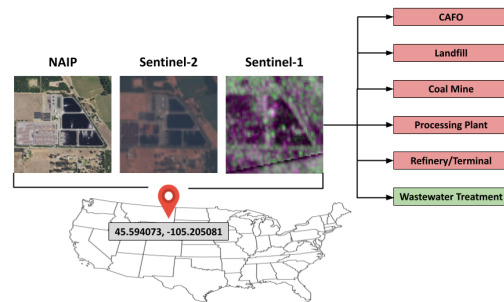
tential to fill in this gap. Several recent works have developed deep learning models to automatically interpret remotely sensed imagery and deploy them at scale to

---

map infrastructure [4, 5, 6, 7]. Methods for mapping methane source infrastructure have been emerging as well, including well pads in the Denver basin [8], oil refineries and concentrated animal feed operations in the U.S. [9, 10], and wastewater treatment plants in Germany [11]. Each of these works depended on the curation of large, labeled datasets to develop the machine learning models, but there is a lack of publicly available, labeled Earth observation data, specifically on methane emitting infrastructure, which prohibits researchers and practitioners from building automated mapping approaches.

In this work, we construct a multi-sensor Earth observation dataset for methane source infrastructure identification called METER-ML. In support of a new initiative to build a global database of methane emitting infrastructure called the **ME**thane **T**racking **E**missions **R**eference (METER) [12], we develop METER-ML to allow the machine learning community to experiment with multi-view/multi-modal modeling approaches to automatically identify this infrastructure in remotely sensed imagery. METER-ML includes georeferenced imagery from three remotely sensed image products, specifically 19 spectral bands in total from NAIP, Sentinel-1, and Sentinel-2, capturing 51,729 sources of methane from six different classes as well as 34,870 negative examples (Figure 1). The dataset includes expert-reviewed validation and test sets for robustly evaluating the performance of derived models. Using the dataset, we experiment with a variety of convolutional neural network models which leverage different spatial resolutions, spatial footprints, image products, and spectral bands. The dataset is freely available[1] in order to encourage further work on developing and validating methane source mapping approaches.

2. Methods

2.1. Methane source locations

We collect locations of methane emitting infrastructure in the U.S. from a variety of public datasets. We focus on the U.S. in this study due to the high availability of publicly accessible infrastructure data and remotely sensed imagery. The infrastructure categories we include are concentrated animal feeding operations (CAFOs), coal mines (Mines), landfills (Landfills), natural gas processing plants (Proc Plants), oil refineries and petroleum terminals (including crude oil and liquified natural gas terminals), and wastewater treatment plants (WWTPs). We group oil refineries and petroleum terminals together due to their high similarity in appearance, and refer to that category as "Refineries & Terminals" (R&Ts). These infrastructure categories were chosen based on their potential for emitting methane along with their consistent,

---

[1]https://stanfordmlgroup.github.io/projects/meter-ml

Table 1

Counts and proportions of each category in METER-ML. The labels on the training set are obtained from public data whereas the labels on validation and test sets are obtained from a consensus of two methane source identification experts. The individual category counts do not add up to the overall train/valid/test counts as some (0.8%) of the positive examples are labeled with more than one methane source category.

Category	Train (%)	Valid (%)	Test (%)	Total
CAFOs	24957 (29.3%)	47 (9.1%)	92 (9.0%)	25096
Landfills	4085 (4.8%)	46 (8.9%)	111 (10.9%)	4242
Coal Mines	1776 (2.1%)	40 (7.8%)	72 (7.1%)	1888
Proc Plants	1900 (2.2%)	38 (7.4%)	107 (10.5%)	2045
R&Ts	4012 (4.7%)	59 (11.5%)	108 (10.6%)	4179
WWTPs	14519 (17.1%)	46 (8.9%)	129 (12.7%)	14694
Negatives	34195 (40.2%)	249 (48.3%)	426 (41.8%)	34870
Total	85066	515	1018	86599

visible differentiating features which make them feasible to identify in high resolution remotely sensed imagery.

The locations are obtained from 18 different publicly available datasets, all of which have licenses that allow for redistribution (see Table 6 in the Appendix). As various datasets may contain the same locations of infrastructure, we deduplicate by considering locations within 500m of each other identical. In total we include 51,729 unique locations of methane source infrastructure in the dataset, which we refer to as positive examples.

2.2. Negative locations

We additionally include a variety of images in the dataset which capture none of the six methane emitting facilities. To do this, we define around 50 classes (see Appendix) of different facilities and landscapes and select characteristic examples of each class. Then we collect locations containing similar facilities and landscapes using the Descartes Labs GeoVisual Search [13], providing up to 1000 similar locations per example. A sample of the similar locations were manually vetted in each case to ensure no locations obtained actually corresponded to the six methane source categories. In total we include 34,870 locations of facilities and landscapes which are not any of the six infrastructure categories, and refer to these as negative examples. The counts and proportions of the positive and negative classes in the dataset are shown in Table 1.

2.3. Remotely sensed imagery

We pair all of the locations in the dataset with three publicly available remotely sensed image sources. Specifically we include aerial imagery from the USDA National Agriculture Imagery Program (NAIP) as well as satellite imagery captured by Sentinel-1 (S1) and Sentinel-2 (S2).

Table 2

Summary of the remotely sensed image products and bands included in METER-ML. RGB are the three visible bands, NIR is a single near-infrared band, RE1-4 are the four red-edge bands, SWIR1-2 are the two shortwave infrared bands, CA is the single coastal aerosol band, WV is the single water-vapor band, C is the single cirrus band, and VH & VV are the two V-transmit bands.

Product	Bands	Image Size	Resolution
NAIP	RGB & NIR	720x720	1m
Sentinel-2	RGB & NIR	72x72	10m
Sentinel-2	RE1-4 & SWIR1-2	36x36	20m
Sentinel-2	CA & WV & C	12x12	60m
Sentinel-1	VH & VV	72x72	10m

NAIP imagery covers the contiguous U.S. and S1 and S2 imagery both have global coverage. For NAIP we use 1m resolution imagery, for Sentinel-2 we use the L1C product at 10m resolution, and for Sentinel-1 we use the Sigma Nought Backscatter product at 10m resolution. We use all spectral bands from each product. Specifically, we use the three visible (RGB) and single near-infrared (NIR) bands from NAIP and S2, the single coastal aerosol (CA) band, four red-edge (RE1-4) bands, single water vapor (WV) band, single cirrus (C) band, and the two shortwave infrared (SWIR1-2) bands from S2, and the V-transmit (VH and VV) bands from S1. We include S1 and S2 in the dataset in order to enable experimenting with coarser resolution satellite imagery which is globally available, unlike NAIP. The details of each imagery product and band are shown in Table 2.

In order to construct images containing each location in the dataset, we consider a 720m x 720m footprint centered around the location. This footprint was chosen to balance the size of the images with the contextual information, but we investigate this choice in the experiments. Due to the geographic coordinate noise in the publicly available datasets, we chose to center the imagery at the locations which increases the likelihood the facilities are captured in the imagery, but still has natural variation in the locations of the facilities in the imagery. We construct a mosaic of the most recently captured pixels in a time range for each image product, where we consider NAIP images captured between 2017 and 2021 and Sentinel-1 and Sentinel-2 images between May and September 2021, where Sentinel-2 images are selected based on lowest cloud cover. We use the Descartes Labs platform to download all of the imagery [14].

The total dataset contains 86,599 images capturing ten spectral bands across the three imagery products. Information about the remotely sensed image products and bands included in the dataset are provided in Table 2 and characteristic examples for each methane source category are shown in Figure 2 in the Appendix.

2.4. Validation and test sets

Two Stanford University postdoctoral researchers with expertise in methane emissions and related infrastructure individually reviewed 1,533 examples to compose the held-out validation and test sets. To determine which examples to include in these held-out sets, we randomly sampled 150 images from each of the six positive classes as well as a random sample of 33 images which have multiple labels, constituting 933 positive examples according to the original public dataset labels. We additionally sampled 12 images from each of the 50 negative categories resulting in 600 negative examples. The experts both manually reviewed these examples and identified the presence or absence of the six methane source categories by using a combination of NAIP imagery as well as Google Maps imagery, which often had finer spatial resolution as well as place names. The facility had to be captured by the NAIP image for the corresponding label to be assigned. If the expert identified no clearly visible methane source categories in the image, the example was labeled "negative", and if the expert was uncertain about any label, the example was labeled "uncertain". The two labels per example were then resolved as follows:

1. If the experts agreed and neither was uncertain, the agreed upon label was taken as the final label.
2. If the experts disagreed, and one was uncertain but the other was not, the expert's certain label was taken as the final label.
3. If the experts disagreed, but one agreed with the original label, the original label was taken as the final label.
4. In all other scenarios, the example was reviewed jointly by the experts and a final label was assigned.

Only 76 examples out of the 1,533 went to another round of review. The resulting datasets have 858 positive examples and 675 negative examples. We split the 1,533 examples into 515 for the validation set and 1,018 for the test set. The label counts on the validation and test sets are shown in Table 1.

3. Experiments

We run a variety of multi-label classification experiments on the curated dataset. In all of our experiments, we use a DenseNet-121 convolutional neural network architecture [15]. Preliminary experiments on the dataset explored various ResNet and DenseNet architectures and found that DenseNets outperformed all ResNet variants [16]. We use a linear layer which outputs six values indicating the likelihood that each of the six methane source

Table 3
Per-class and overall (macro-average) validation AUPRC for different remotely sensed image products and bands. All of these experiments use images of size 720x720 at a spatial resolution of 1m per pixel, with S1 and S2 upsampled to that resolution.

Image Product	Bands	CAFOs	Landfills	Mines	Proc Plants	R&Ts	WWTPs	Overall
S1	VH&VV	0.519	0.107	0.152	0.218	0.487	0.119	0.267
S2	RGB	0.889	0.268	0.305	0.374	0.694	0.204	0.456
S2	All	0.889	0.189	0.382	0.368	0.690	0.183	0.450
S2 & S1	All	0.923	0.152	0.379	0.391	0.612	0.231	0.448
NAIP	RGB	0.903	0.270	0.348	0.327	0.849	0.182	0.480
NAIP	All	**0.945**	**0.276**	0.401	**0.508**	**0.857**	**0.303**	**0.548**
NAIP & S2 & S1	All	0.889	0.214	**0.473**	0.457	0.796	0.272	0.517

categories are present in the input image, which outperformed individual models across all classes in our preliminary experiments. Although the model does not explicitly produce a value indicating the likelihood that the image is negative, a low value assigned to all classes indicates a negative prediction. The loss function is the mean of six unweighted binary cross entropy losses, where the label is 1 if the class if present in the image and 0 otherwise. All six labels in the negative examples are 0. The network weights are initialized with weights from a network pretrained on ImageNet [17]. Before inputting the images into the networks, we upscale the Sentinel-1 and Sentinel-2 images to match the size of NAIP images using bilinear resampling and normalize the values by the display range of the bands (see Table 7 in the Appendix). When using inputs with less than or more than 3 channels, we replace the first convolutional neural network layer with one which accepts the corresponding number of channels. Each model is trained for 5 epochs with a batch size of 4. For each model we use the checkpoint saved after an epoch which led to the lowest validation loss. We use an Adam optimizer with standard parameters [18] and a learning rate of 0.02. All models are trained using a GeForce GTX 1070 GPU.

The baseline setting for all experiments uses images capturing a footprint of 720m x 720m with 1m spatial resolution (720 x 720 image dimensions). After the models are trained, each of the six values output by the model are fed through an element-wise sigmoid function to produce a probability for each of the six categories. To evaluate the performance of the models, we compute the per-class area under the precision recall curve (AUPRC) and summarize the performance over all classes by taking the macro-average of the per-class AUPRCs.

3.1. Impact of using different imaging products and bands

We investigate the impact of using different combinations of image products and bands in the dataset (Table 3). Specifically, we experiment with NAIP, S2, and S1 alone, only visible bands and all spectral bands for S2 and NAIP,

all spectral bands from S1 and S2 together (representing the model closest to public global transferability due to the global coverage of S1 and S2), and all spectral bands from the three products together.

The best model according to macro-average AUPRC is the one which uses NAIP with all bands (the three visible bands and NIR band), achieving an overall AUPRC of 0.548 and the highest performance on CAFOs, Landfills, Proc Plants, R&Ts, and WWTPs compared to all other tested product and band combinations. Notably, it achieves very high performance on CAFOs (AUPRC=0.945) and high performance on R&Ts (AUPRC=0.857). The second best model is the joint NAIP+S2+S1 model, achieving an overall AUPRC of 0.517 and the highest performance on Mines (AUPRC=0.473) compared to all other tested product and band combinations.

S1 alone underperforms all other combinations of products and bands, followed by S2 and S1 jointly, which performed similarly overall to S2 with only the visible bands and all spectral bands. Importantly, the S2 and S1 joint model still achieves high performance on CAFOs (AUPRC=0.923), although the performance is lower than performance on CAFOs using NAIP imagery (AUPRC=0.947). There is a significant drop in performance on all classes when moving from NAIP to S2, highlighting the benefit of using high spatial resolution imagery.

The inclusion of the non-visible information substantially improves overall AUPRC for NAIP (AUPRC=0.480 → 0.548) but not for Sentinel 2 (AUPRC=0.450 → 0.448). For NAIP, the improvement is observed for all classes, with substantial gains on CAFOs, Mines, Proc Plants, and WWTPs. For Sentinel 2, the inclusion of non-visible bands substantially improves performance on Mines but substantially degrades performance on Landfills. For both products, minimal change on R&Ts performance is observed when including the non-visible bands.

Table 4
Per-class and overall (macro-average) validation AUPRC at varying image footprints and spatial resolutions.

Image Footprint	Resolution	CAFOs	Landfills	Mines	Proc Plants	R&Ts	WWTPs	Overall
240x240	1m	0.773	0.217	0.407	0.438	0.735	0.337	0.485
480x480	1m	0.772	0.226	0.260	0.371	0.855	**0.506**	0.498
720x720	3m	0.891	0.245	0.378	**0.566**	0.837	0.269	0.531
720x720	1.5m	0.927	0.244	**0.426**	0.366	0.831	0.449	0.541
720x720	1m	**0.945**	**0.276**	0.401	0.508	**0.857**	0.303	**0.548**

Table 5
Per-class and overall (macros-average) test metrics of the per-class expert model. The per-class expert model consists of one model per class, where the model used for each class is selected based on the highest performing settings for that class across the product, bands, footprint, and resolution experiments.

Category	AUPRC	AUROCC	Precision	Recall	F1
CAFOs	0.915	0.989	0.822	0.902	0.860
Landfills	0.259	0.754	0.246	0.523	0.334
Mines	0.470	0.905	0.558	0.403	0.468
Proc Plants	0.350	0.787	0.336	0.477	0.394
R&Ts	0.821	0.956	0.752	0.787	0.769
WWTPs	0.534	0.836	0.633	0.477	0.544
Overall	0.558	0.871	0.558	0.595	0.562

3.2. Impact of image footprint and spatial resolution

As image footprint (i.e. the amount of area on the ground captured by the image) and spatial resolution likely impact model performance due to the variation in the sizes of the methane-emitting facilities and equipment, we conduct experiments to test these effects (Table 4). To investigate the impact of footprint, we center crop the 720 x 720 1m images to obtain 480 x 480 and 240 x 240 1m images corresponding to 480m x 480m and 240m x 240m footprints respectively. Note that this reduces the area on the ground with spatial resolution held constant. To investigate the impact of spatial resolution, we use cubic spline interpolation [19] to downsample the 720 x 720 images to 480 x 480 (1.5m resolution, corresponding to Airbus SPOT imagery) and 240 x 240 (3m resolution, corresponding to PlanetScope imagery). Note that this reduces the spatial resolution without modifying the image footprint. In all experiments, we up-sample the images back to 720 x 720 to avoid any differences in performance due to varying image size. We use NAIP with RGB + NIR bands for these experiments as this setting produced the best overall performance compared to the other combinations of products and bands.

We find that the largest tested image footprint achieves the highest overall performance (0.548) and substantially outperforms both smaller spatial footprints across all classes except for WWTPs. This may be explained by the fact that a significant number of smaller wastewater treatment plants are surrounded by industrial buildings and other infrastructure, so cropping out this infrastructure improves the model's ability to identify the salient features of the wastewater treatment facilities.

We further find that the highest spatial resolution achieves the best overall performance (AUPRC=0.548), outperforming the coarser resolution models on CAFOs, Landfills, and R&Ts. The 1.5m resolution model closely follows with an overall AUPRC of 0.541 and outperforms the 1m resolution model on Mines. The 3m resolution model also closely follows the 1.5m resolution model achieving an overall performance of 0.531, and substantially outperforms both the higher resolution models on Proc Plants. This result suggests that models developed at 1.5m and even 3m resolution have the potential to perform almost as well as 1m resolution models, which has implications on global applicability as Airbus SPOT and PlanetScope are globally (privately) available at 1.5m and 3m resolution respectively.

3.3. Per-class expert model test set results

For each methane source category, we select the experimental configuration (product/band/footprint/resolution) that achieved the highest validation AUPRC for that class to serve as the "class expert". We refer to the combination of the different class experts as the per-class expert model.

We evaluate the per-class expert model on the hold-out test set using a variety of metrics including AUPRC and area under the receiver operating characteristic curve (AUROCC) as well as precision, recall, and F1 at the threshold which achieves the highest F1 on the validation set. The results are shown in Table 5. The per-class expert model obtains a macro-average AUPRC of 0.558. The model does especially well on CAFOs (AUPRC=0.915) and R&Ts (AUPRC=0.821), possibly because these sources have very distinctive features (e.g., long barns in CAFOs and storage tank farms in R&Ts). It performs more poorly on the other sources, especially landfills which do not have many clear distinctive features visible at 1m resolution. Notably it achieves the lowest performance on the categories with the least number of examples in the dataset, excluding R&Ts which may be simpler to identify due to their homogeneity and discernible features.

4. Discussion

The experiments suggest that the choice of imaging product, spectral band, image footprint, and spatial resolution can lead to substantial differences in model performance, with the effect often depending on the methane source category. In particular, this suggests that there is significant room to explore approaches which leverage the multi-sensor and multi-spectral aspects of METER-ML. For example, the NAIP & S2 & S1 model underperformed the model which used NAIP alone, and using all 13 spectral bands in the S2 model did not lead to substantial performance differences compared to the S2 model which only used the three visible bands. We also do not leverage the geographic information explicitly in the models, but this has been shown to improve performance on other Earth observation tasks [20, 21]. Furthermore, there is potential to augment the dataset with other sources of imagery and information available at the provided geographic locations. We hope to help create new versions of METER-ML which may include other sources of input data and methane emitting infrastructure categories.

The best model from our experiments achieves high performance on identifying CAFOs and R&Ts, suggesting the potential to map these facilities with NAIP imagery in the U.S. which aligns with findings from prior studies [9, 10]. The performance for identifying CAFOs remains high when using S1 and S2, which are globally and publicly available. This suggests the potential to use these lower spatial resolution imagery sources to map CAFOs in other countries besides the U.S., but future work should investigate whether these findings generalize to other regions. There is still a large gap to achieving high performance for each of the other methane source categories and further improve performance on the high performing categories, so METER-ML is a challenging benchmark to test new infrastructure identification approaches.

There are many other publicly available remote sensing datasets for classification, with some of the most common being UC Merced [22], SAT-4 and SAT-6 [23], AID [24], NWPU-RESISC45 [25], EuroSAT [26], and BigEarth-Net [27]. Few of these datasets have georeferenced multi-sensor images, which limits their utility for new modeling approaches and downstream use. The OGNet dataset [9] is the most similar publicly available dataset to METER-ML and is essentially a subset of it, containing NAIP imagery of refineries in the contiguous U.S.

We identify four limitations of this work. First, we limit the geographic scope of METER-ML to the U.S. due to the availability of disseminatable infrastructure data and publicly available, high resolution imagery. Future work should include data in other regions worldwide. Second, we do not include longitudinal imagery in the dataset to reduce the size and complexity of the dataset as most infrastructure is static over time. However, longitudinal information has the potential to provide additional signal to help differentiate certain facilities, e.g. waste pile evolution at landfills. Third, we use a DenseNet121 model that is pre-trained on ImageNet, but the shape and number of channels of remote sensing imagery can be significantly different from ImageNet. It would be worthwhile to train a network from scratch on METER-ML, and compare its performance against a network that is pre-trained on ImageNet and fine-tuned on METER-ML. Fourth, our approach to combine the multi-sensor data may not be optimal as the products and spectral bands have different spatial resolutions and sensor types (e.g. active vs. passive sensors). One alternate approach may be to dedicate different network branches for the inputs and combine the representations from each branch.

5. Conclusion

In this work, we curate a large georeferenced multi-sensor dataset called METER-ML to test automated methane source identification approaches. We conduct a variety of experiments investigating the impact of remotely sensed image product, spectral bands, image footprint, and spatial resolution on model performance measured against a consensus of expert labels. We find that a model which leverages NAIP with all four bands achieves the highest overall performance across the tested image product and spectral band combinations, followed closely by a joint NAIP, Sentinel-2, and Sentinel-1 model. We also find that the highest spatial resolution and footprint leads to the best overall performance, although performance can depend on the methane source category. Finally we show that the best model achieves high performance in identifying concentrated animal feeding operations and oil refineries and petroleum terminals, suggesting the potential to map them at scale, but substantially lower performance on the other four categories with notably lower performance identifying processing plants and landfills. We make METER-ML freely available in order to encourage and support future work on developing Earth observation models for mitigating climate change.

Acknowledgments

References

[1] Z. Zhang, B. Poulter, S. Knox, A. Stavert, G. Mc-Nicol, E. Fluet-Chouinard, A. Feinberg, Y. Zhao, P. Bousquet, J. G. Canadell, et al., Anthropogenic emission is the main contributor to the rise of atmospheric methane during 1993–2017, National science review 9 (2022) nwab200.

[2] P. Agreement, Paris agreement, in: Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change (21st Session, 2015: Paris). Retrived December, volume 4, HeinOnline, 2015, p. 2017.

[3] D. J. Jacob, D. J. Varon, D. H. Cusworth, P. E. Dennison, C. Frankenberg, R. Gautam, L. Guanter, J. Kelley, J. McKeever, L. E. Ott, et al., Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane, Atmospheric Chemistry and Physics Discussions (2022) 1–44.

[4] J. Yu, Z. Wang, A. Majumdar, R. Rajagopal, Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states, Joule 2 (2018) 2605–2617.

[5] J. Lee, N. R. Brooks, F. Tajwar, M. Burke, S. Ermon, D. B. Lobell, D. Biswas, S. P. Luby, Scalable deep learning to identify brick kilns and aid regulatory capacity, Proceedings of the National Academy of Sciences 118 (2021).

[6] L. Kruitwagen, K. Story, J. Friedrich, L. Byers, S. Skillman, C. Hepburn, A global inventory of photovoltaic solar energy generating units, Nature 598 (2021) 604–610.

[7] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, J. Quinn, Continental-scale building detection from high resolution satellite imagery, arXiv preprint arXiv:2107.12283 (2021).

[8] S. Dileep, D. Zimmerle, J. R. Beveridge, T. Vaughn, Automated identification of oil field features using cnns, in: NeurIPS, Workshop on Tackling Climate Change with Machine Learning, 2020, 2020.

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

[10] C. Handan-Nader, D. E. Ho, Deep learning to map concentrated animal feeding operations, Nature Sustainability 2 (2019) 298–306.

[11] H. Li, J. Zech, D. Hong, P. Ghamisi, M. Schultz, A. Zipf, Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection, International Journal of Applied Earth Observation and Geoinformation 110 (2022) 102804.

[12] Stanford University, Methane tracking emissions reference platform, 2022. URL: https://meterplatform.web.app/, accessed: 2022-06-30.

[13] R. Keisler, S. W. Skillman, S. Gonnabathula, J. Poehnelt, X. Rudelis, M. S. Warren, Visual search over billions of aerial and satellite images, Computer Vision and Image Understanding 187 (2019) 102790.

[14] D. Labs, Descartes labs platform, 2022. URL: https://descarteslabs.com/, accessed: 2022-06-26.

[15] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[19] P. S. Parsania, P. V. Virparia, A comparative analysis of image interpolation algorithms, International Journal of Advanced Research in Computer and Communication Engineering 5 (2016) 29–34.

[20] O. Mac Aodha, E. Cole, P. Perona, Presence-only geographical priors for fine-grained image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9596–9606.

[21] J. Irvin, H. Sheng, N. Ramachandran, S. Johnson-Yu, S. Zhou, K. Story, R. Rustowicz, C. Elsworth, K. Austin, A. Y. Ng, Forestnet: Classifying drivers of deforestation in indonesia using deep learning on satellite imagery, arXiv preprint arXiv:2011.05479 (2020).

[22] Y. Yang, S. Newsam, Geographic image retrieval using local invariant features, IEEE Transactions on Geoscience and Remote Sensing 51 (2012) 818–832.

[23] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, R. Nemani, Deepsat: a learning framework for satellite imagery, in: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, 2015, pp. 1–10.

[24] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, Aid: A benchmark data set for performance evaluation of aerial scene classifica-

tion, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 3965–3981.

[25] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proceedings of the IEEE 105 (2017) 1865–1883.

[26] P. Helber, B. Bischke, A. Dengel, D. Borth, Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12 (2019) 2217–2226.

[27] G. Sumbul, M. Charfuelan, B. Demir, V. Markl, Bigearthnet: A large-scale benchmark archive for remote sensing image understanding, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2019, pp. 5901–5904.

[28] CA Governor's Office of Emergency Services, Ca energy commission - oil refineries and terminals, 2021. URL: https://arcg.is/1rXn4G0, accessed: 2022-02-01.

[29] CA State Water Resources Control Board, 2022. URL: https://www.waterboards.ca.gov/resources/data_databases/site_map.html, accessed: 2022-02-01.

[30] D. McKernan, 2012 factory farms in michigan, 2012. URL: https://data.world/dmckernan/2012-factory-farms-in-michigan, accessed: 2022-02-01.

[31] O. Tsubiks, Concentrated animal feeding operations (cafo), us, 2017. URL: https://data.world/dataforacause/concentrated-animal-feeding-operations-cafo, accessed: 2022-02-01.

[32] U.S. Energy Information Administration, U.s. energy mapping system, 2021. URL: https://www.eia.gov/state/maps.php, accessed: 2022-02-01.

[33] U.S. Environmental Protection Agency, Greenhouse gas reporting program (ghgrp), 2021. URL: https://www.epa.gov/ghgreporting, accessed: 2022-02-01.

[34] K. Rose, J. Bauer, V. Baker, A. Barkhurst, A. Bean, J. DiGiulio, K. Jones, T. Jones, D. Justman, R. Miller, et al., Global Oil & Gas Features Database, Technical Report, 2018.

[35] U.S. D of Homeland Security, Homeland infrastructure foundation-level data (hifld), 2021. URL: https://hifld-geoplatform.opendata.arcgis.com/, accessed: 2022-02-01.

[36] H. Ehalt Macedo, B. Lehner, J. Nicell, G. Grill, J. Li, A. Limtong, R. Shakya, Distribution and characteristics of wastewater treatment plants within the global river network, Earth System Science Data 14 (2022) 559–577.

[37] Indiana D of Environmental Management, Confined feeding operation facilities: Indiana,

2020. URL: https://maps.indiana.edu/metadata/Environment/Agribusiness_Confined_Feeding_Operations.html#Identification_Information, accessed: 2022-02-01.

[38] U.S. Environmental Protection Agency, Landfill methane outreach program, 2022. URL: https://www.epa.gov/lmop/project-and-landfill-data-state, accessed: 2022-02-01.

[39] A. J. Marchese, T. L. Vaughn, D. J. Zimmerle, D. M. Martinez, L. L. Williams, A. L. Robinson, A. L. Mitchell, R. Subramanian, D. S. Tkacik, J. R. Roscioli, et al., Methane emissions from united states natural gas gathering and processing, Environmental science & technology 49 (2015) 10718–10727.

[40] V. Maus, S. Giljum, D. M. da Silva, J. Gutschlhofer, R. P. da Rosa, S. Luckeneder, S. L. Gass, M. Lieber, I. McCallum, An update on global mining land use, Scientific data (2022).

[41] Metropolitan Council, Wastewater treatment plants, 2020. URL: https://gisdata.mn.gov/dataset/us-mn-state-metc-util-wastewater-treatment-plants, accessed: 2022-02-01.

[42] MN Pollution Control Agency, Feedlots in minnesota, 2016. URL: https://gisdata.mn.gov/dataset/env-feedlots, accessed: 2022-02-01.

[43] ORNL DAAC, Sources of methane emissions (vista-ca), state of california, usa, 2019. URL: https://daac.ornl.gov/NACP/guides/NACP_Vista_CA_CH4_Inventory.html, accessed: 2022-02-01.

[44] Sierra Club Michigan Chapter, Michigan cafo mapping report, 2017. URL: https://www.sierraclub.org/michigan/michigan-cafo-mapping-report, accessed: 2022-02-01.

[45] Stanford Regulation, Evaluation, and Governance Lab, Cafo training dataset, 2019. URL: https://reglab.stanford.edu/data/cafo-training-dataset/, accessed: 2022-02-01.

A. Appendix

A.1. Methane-Emitting Infrastructure Datasets

Table 6
Summary of the datasets containing the locations of methane-emitting infrastructure that are included in METER-ML. All datasets are subsetted to the locations within the contiguous U.S. We use the centroids of polygons for any datasets provided in polygon format.

Dataset Source	Scope	Methane Source Categories
CA Energy Commission [28]	California	R&Ts
CSWRCB [29]	California	WWTPs
data.world [30]	Michigan	CAFOs
Data For Cause Challenge [31]	US	CAFOs
EIA [32]	US	Proc Plants, R&Ts, WWTPs
GHGRP [33]	US	Landfills, Mines, R&Ts, WWTPs
GOGI [34]	Global	Proc Plants, R&Ts
HIFLD [35]	US	R&Ts, WWTPs
HydroWASTE [36]	Global	WWTPs
IndianaMap [37]	Indiana	CAFOs
LMOP [38]	US	Landfills
Marchese et al. [39]	US	Proc Plants
Maus et al. [40]	Global	Mines
Minnesota Metropolitan Council [41]	Minnesota	WWTPs
Minnesota Pollution Control Agency [42]	Minnesota	CAFOs
ORNL DAAC [43]	California	CAFOs
Sierra Club [44]	Michigan	CAFOs
Stanford RegLab [45]	North Carolina	CAFOs

A.1.1. Coal Mines Data

The mines data from [40] were subsetted to coal mines in order to capture the mines responsible for the vast majority of methane emissions. To do this, the polygons and coal mine coordinates obtained from S&P Global Commodity Insights were matched to determine which polygons were spatially related to coal mine coordinates. Then a visual check and hand cleaning was performed on the polygons assigned a coal mining label to ensure correctness.

A.2. Negative Classes

We identify a variety of infrastructure which are not any of the six infrastructure categories to use as negatives in the dataset. Specifically we include football fields, marinas, solar panels, large bodies of water, parking lots, windmills, baseball fields, airport runways, clouds, neighborhoods, golf courses, roundabouts, mountainous terrain, trees, boats, islands, rocks, rivers, roads, bridges, ripples in water, snow, canyon formations, sparse forests, suburban neighborhoods, beaches, clear water, swimming pools, sand, corn farms, soy farms, trees on mountainside, farm houses, grass, airplanes, turning roads, intersections, multifamily residential facilities, rapids, docks, highway loops, mowed grass, container yards, soccer fields, greenhouses, crops, personal watercrafts, pivot irrigation systems, and concrete plants. Characteristic examples of each type were selected and a variety of similar examples per type were obtained using the Descartes Labs GeoVisual Similarity tool [13].
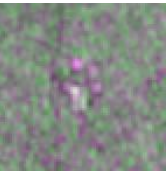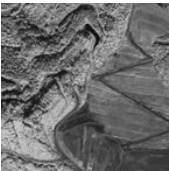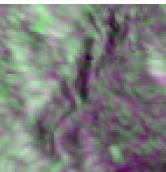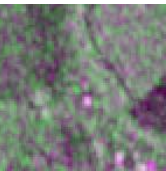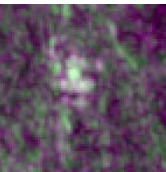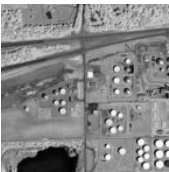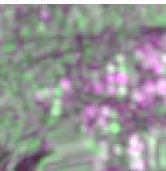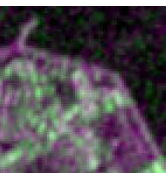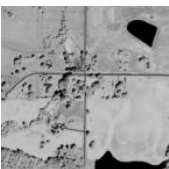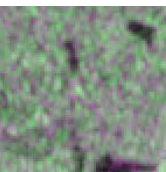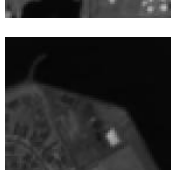
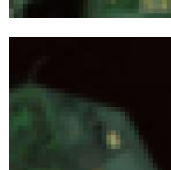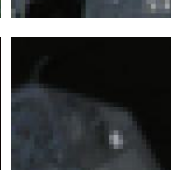## A.3. Remotely Sensed Image Statistics and Examples
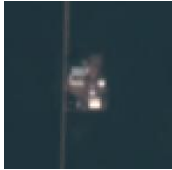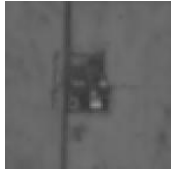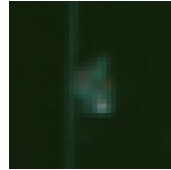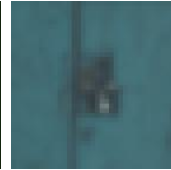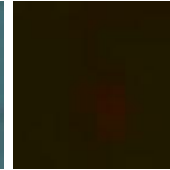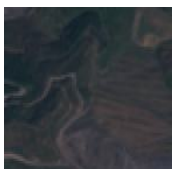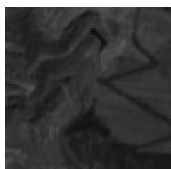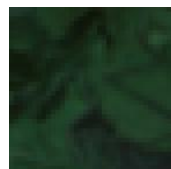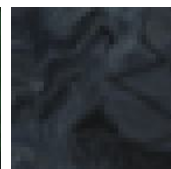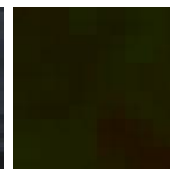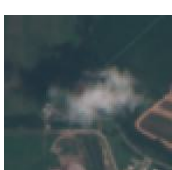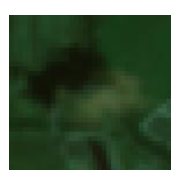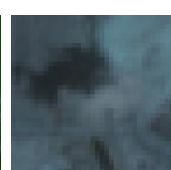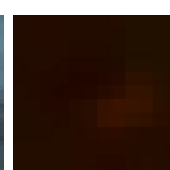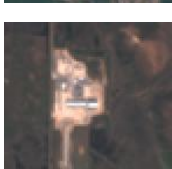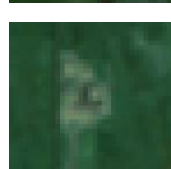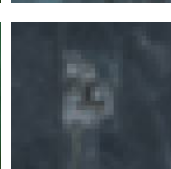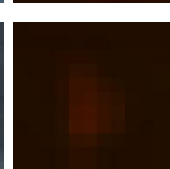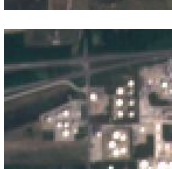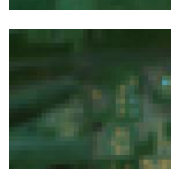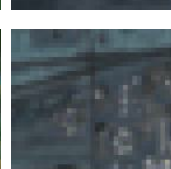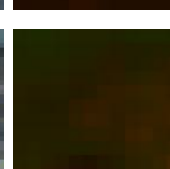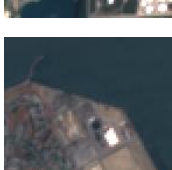
Table 7
Summary of the remotely sensed image products and bands included in METER-ML. The raw image data contains values from the data range, while the display range is used to normalize values before displaying imagery or inputting into models.

Product/Bands	Image Size	Resolution	Data Range	Display Range
NAIP RGB & NIR	720x720	1m	[0,255]	[0,255]
Sentinel-2 CA	12x12	60m	[0,10000]	[0,10000]
Sentinel-2 RGB	72x72	10m	[0,10000]	[0,4000]
Sentinel-2 RE1-4	36x36	20m	[0,10000]	[0,10000]
Sentinel-2 NIR	72x72	10m	[0,10000]	[0,10000]
Sentinel-2 WV	12x12	60m	[0,10000]	[0,10000]
Sentinel-2 C	12x12	60m	[0,10000]	[0,10000]
Sentinel-2 SWIR1-2	36x36	20m	[0,10000]	[0,10000]
Sentinel-1 VH	72x72	10m	[1,4095]	[585,2100]
Sentinel-1 VV	72x72	10m	[1,4095]	[585,2926]

Figure 2: Characteristic example images of each category in METER-ML. METER-ML contains 19 spectral bands across 3 image products, visualized here in grayscale for single bands and false color composites for multiple bands (the first listed band is used in the red channel, second in the green, third in the blue).

Category	S2 RGB	S2 NIR	S2 RE1&SWIR1-2	S2 RE2-4	S2 CA&WV&C
CAFOs					
Coal Mines					
Landfills					
Proc Plants					
R&Ts					
WWTPs					
Negatives					