

Physics-Guided Multitask Learning for Estimating Power Generation and CO₂ Emissions From Satellite Imagery

Joëlle Hanna¹, Graduate Student Member, IEEE, Damian Borth, Member, IEEE, and Michael Mommert¹, Member, IEEE

Abstract—Fossil fuel combustion produces large quantities of carbon dioxide (CO₂), a major greenhouse gas (GHG), which is one of the main drivers of climate change. A quantitative assessment of GHG emissions is fundamental to predicting climate change effects, enforcing emission regulations, and monitoring pollution trading schemes. Unfortunately, the reporting of GHG emissions is only required in some countries, resulting in insufficient global coverage. At the same time, the transition from fossil fuels to zero carbon to limit climate change is at the heart of several ecological movements, hence the need for quantifying energy production, as well. In this work, we propose an end-to-end method to estimate power generation rates for fossil fuel power plants from satellite images, based on which we approximate GHG (CO₂) emission rates. We present a physics-guided multitask deep-learning approach able to simultaneously predict from a single-satellite image of a power plant: 1) the pixel-area covered by plumes; 2) the type of fired fuel; and 3) the power generation rate. To ensure physically realistic predictions from our model we account for environmental conditions and empirical physical constraints. We then convert the predicted power generation rate into estimates for the rate at which CO₂ is being emitted, using a fuel-dependent conversion factor. Experimental results show that our multitask learning approach improves the power generation estimation mean absolute error (MAE) by 23% compared to a single-task network trained on the same dataset.

Index Terms—CO₂ estimation, physics-guided deep learning, power generation, remote sensing.

I. INTRODUCTION

CLIMATE change is affecting the entire world, leading to polar ice shields melting and sea levels rising, as well as extreme weather events. Some regions are more prone to increased rainfall, while others are facing severe heat waves and droughts. These climate disorders are primarily caused by the excessive release of greenhouse gases (GHG), trapping heat in the atmosphere and the oceans. Carbon dioxide (CO₂), one of the long-lived GHG released by the combustion of fossil fuels, plays the most significant role in climate change [1].

One solution to reduce—or at least control—GHG emissions would be to put in place a system of environmental policies such as taxing pollution or using an emission permit market, which would force certain companies to reduce their emissions or otherwise buy permits from other companies in

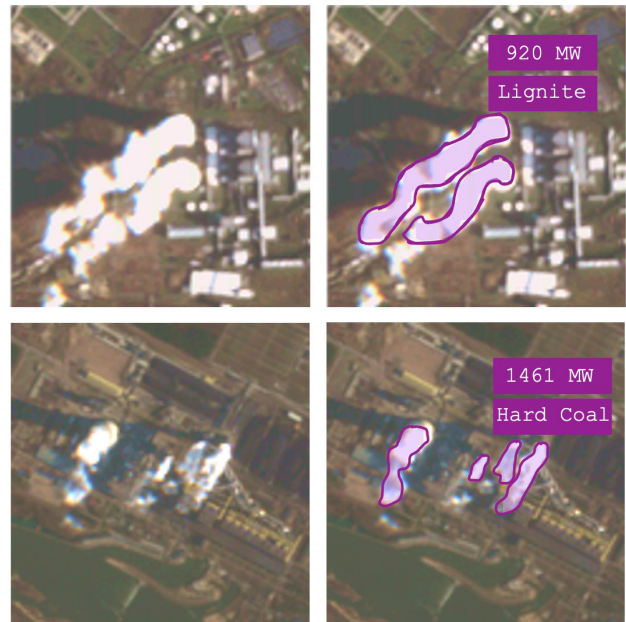


Fig. 1. Two examples of RGB satellite images (left) with their corresponding predictions by (right) our multitask approach. Our model predicts plume segmentation maps, type of fired fuel, and power generation rate (in MW). Using these predictions, we can estimate the amount of emitted CO₂.

the sector. Putting this solution in place on a global level would require precise quantification of emissions, which is not scalable as dedicated CO₂ measuring devices are not always available, nor in place.

At the same time, the objective monitoring of the planned out-phasing of fossil fuel power plants requires independent methods to quantify the generation of power from such plants.

To address both blind spots, we propose a method to estimate power generation and CO₂ emissions from fossil fuel power plants using observations of plumes acquired by Earth observation (EO) satellites. We focus on power plants given that electricity generation is one of the biggest sources of GHG emissions in Europe [2] and across the world. Moreover, quantifying fossil fuel-based power production is an equally important aspect for climate change monitoring, and more specifically for the energy transition, since the latter aims to minimize the environmental impact of the energy sector and promotes a more sustainable and carbon-free model. Therefore, we propose to predict actual power generation from fossil-fuel power plants using data from the European Network of Transmission System Operators for Electricity

Manuscript received 28 June 2022; revised 2 February 2023 and 18 April 2023; accepted 17 May 2023. Date of publication 15 June 2023; date of current version 29 June 2023. (Corresponding author: Joëlle Hanna.)

The authors are with the AIML Laboratory, School of Computer Science, University of St. Gallen, CH-9000 St. Gallen, Switzerland (e-mail: joelle.hanna@unisg.ch).

Digital Object Identifier 10.1109/TGRS.2023.3286444

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See <https://www.ieee.org/publications/rights/index.html> for more information.

(ENTSOE) [3]. We base our method on the assumption that GHG emission rates are correlated to power generation rates and as such can be related to the projected spatial extent of plumes emerging from smokestacks or cooling towers of power plants. Our definition of plumes includes plumes of steam released from cooling towers, and, to a lesser degree, plumes of smoke released from smokestacks. We use imaging data obtained by the Sentinel-2 [4] EO satellites for this problem, since they have good spatiotemporal coverage, which allows us to retrieve multiple observations from different power plants in Europe, at various points in time.

In this work, we propose a novel multitask deep-learning approach to predict the rate of power generation (treated as a regression problem), the type of fired fuel (treated as a classification problem), as well as the plume footprint (treated as a segmentation problem), all from a single-satellite image centered on a power plant (see Fig. 1). Our approach leverages various relationships between the different tasks, such as larger smoke plumes or certain fuel types being linked to higher power outputs. Using the predicted power generation rate and the type of fired fuel, we estimate the amount of CO₂ emitted at a given time. This further highlights the benefits of our multitask approach, as it is more efficient to have one model produce all the required variables (type of fuel and power production) for our target task, rather than having to train and utilize multiple models. In addition, to account for meteorological and physical phenomena that may affect the visibility and extent of plumes, and consequently affect our predictions, we incorporate external environmental data into our model, leading to physically realistic values. Through the use of a multitask approach with a shared backbone, we were able to utilize *labeled data*, which can be difficult to obtain, more efficiently, reducing the possibility of overfitting [5]. Code¹ and dataset² utilized in this work are publicly available.

Both the predicted power generation rates and CO₂ emission rates in this work are of interest to various stakeholders and inform efforts to minimize the causes of climate change. Furthermore, the free availability of EO data enables the application of such an approach on a global scale and specifically in areas where no emission estimates are available.

The main contributions of this article can be summarized as follows.

- 1) We compile a dataset of satellite images of active power plants in Europe along with hand-labeled plume segmentation masks, corresponding actual power generation rate, type of fired fuel, and environmental conditions, at a given time.
- 2) We propose a novel multitask learning approach able to simultaneously segment plumes, predict the type of fired fuel, as well as the power generation rate—significantly outperforming single-task approaches (baselines) for all these tasks—guided by physical considerations to ensure realistic predictions.
- 3) We propose a novel method to estimate CO₂ emission rates using the predicted power generation rates and derived fuel-dependent emission factors.

¹The code base for this work is available at github.com/HSGAIML/RemoteSensingCO2Estimation

²The complete dataset is available at <https://doi.org/10.5281/zenodo.5644746>

The article is organized as follows. Section II introduces related work. Section III describes the dataset compiled for this work. Section IV details the proposed approach. Sections V and VI present results and discussions. And, finally, Section VII summarizes our findings.

II. RELATED WORK

A. Estimating Power Generation and GHG Emissions

There have been different approaches to estimating GHG emissions for power plants. One can distinguish two types of approaches: observation data-based (using satellite data) and statistical data-based. The latter has mainly been restricted to analyzing annual data. As an example, Yin et al. [7] estimate the deviation in power generation of each plant from the average generation of comparable plants based on detailed information on plant-level (such as its size, fuel type, or installed capacity) and environmental factors using machine-learning algorithms such as gradient boosting trees. However, they note that their method can be highly inaccurate for small power plants. Regarding observation data-based methods, some works first predict an observable value which, in a second step, can be used as a proxy for estimating GHGs. For example, Liu et al. [8] infer the amount of nitrogen oxides (NO_x) emitted from a small number of U.S. power plants using the ozone-monitoring instrument (OMI), a nadir-viewing visual and ultraviolet spectrometer, and use this value as a proxy to calculate the amount of CO₂ emissions using a predetermined relationship between NO_x and CO₂ and measurements provided by the continuous emissions monitoring system (CEMS). They find that their predictions are on average relatively close to the real measurements, but with a high standard deviation. Hu and Shi [9] estimate CO₂ emissions from two individual coal-fired power plants in China, using the XCO₂ values of the Orbiting Carbon Observatory-2 satellite (OCO-2), as well as an enhanced Gaussian plume model to simulate pollutant movement and spread. In their work, they showed that previous emission inventories might have overestimated the CO₂ emissions of power plants. Gray et al. [10] estimate the utilization of fossil fuel power plants by identifying plumes in satellite image data and then counting the number of hours it is actually generating power over a year. Finally, they compute the capacity factor by dividing the annual generation by an empirical relationship between the annual mean plume area and the mean hourly generation rate. Couture et al. [11] show that it is possible to achieve high-quality estimates of a power plant's operational state, that is, whether it is on or off, using satellite images, which would in future work be helpful to estimate total emissions for a plant over some period of time. Moreover, Scheibenreif et al. [12] can estimate surface-level NO₂ pollution with a high spatial resolution for any location on earth, using satellite data.

In contrast to previous works, our approach allows us to estimate CO₂ emissions using predicted power generation rates as a proxy, based on freely and globally available satellite imagery for hundreds of power stations in Europe.

B. Multitask Learning

Multitask architectures have proven successful in a number of remote-sensing applications. For instance, Wang et al. [13]

perform three tasks on remote-sensing imagery within a unified model: semantic segmentation, height estimation, and boundary detection (treated as an auxiliary task). They use a shared encoder, with task-specific heads and cross-task interactions to push the network into learning the high-frequency details. Their model improves the performance of all tasks and outperforms state-of-the-art approaches. Another example is [14], in which a multitask learning-based object detector is proposed to distinguish ships in synthetic aperture radar (SAR) images. This model is thus able to learn more discriminating object-specific features without the additional cost of manual labeling. Moreover, Bischke et al. [15] propose a multitask cascade architecture for segmenting buildings and predicting pixel distance to boundaries. They also introduce an uncertainty-weighted multitask loss, as suggested in [16], and show that their method outperforms state-of-the-art by a good margin.

Motivated by Mommert et al. [17], who successfully segment plumes over industrial sites (including power plants) and distinguish them from natural clouds and Mommert et al. [18], who show that classifying different types of power plants is possible with high confidence, we propose a multitask model which uses a shared Unet-like backbone with three task-specific heads: plume segmentation, fuel type classification, and power generation estimation. It takes as input satellite images with concurrent environmental data. Inspired by the dynamic task prioritization work of Guo et al. [19], we use a multitask loss-weighting strategy based on task performances during the training process.

C. Science-Guided Neural Networks

Incorporating scientific knowledge into models to guide the learning process has become increasingly popular, especially in remote-sensing applications, arising from the fact that models often yield physically implausible results. Guidance can take the form of constraints in the loss function, or as dedicated encoding layers in the network [20]. As an example, Seo et al. [21] propose to use a separate rule encoder in the model, with controllable rule strength. Their approach is general and can be used to encode any rules, such as physics or economics ones. Moreover, Karpatne et al. [22] formalize the general concept of theory-guided learning and describe several approaches for integrating domain knowledge for different research directions. Furthermore, Karpatne et al. [23] use a physics-based loss function in the learning objective for lake temperature modeling, constraining it to follow the physical relationships. Additionally, De Bézenac et al. [24] show how incorporating general background knowledge gained from physics can help design efficient deep-learning models, providing experiments and comparisons to traditional approaches. Finally, Shen and Zhang [25] discuss how combining traditional physics-based models with machine-learning models can improve the accuracy of predicting Earth surface systems and provide a framework for combining these two approaches in various ways to create more effective models.

Following these works, we propose to incorporate empirical physical constraints related to plume formation into our objective function, constraining the model to output physically plausible estimates.

III. DATASETS

In this section, we first describe the three datasets used in this work (Sections III-A–III-C), followed by an explanation of the preprocessing steps.

A. Satellite Images

We use remote imaging data of fossil fuel power plants taken by ESA's Sentinel-2 Earth-observing satellites [4]. They consist of two polar-orbiting satellites and provide multi-spectral data (13 bands) in the visible, near-infrared, and short-wave infrared parts of the spectrum, with a pixel resolution of up to 10 m. Together, they achieve a revisit rate of five days at the equator, if weather permits. Using Level-2A data, which includes atmospheric corrections, we interpolate images to the same spatial resolution (10 m) for all bands; then, we crop 120×120 pixels regions centered on the geographic coordinates of the power plants. Extending the dataset from [17], we acquire geographic coordinates of 300 European fossil fuel power plants from the *Joint Research Centre (JRC) Open Power Plants Database* [26]. For each site, we retrieve several Sentinel-2 images taken during 2020, for which we manually annotate plume segmentation labels using the open-source tool Label Studio [27].

B. Power Plant Metadata

We supplement our satellite image dataset with power plant-related information such as the type of fossil fuel used and the power generation rate, both taken from [3]. This external dataset contains actual production data (in MW) per generation unit for each European power plant, allowing us to sum the production of all generation units in each plant at the time the satellite image was acquired. The installed capacity of these power plants ranges from 29 to 5230 MW with a mean of 1177 MW. Our sample includes power plants that use four different types of fuel: hard coal (41%), gas (29%), lignite (29%), and peat ($\leq 1\%$).

C. Environmental Data

We complete our dataset with contemporaneous weather data: the temperature at the surface level (measured in K), relative humidity (%), and wind vector ($\vec{u}, \vec{v}$) from the *ERA-5 dataset* [28], provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). We include these weather variables since they may influence the visual characteristics of plumes: for example, the norm of the wind vector may affect plume width, and humidity and temperature may affect plume visibility.

Our final dataset contains ~ 3000 annotated satellite observations (~ 2000 of them contain plumes) of 153 different fossil fuel power plants, which we carefully divided into train validation (80%) and test (20%) sets. We scale image data channel-wise and weather variables such that all feature values are in the range $[0, 1]$. Moreover, since values of power generation outputs are widely spread, we use the natural logarithm of one plus the target, then we scale them such that they lie in the range $[0, 1]$. Although this step is not mandatory for convergence, it is recommended to avoid large error gradient values making the learning process unstable. This also results in losses that fall within the same order of magnitude, leading

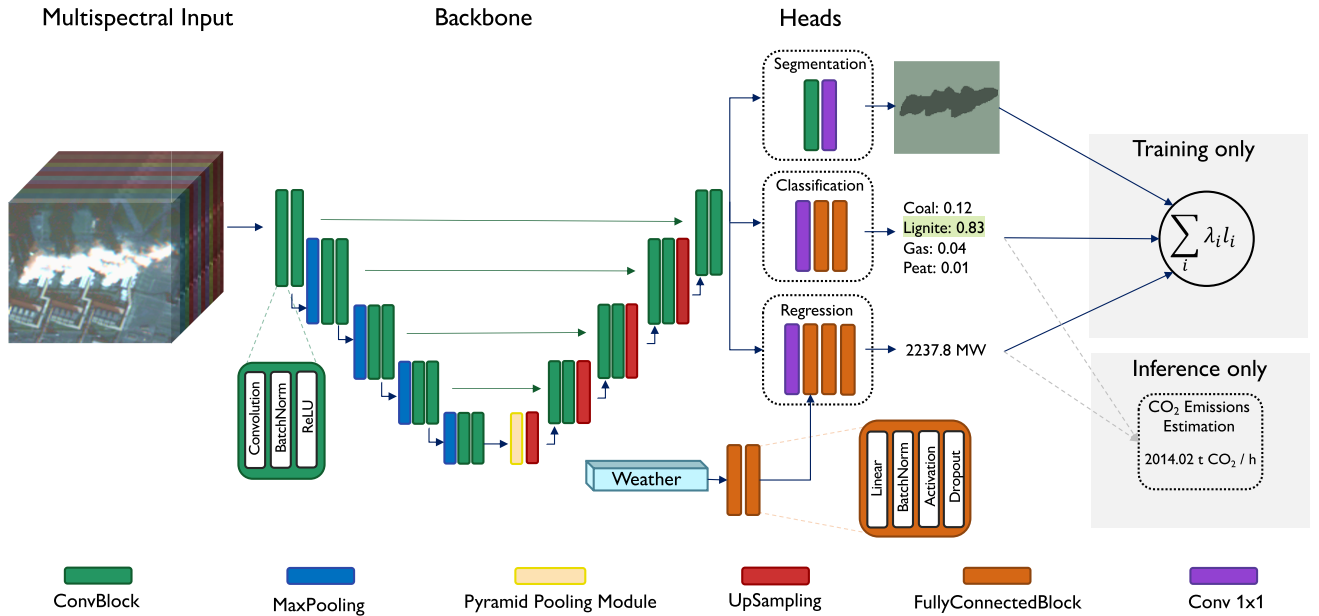


Fig. 2. Network architecture for our proposed multitask learning method. The model takes as input a multispectral satellite image and trains a UNet-like backbone to learn a unique representation that will be shared across the different tasks. A pyramid pooling module [6] is added at the last layer of the backbone's contracting path, as detailed in Section IV. The network is trained to simultaneously learn three different tasks: 1) semantic segmentation of plumes; 2) classification of the type of fired fuel; and 3) regression with respect to power generation rate. For the latter, we append to the shared representation a vector of environmental (weather) variables. Using the predicted power generation rates and the predicted type of fuel, CO₂ emissions can then be estimated.

to a more consistent multitask configuration. Note that all evaluation metrics shown in this work are computed on the unscaled versions of the outputs. For augmentations, we adopt traditional techniques, such as random horizontal and vertical flipping, and random rotations.

In summary, for each observation, our dataset consists of a multispectral Sentinel-2 satellite image centered on a power station, with its plume segmentation mask, the type of fuel used, the actual power generation rate at the corresponding timestamp, and concurrent environmental (weather) data.

IV. APPROACH

In this section, we present our method to estimate power generation and GHG emissions from multispectral satellite images using deep neural networks. Our main objective is to estimate power generation rates, knowing that these values can be used as proxies for GHG emissions.

In Section IV-A, we describe the proposed network architecture. In Section IV-B, we introduce the empirical physical constraints, and, in Section IV-C, we present the task weighting strategy used.

A. Our Proposed Network

Fig. 2 illustrates our overall approach. The proposed approach is a multitask network including a unique shared backbone and three distinct task heads: a regression head to predict power generation rates, a segmentation head to semantically segment plumes, and a classification head to predict the type of fuel used. In our case, where the amount of labeled data is very limited, the use of a shared backbone has several advantages. The first advantage is to reduce the number of trainable parameters and, therefore, to reduce the risk of overfitting on the small training set. The second advantage is to push the network toward learning a robust representation that generalizes well to different related tasks. We detail our approach in the following.

1) *Backbone*: The first part of the proposed method is a UNet-like backbone used as a feature extractor. Its architecture consists of a contracting path (downsampling) and an expanding path (bilinear upsampling), as described in the original paper [29]. Inspired by [30], who apply a pyramid pooling module (PPM) from PSPNet [6] on the last layer of their feature pyramid network (FPN) backbone, we append a similar module at the bottleneck (last layer of the contracting path) of the feature extractor. Essentially, a PPM fuses information from features with different scales, working as a global contextual prior. PPMs are proven to improve pyramid-like model performances by artificially increasing the receptive field [30], [31]. We take advantage of hard parameter sharing through the entire UNet backbone leading to a representation that will be used across all the tasks, reducing the risk of overfitting.

2) *Specific Task Heads*: The three tasks' branches receive the shared representation and specialize each on one task. The *first task* is a spatially dense prediction task, whose role is to segment plumes. Since the fully convolutional backbone leads to coarse outputs, we must connect these outputs back to the pixel-level to produce pixelwise predictions. As a result, we add a 1×1 convolutional layer, which outputs a segmentation map of the same height and width as the input image. The *second task* is a classification task, whose role is to predict the type of fired fuel: hard coal, lignite, gas, or peat. In this case, the shared backbone's output is treated as a feature map and is transformed using linear blocks (dense layers, batch normalization, and activation function) to return a unique value corresponding to the probability that a particular fuel type is used in the power plant. To calculate these probabilities, we use a softmax function. The dense layers and the softmax function represent the prediction head for this classification task. The *third task*, in our case, the most important one, is to estimate actual power generation rates. This task is treated as a regression problem. The setup is very similar to the

classification task described above (second task) in the sense that a regression problem involves the same function as a classification one, but with the difference that the output is a continuous value and not a discrete one. This implies that following a number of linear blocks, we use a ReLU as the activation function, which also constrains the prediction to be positive. These layers constitute the prediction head for the regression task. Additionally, to account for environmental effects, we add weather vector embeddings, generated by a group of fully connected blocks, to the flattened output of the convolutional layer in the prediction head. Our targets for the regression task are power generation rates reported through [3] that are contemporaneous with the Sentinel-2 observations and available for each power plant in our sample.

3) *Losses and Evaluation Metrics:* We perform single-task and multitask training. Each of the three tasks is learned using a specific loss function. For the plume segmentation task (l_{seg}), we use the binary cross-entropy loss, for the regression task (l_{reg}), we use the $l1$ loss, and, for the fuel classification task (l_{cls}), we use the cross-entropy loss. Since the fuel-type classes are very unbalanced, we apply manual class-weighting to the loss function. We regularize our total loss function by adding the L2 norm of all model's weights W as a penalty term. In the multitask learning setup, we weight each loss (λ_i) either manually or using a dynamic weighting strategy (details can be found in Section IV-C and Fig. 3).

The general formulation of the multitask loss is as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \sum_{t=1}^T \lambda_t l_t + \mu \|W\|_2 \\ &= \lambda_1 l_{\text{seg}} + \lambda_2 l_{\text{reg}} + \lambda_3 l_{\text{cls}} + \mu \|W\|_2 \end{aligned} \quad (1)$$

where T is the number of tasks and μ is the regularization hyperparameter. In Section IV-B, we extend our loss function to follow physical principles.

We quantitatively measure the performance of the segmentation task with the intersection over union (IoU), fuel classification task with multiclass accuracy, and the regression task with the mean absolute error (MAE), the R -squared (R^2), and the mean absolute percentage error (MAPE), which is only defined for nonzero inputs, as follows. Note that for the different metrics, y represents the vector of ground-truth values, and $\hat{y}$ the predicted one

$$\text{IoU}(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} \quad (2)$$

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_i^N |\hat{y}_i - y_i| \quad (3)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_i^N (\hat{y}_i - y_i)^2}{\sum_i^N \left(y_i - \frac{1}{N} \sum_i^N y_i \right)^2} \quad (4)$$

$$\text{MAPE}(y, \hat{y}) = \frac{1}{M} \sum_i^N \begin{cases} \frac{|\hat{y}_i - y_i|}{y_i}, & \text{if } y_i \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $M = N(y_i \neq 0)$, with N being the size of the dataset.

4) *CO₂ Emission Estimation:* There is a direct correlation between CO₂ emission rates and power generation rates, as well as other plant-specific properties, such as the fuel type.

Our idea is to derive a conversion factor, which would allow us to estimate the rate of CO₂ emissions by directly using the rate of power generation produced by the power plants. This factor would depend on the type of fuel used. Based on reported annual CO₂ emissions and power generation rates for a selection of hard coal and lignite power plants extracted from the “EU’s dirty 30” report [32], we derive average conversion factors from emissions (0.9 ± 0.1 t CO₂/MWh for lignite and 0.6 ± 0.1 t CO₂/MWh for hard coal) to convert our predicted power generation rates to CO₂ emission rates. These factors represent the average amount of CO₂ emitted per 1 MWh. Multiplying the predicted power generation rate with the conversion factor corresponding to the predicted type of fuel provides an estimate of the total emissions of CO₂. Unfortunately, conversion factors can only be derived for hard coal and lignite, since we only have ground-truth data on CO₂ emissions from power plants using these fuels.

Among the power plants in our test dataset, nine are cited in the report [32], so we know their exact emission factors. We thus compare in Section V the predicted CO₂ emission rates with the ground truth of these nine hard coal and lignite power plants, corresponding to 97 observations in total. Ground-truth emission rates are derived for these nine plants by multiplying plant-specific emission factors with actual power generation rates, while predicted emission rates are derived by multiplying average conversion factors specific to the fuel, depending on the predicted type of fuel, by the predicted power production rate.

B. Empirical Physical Constraints

As stated previously, plume visibility is strongly affected by weather conditions. Since we use visual characteristics of plumes to estimate the rate of power generation, it is important to take into account these environmental conditions and therefore to integrate some physical knowledge into the learning process, to make the prediction more physically realistic.

In this work, we do not distinguish between actual smoke resulting from the combustion of fossil fuels and water vapor from cooling devices in our plume segmentation model. However, the majority of plumes that we observe in the satellite images emerge from cooling towers.

In this section, we will briefly explore the physical processes that lead to plume formation. At the outlet of cooling towers, exhaust gases are much warmer, more humid, and have a higher dew point than ambient air. As soon as the flow of warm moisture-saturated air comes into contact with the external atmosphere, especially during the winter period when the temperature outside drops, it suddenly cools. When the stream temperature drops below the dew point, some of the exhaust gas condenses into characteristic vapor plumes. This means that above a certain temperature, or below a certain relative humidity, the likelihood of observing plumes is greatly reduced. This hypothesis is also confirmed empirically by our observations: very weak plumes are observed when temperatures are above a threshold between 30 °C and 35 °C, and the same is true for relative humidity below about 30%, despite actual power generation at the time of observation.

We include these physical considerations in the learning process by incorporating them as constraints in the objective

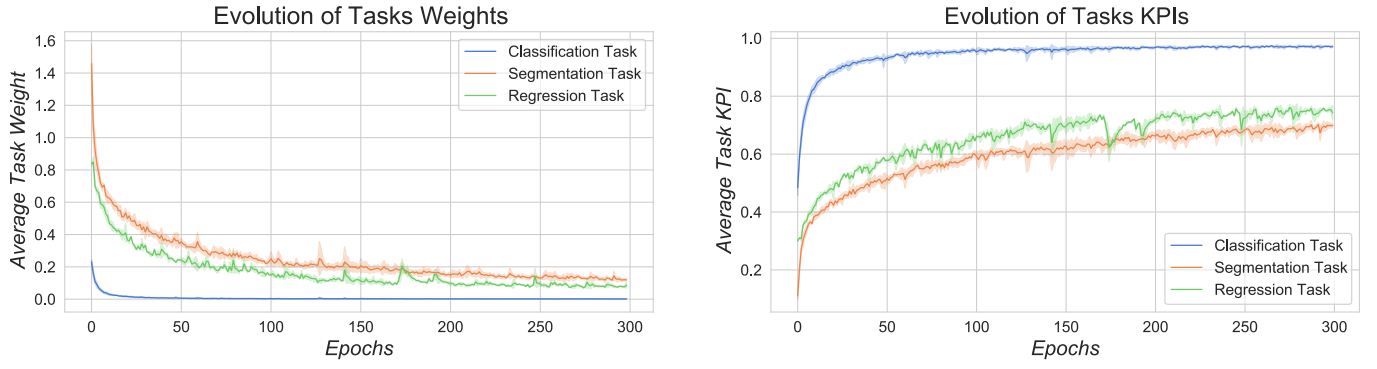


Fig. 3. Dynamic task weighting strategy (Section IV-C). The x -axis denotes the training epoch number. (Left) Average weight over all iterations per epoch, for each task, and (right) average KPI over all iterations per epoch, for each task. All plots are displayed with their corresponding standard deviation over five different runs (shaded areas). We note that (left) tasks' weights decrease at the same rate as (right) KPIs increase.

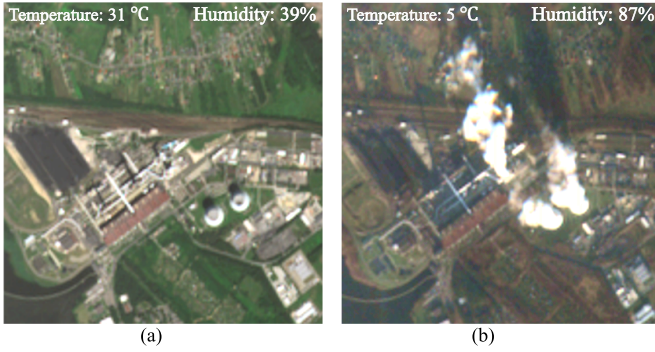


Fig. 4. Two examples of the effect of weather conditions on plumes' visual appearance. Due to the condensation phenomenon, plumes are more visible when the weather is cold and moist. Note that in both images, the power plant is running and producing electricity: (left) 405 MW and (right) 648 MW. (a) Low humidity, high temperature. (b) High humidity, low temperature.

function, which impacts both the segmentation and regression tasks. This section specifically addresses scenarios where temperatures are high and/or humidity is low, as under these conditions, large plumes are unlikely to occur, but power plants could still be emitting power (see Fig. 4). To account for this, we penalize the learning process if the predicted plume size or power generation predictions deviate considerably from their actual values. More precisely, we will penalize the learning of the regression task **if** the power generation prediction is much lower than the true one by scaling the loss l_{reg} , and we will penalize the learning of the segmentation task **if** the predicted size (area) of plumes is much larger than the real one. The penalties are proportionate to the deviation of humidity and temperature weather variables (h , t) from their extreme values (H_{max} , T_{min}). Essentially, this would allow the model to adjust its level of caution based on specific environmental conditions and patterns observed in the image, thus increasing the overall accuracy and reliability of the model's predictions and actions

$$l_{seg} = l_{bce} \cdot (1 + (H_{max} - h) \cdot k_{1seg} + (t - T_{min}) \cdot k_{2seg}) \quad (6)$$

$$l_{reg} = l_1 \cdot (1 + (H_{max} - h) \cdot k_{1reg} + (t - T_{min}) \cdot k_{2reg}) \quad (7)$$

where

- H_{max} , T_{min} max. humidity and min. temperature;
- h , t humidity and temperature of the observation;
- k_{1seg} , k_{2seg} scaling factors;
- k_{1reg} , k_{2reg} scaling factors.

This means that the smaller the h or the higher the t , the larger the penalty. The penalty is implemented as a scaling of the losses l_{bce} and l_1 , which is equivalent to scaling (increasing by the same factor) the learning rate and consequently increasing the amount that the weights are updated during training. We are aware that too high a learning rate could cause the model to converge too quickly toward a suboptimal solution. Nevertheless, this is not an issue for us, since we use additional subsequent loss weighting (Section IV-C). In addition, our ablation study found that the optimal values for k_{1seg} , k_{2seg} , k_{1reg} , and k_{2reg} were 0.5, and as such, these were set accordingly. It may be advantageous to adjust these four scaling factors when there is a significant disparity in precision between the two variables (temperature or humidity), to reduce the impact of the less-precise variable on the applied penalty. We show the utility of these constraints in Section V-C.

C. Dynamic Task Weighting

In a multitask setup, each of the tasks has its own separate loss function, and these are combined to form a single aggregate loss function for the model to learn to minimize. The question is how to combine these multiple loss functions into one. The most straightforward approaches are to first, naively give the same weight to each of the losses, then in a second step, to manually tune these weights one at a time. It is obvious that this technique is extremely time-consuming, since the search space is infinitely large. Motivated by [19], we implement a self-paced loss weighting strategy by dynamically prioritizing tasks based on their key performance indicator (KPI), denoted by κ .

Namely, the goal is to dynamically adjust a task's weight during training depending on its training performance using a task difficulty indicator. The proposed indicator is a focal loss [33] of κ and a focusing parameter γ : $FL(\bar{\kappa}_t, \gamma_t)$. γ 's role is to adjust the rate at which easy tasks are downweighted: a bigger γ implies a stronger downweighting of easy tasks. Essentially, this approach allows the easiest tasks (those which have a high KPI) to be given less importance during the training, and the difficult tasks to be given more importance. At each training iteration τ , the weights λ_t associated with the loss for each task t [see (1)] are computed in two steps. First, the exponential moving average of the KPIs $\bar{\kappa}$ is calculated to track the evolution of the task performance with respect to previous iterations [see (8)], then the corresponding loss

TABLE I
RESULTS FOR SINGLE VERSUS MULTITASK NETWORKS USING RGB INPUTS (TOP BLOCK) AND MULTISPECTRAL INPUTS (BOTTOM BLOCK)

		Task Weights (λ_i)			Segmentation	Regression		Classification	
		Segmentation	Regression	Classification	IoU (%)	MAE	R ² (%)	MAPE (%)	Accuracy (%)
RGB	Single	1	0	0	55 ± 2	-	-	-	-
		0	1	0	-	218 ± 21	55 ± 5	60 ± 2	-
		0	0	1	-	-	-	-	87 ± 1
	Multi	0.33	0.33	0.33	57 ± 1	232 ± 17	48 ± 2	61 ± 3	88 ± 1
		0.15	0.7	0.15	53 ± 1	202 ± 6	62 ± 5	53 ± 2	89 ± 1
		Dynamic Weighting			57 ± 1	178 ± 5	70 ± 4	50 ± 5	88 ± 1
Multispectral	Single	1	0	0	59 ± 1	-	-	-	-
		0	1	0	-	202 ± 20	65 ± 5	60 ± 1	-
		0	0	1	-	-	-	-	90 ± 1
	Multi	0.33	0.33	0.33	61 ± 1	194 ± 9	63 ± 5	57 ± 5	92 ± 1
		0.15	0.7	0.15	62 ± 1	181 ± 6	69 ± 3	56 ± 1	94 ± 1
		Dynamic Weighting			64 ± 0	157 ± 4	78 ± 3	43 ± 5	93 ± 1

Note: Multitask networks are trained using different loss weighting strategies: equal weighting (first row), manual weighting by giving more importance to the regression task (second row), and dynamic weighting based on the performance of each task during training (see Section IV-C) (third row). Results are reported with their standard deviations from 5 runs. The best performing model is highlighted in bold.

weight is calculated [see (9)]

$$\bar{\kappa}_t^{(\tau)} = \alpha \kappa_t^{(\tau)} + (1 - \alpha) \bar{\kappa}_t^{(\tau-1)} \quad (8)$$

$$\begin{aligned} \lambda_t^{(\tau)} &= \text{FL}(\bar{\kappa}_t^{(\tau)}, \gamma_t) \\ &= -(1 - \bar{\kappa}_t^{(\tau)})^{\gamma_t} \log(\bar{\kappa}_t^{(\tau)}) \end{aligned} \quad (9)$$

where

$\bar{\kappa}_t^{(\tau)}$ exponential moving average of KPI at iteration τ ;

α discount factor, between 0 and 1;

γ_t focusing parameter of task t .

γ_t was fixed to 1 and α was set to 0.8. A large α gives more importance to recent values of KPIs (recent iterations).

Since the KPIs (κ_t) should be in the $[0, 1]$ range, we define a KPI κ_t for each task t in the following way, with N being the sample size and $\mathbb{1}(\cdot)$ the indicator function.

- 1) For the fuel-type classification task, we use the multi-class accuracy value

$$\kappa_{\text{cls}} = \frac{1}{N} \sum_i \mathbb{1}_{\{y_i = \hat{y}_i\}}. \quad (10)$$

- 2) For the plume segmentation task, we use the proportion of points that have an IoU higher than a threshold. We used a threshold $T = 0.7$

$$\kappa_{\text{seg}} = \frac{1}{N} \sum_i \mathbb{1}_{\{\text{IoU}(y_i, \hat{y}_i) \geq T\}}. \quad (11)$$

- 3) For the power generation regression task, we use the proportion of points that have an MAE lower than a threshold. We used a threshold $T = 120$ MW

$$\kappa_{\text{reg}} = \frac{1}{N} \sum_i \mathbb{1}_{\{\text{MAE}(y_i, \hat{y}_i) \leq T\}}. \quad (12)$$

In Fig. 3, we display the evolution of task weights (on the left) with the evolution of their KPIs (on the right). For the sake of clarity, we only display the average value per

epoch (and not per iteration). Quantitative results of those different weighting strategies (naive approach, manual tuning, and dynamic prioritization) are available in Table I.

D. Sensitivity Maps

To get a better understanding of the features and regions that our model focuses on for each of the tasks, we look at the derivative of each of the responses (losses) with respect to the input. We compute a sensitivity map ($\mathcal{S}$), which is the gradient of the maximum responding output with respect to the input. Note that this definition is different from the gradient computed in backpropagation, which is derived with respect to the parameters W

$$\mathcal{S} = \nabla_x f(x; w). \quad (13)$$

However, sensitivity maps can be very noisy, due to the local irregularity of the network's response as a function of the input. For this reason, we apply the SmoothGrad technique [34]: for each input image, we create N slightly noisy versions of it by adding a small amount of Gaussian noise and compute the average gradient as follows:

$$\tilde{\mathcal{S}} = \bar{\nabla}_x f(x; w) = \frac{1}{N} \sum_{n=1}^N \nabla_x f(x + \epsilon_n; w) \quad (14)$$

where ϵ_n is the i.i.d of distribution $\mathcal{N}(0, \sigma^2 \mathbf{I})$. σ should be set to a fraction of the gap between the maximum and the minimum of the pixel values (we use $\sigma = 0.25$).

V. EXPERIMENTS

A. Implementation Details

Our method outlined in Section IV and Fig. 2 is implemented using the PyTorch framework [35]. The model is trained on the training set, and a grid search is used to determine the hyperparameters based on the validation set.

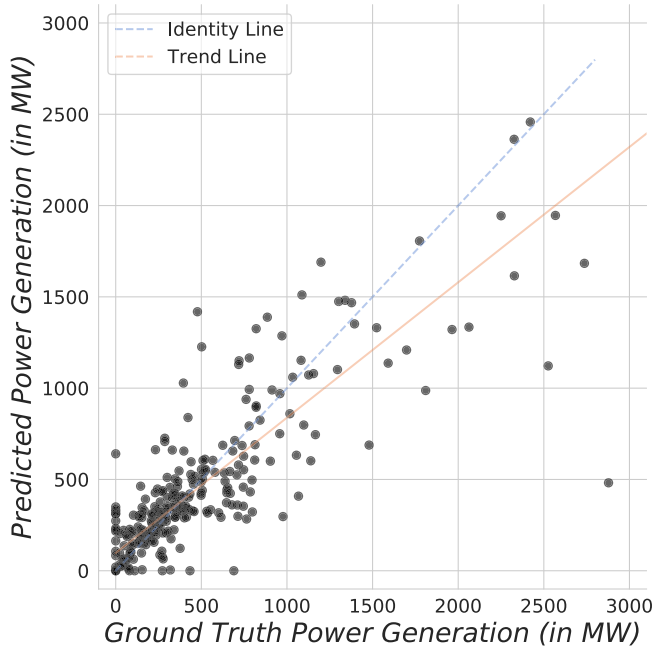


Fig. 5. Predicted versus ground-truth power generation rates in MW, on our full test set. Data points are displayed in transparent black. In blue, the identity line ($f(x) = x$), in red, the trend line fit on the data points, using a least-squares polynomial fit. Our model reached an MAE of 157 MW, RMSE of 254, and an R^2 of 78%.

The SGD optimizer was used during training. The learning rate was set to 0.01. We trained our model for 300 epochs with a batch size of 32 using 1 Nvidia Tesla V100 GPU.

B. Model Performance

We compare our multitask model to the three single-task models (fuel-type classification only, power generation regression only, and plume segmentation only), which correspond to our separate baselines. Evaluation metrics (Section IV-A3) are shown in Table I. In Fig. 5, we display the predicted power generation rate versus the ground truth for our test sample based on the multitask approach using multispectral inputs and the dynamic weighting strategy (last row of Table I), resulting in an R^2 of 0.78 and an MAE of 157 MW. Our multitask model achieves the best average performance in the multispectral setup and surpasses the segmentation baseline by 5% points (which corresponds to a relative increase of +8%) and the classification baseline by 3% points (+3%). It also surpasses the regression baseline by 45 points of MAE (−22%), 13% points of R^2 (+20%), and 17% points of MAPE (−28%). CO₂ emission rates are estimated from predicted power generation rates utilizing fuel-specific emission factors (see Section IV-A4) for nine plants in our test set for which such ground-truth information is available; the derived MAE for CO₂ is 166 t/h.

C. Importance of Empirical Physical Constraints

In Table II, we validate the effectiveness of the empirical physical constraints. We first apply the penalty-scaled loss functions, as described in Section IV-B in (6) and (7) and display the results in the top row. Note that these values correspond to those we find in the last row of Table I. In the second step, we decide to use only the nonscaled loss functions

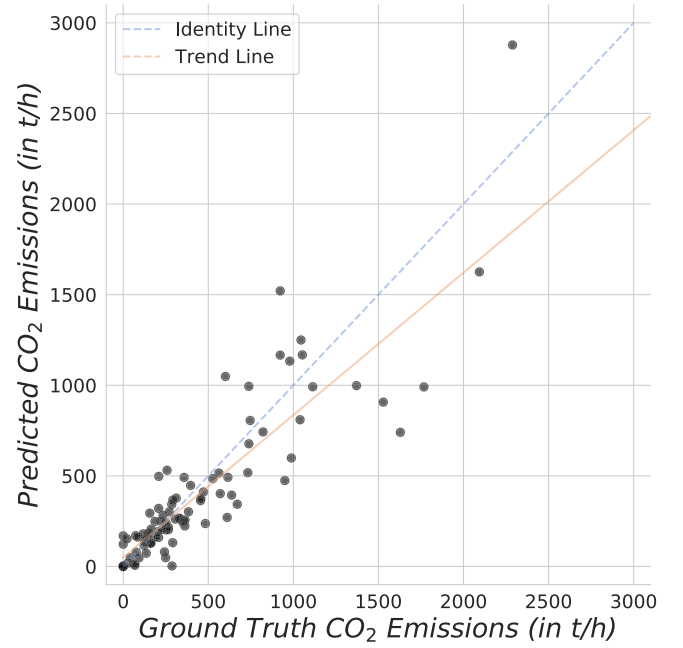


Fig. 6. Predicted versus ground-truth CO₂ emissions in t/h, on a subset of nine different power plants (97 observations) from our test set for which CO₂ conversion factors are available (see Section IV-A4). Data points are displayed in transparent black. In blue, the identity line ($f(x) = x$), in red, the trend line fit on the data points, using a least-squares polynomial fit. Our model reached an MAE of 166 t/h, RMSE of 237, and an R^2 of 73%.

TABLE II
IMPACT OF EMPIRICAL PHYSICAL CONSTRAINTS

	Segmentation IoU (%)	MAE	Regression R^2 (%)	MAPE (%)	Classification Accuracy (%)
guided	64 ± 0	157 ± 4	78 ± 3	43 ± 5	94 ± 1
un-guided	63 ± 1	164 ± 6	75 ± 3	48 ± 5	94 ± 2

Note: Results in the multitask setup using dynamic weighting for multispectral inputs. Results are reported with their standard deviations from 5 runs.

(l_{bce} for the segmentation task, and l_1 for the regression task; see Section IV-A3) and report the results in the bottom row of Table II. We notice that the physics-related penalty term in the loss function improves the performances, especially for the regression task, where we observe a relative improvement of 4% for the MAE, 4% for the R^2 , and 10% for the MAPE.

D. Weather Ablation Study

Table III compares the effect of different combinations of weather variables on the performance of the regression task. We use a concatenation of the candidate set of weather embeddings (see Section IV-A2) with the flattened shared representation as the input of the regression task head (see Fig. 2). For a fair comparison, we do not use any physical constraints (Section IV-B) in the loss functions. As a baseline, we do not use any weather variables. This leads to the worst performance among all combinations. As can be seen from Table III, all other combinations perform significantly better than the baseline, validating our hypothesis that adding external environmental knowledge into the model is useful for the power generation estimation. Indeed, the norm of the wind affects the width of the plume and the humidity and temperature affect the visibility of the plume; ignoring these variables could distort the estimate. Including all variables lead to the best estimate in terms of MAE and R^2 .

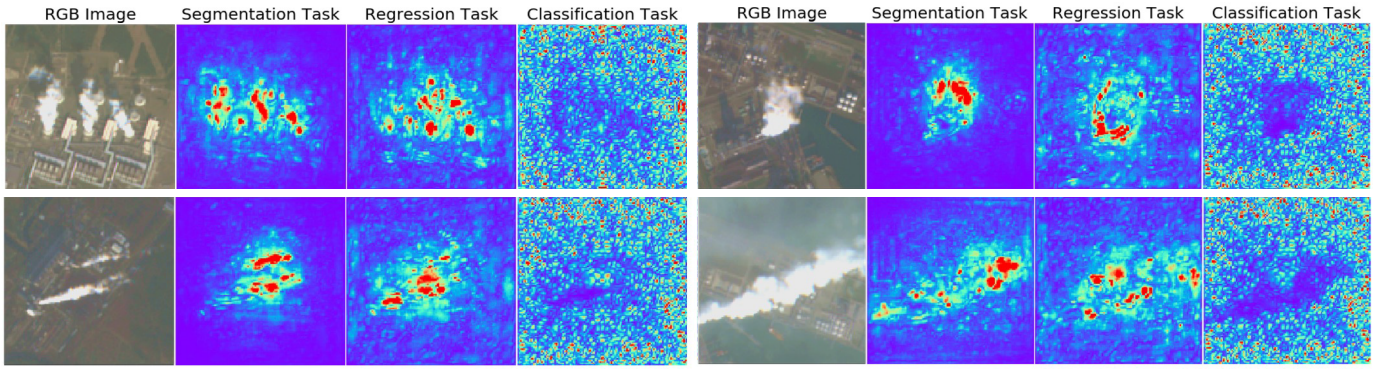


Fig. 7. Task-specific sensitivity maps ($\tilde{S}$) for four random input images. Color red denotes more important regions in the image and color blue highlights less important ones. Our multitask model focuses on different areas in an image for the different tasks, such as only plume extent for segmentation, plume and some surroundings for regression, and anything but plume for classification.

TABLE III
ABLATION STUDY ON IMPACT OF WEATHER VARIABLES
ON THE REGRESSION TASK PERFORMANCE

Temperature	Humidity	Wind	MAE	R ²	MAPE
			187 ± 4	66 ± 6	58 ± 2
✓			177 ± 2	69 ± 3	52 ± 4
	✓		179 ± 8	68 ± 6	52 ± 6
		✓	175 ± 3	70 ± 2	49 ± 5
	✓	✓	174 ± 4	68 ± 2	51 ± 3
✓		✓	169 ± 5	71 ± 3	50 ± 4
✓	✓		171 ± 3	73 ± 2	47 ± 4
✓	✓	✓	164 ± 6	75 ± 3	48 ± 5

Note: Results on the test set, using multi-spectral inputs and the multitask model. Results are reported with their standard deviations from 5 runs. The best performing model is highlighted in bold.

E. Comparison of Loss Weighting Strategies

In Table I, we study the impact of dynamically weighting our losses during training using each task's KPI (Section IV-C) compared to weighting them manually. The manual weighting was tested with a fixed number of combinations, and the one that led to the best average results was displayed. We thus compare the equal weighting and the manual tuning of the weight with the dynamic weighting introduced above. As validated by Table I, dynamically weighting our losses leads to the highest (or comparable) performances across all tasks: compared to manually weighting the losses, we find a relative decrease of 1% for the classification task and a relative increase of 3% for the segmentation task. Same behavior for the regression task with a relative decrease of 13% for the MAE, 23% for the MAPE, and an increase of 13% for the R^2 .

F. Importance of Multispectral Inputs

In this work, we use a multispectral satellite imagery dataset, containing bands well beyond the visible spectral range (see Section III). Especially near-infrared and short-wave infrared bands are very useful for the distinction and identification of certain materials that may look similar in the visible range: for example, a dense natural cloud and a plume of smoke coming out of a cooling tower may look the same in the visible spectrum, but would definitely differ in the full spectrum. In Table I, we validate the importance of using all bands available rather than simply limiting ourselves to the

RGB (visible) bands. We note that using RGB inputs leads to the worst performance among all setups (baselines and multitasks). For example, for the multitask model using the dynamic weighting strategy, there is a relative improvement of +12% for the segmentation task, +7% for the classification task, and −12% for the MAE of the regression task, when using multispectral inputs compared to RGB ones. Nevertheless, this spectral ablation study shows the feasibility of our approach even with RGB-only data within certain limitations.

G. Visualization of Sensitivity Maps

Fig. 7 presents sensitivity maps (see Section IV-D) with respect to the three task-specific losses for four different inputs. The high-valued regions of $\tilde{S}$ highlight the areas of the inputs that matter the most for the corresponding task, that is, regions in the input where a small change would result in a significant change in the loss. Notably, we observe that the plumes are the most important parts of a satellite image for the plume segmentation task. However, for the fuel classification task, all the parts related to the power plant are important except for the plumes that emanate from it. Concerning the power generation regression task, we note that while the size and shape of the plumes are important factors, they do not seem to be the only determinants.

H. Qualitative Comparison

We conduct a visual inspection of the results to supplement our numerical evaluation of the approach. Fig. 8 presents a qualitative comparison of our results under various configurations. It displays the predicted plume segmentation with the IoU, predicted fuel type, and predicted power production. The single-task baselines are presented first, followed by the multitask model trained on the RGB bands of Sentinel-2 data, the multitask model trained on all bands of Sentinel-2 data, and finally the multitask model trained on all bands of Sentinel-2 data and guided by physical considerations.

VI. DISCUSSION

This work presents a novel method for predicting power generation and GHG emission rates from fossil fuel power plants using freely and globally available satellite imagery and environmental information and a multitask deep-learning

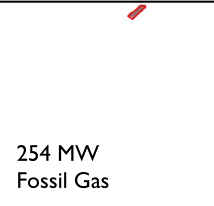
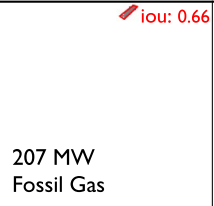
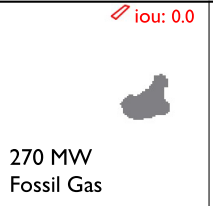
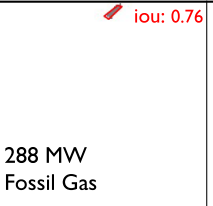
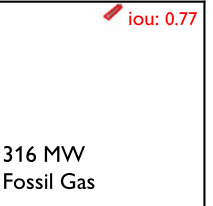
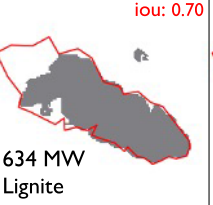
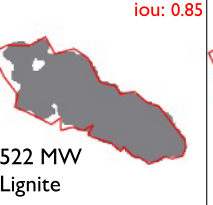
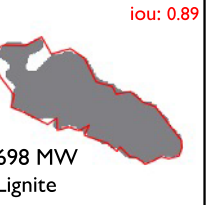
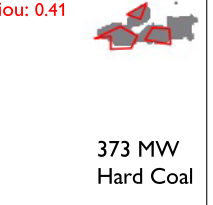
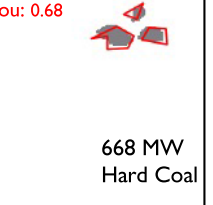
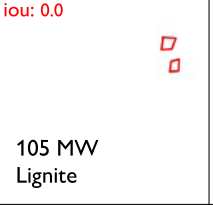
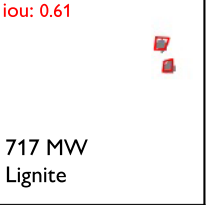
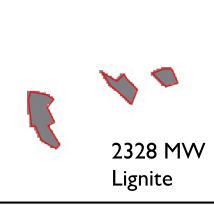
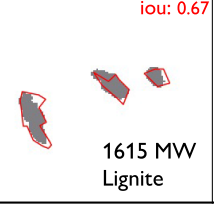
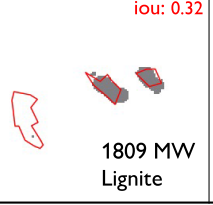
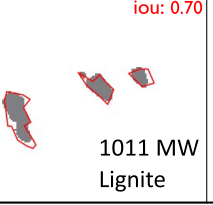
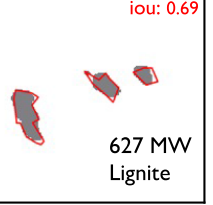
RGB Image	Ground Truth	Single-task Baseline	Multitask RGB Model	Multitask Model	PG-Multitask Model
		 iou: 0.66	 iou: 0.0	 iou: 0.76	 iou: 0.77
		 iou: 0.85	 iou: 0.70	 iou: 0.85	 iou: 0.89
		 iou: 0.52	 iou: 0.43	 iou: 0.41	 iou: 0.68
		 iou: 0.49	 iou: 0.0	 iou: 0.38	 iou: 0.61
		 iou: 0.67	 iou: 0.32	 iou: 0.70	 iou: 0.69

Fig. 8. Qualitative and quantitative comparison of results for five different randomly selected inputs. Results from left to right: Sentinel-2 true color (RGB), ground truth, single-task baselines, multitask model trained on RGB inputs, and multitask model trained on multispectral inputs, and finally physics-guided multitask model trained on multispectral inputs. For all results, we show the power generation in MW, the type of fossil fuel, the segmentation mask output with ground-truth plume boundary in red, as well as the corresponding IoU value.

approach. Model to predict those two values, based on one satellite image. Based on Table I, we find that our multitask model outperforms the single-task baseline models trained on the same data for all tasks. Among the factors leading to this success is the self-paced learning of task weights by **dynamically prioritizing tasks**. Rather than keeping losses' weights static throughout training, they are updated based on the performance of their respective tasks, prioritizing those that are difficult, and giving less importance to those that are easier. Fig. 3 shows that the classification task is much easier to learn than the other two, and, unexpectedly, the segmentation task seems more difficult than the regression. Note that these conclusions can only be made relative to the threshold T we used in the KPI calculation [see Section IV-C; (11) and (12)]: if we use a smaller T for the IoU, the KPI would be naturally higher, and vice versa. Nevertheless, this made it necessary to understand the decision process related to each of these tasks, such as the regions in the satellite image that matter most. For this purpose, we generate **sensitivity maps** of different inputs, as described in Section IV-D. Fig. 7 indicates that, as expected, the most important parts of a satellite image

for the plume segmentation task are the plumes themselves. We notice the complete opposite for the fuel classification task: all the parts related to the power plant are important, excluding the plumes emanating from it. This makes sense, considering that most visible plumes in our dataset contain only water vapor and do not provide any insights into the type of fuel used, whereas the surrounding area might provide additional information such as the plant size, number of generation or cooling units, or geographical location. As for the power generation regression task, we notice that the size and shape of the plumes are not the only determining factors. This is consistent with Fig. 8, which shows that large plumes are not always associated with more power generation, otherwise a simple calculation would make it possible to estimate the power generation directly from plume areas. For example, other factors that might affect power generation are the number of power units in each plant and environmental conditions at the time of observation. Another important factor enabling the success of our approach is the use of **multispectral satellite images**, rather than only the usual RGB bands. Indeed, our quantitative (Table I) and qualitative (Fig. 8) results highlight

the importance of using all available bands. With only RGB inputs (Fig. 8), some natural clouds (first image) and whitish areas (second and third images) are confused with plumes, which does not happen when the multispectral data are used. More precisely, using all spectral bands improves the accuracy of our model by providing more detailed information about the satellite image, thus allowing for more accurate identification of plumes, and reducing the number of false positives. Finally, we notice that integrating **empirical physical constraints** into the learning process improves all tasks (Table II), particularly the regression task. However, it should be noted that the low resolution of the environmental data used in this study could introduce uncertainties in the results. Despite this limitation, the results in Table III clearly show that the **environmental data** significantly affects the model's performance. Further research with higher-resolution environmental data would be necessary to improve upon these findings.

Regarding the quantitative performance of the regression task, we can observe the quality of the predictions in Fig. 5 and Table I. The trend of the predictions follows that of the ground truths rather well. Moreover, the value of R^2 (0.78, from Table I) also highlights the goodness of fit for the regression task. Visually, we notice that the smaller the power generation rate, the better the fit, and that data points tend to spread out for large values. This is also what the red (trend) and blue (identity) lines suggest, with a visible divergence for large values. This behavior is to be expected given the **large range of power generation values** and a significant concentration in the smaller range. However, it is important to note that the dataset used for this analysis is limited in size due to the cost of labeling, which could introduce uncertainties in the predictions. Therefore, while the evaluation metrics (R^2 , MAE, etc.) and the visual inspection of the predictions appear to indicate a good fit, gathering more data in the future would be necessary to strengthen and confirm these findings. Regarding the segmentation task, our model is designed to segment both **large plumes and very small ones** (e.g., the first image of Fig. 8) in relatively large images. Although the IoU may not seem high, this value should be interpreted in context: empirically, we found that the performance of the segmentation task is on par with that of human annotation, meaning that different annotators have labels that overlap with a similar IoU. Lastly, emission conversion factors utilized in the conversion from power generation to CO₂ emission are empirically based on fuel-specific statistical considerations. While uncertainties introduced by this simplification are propagated into our CO₂ emission rates (see Fig. 6), power plant-specific emission factors might lead to better results, but for now, are only available for few power plants (representing less than 10% of our dataset). Note that plant efficiency is not considered in this work due to the unavailability of data, but we are aware that this information may affect our estimates. Nevertheless, our approach allows an estimation of CO₂ emission rates to within 166 t/h, and knowing that the power plants in our sample have an average capacity of 1177 MW, we can therefore say that our power generation predictions are of high confidence and that, consequently, our method can contribute to the estimation of CO₂ emission and power generation rates from individual power plants on a global scale.

On a final note, the complexity of our dataset construction, its relatively small size, and the lack of comprehensive

ground-truth emission data for all fuel types corroborate the challenges posed in the introduction on the lack of data availability. Data were gathered from three sources: the JRC Open Power Plants Database [26], ENTSOE [3], and E-PRTR [36]. Even using these three sources, only 153 out of 300 power plants in the JRC database had power generation reports, and only nine out of those had CO₂ emission data in the dirty 30 report [32], and for only two fuel types, coal and lignite. We believe this study brings us a step closer to the objective of independent monitoring and transparency in production and emission data for power plants.

VII. CONCLUSION

This work is the first to show the possibility to predict power generation rates with high confidence and CO₂ emission rates with some confidence from Sentinel-2 satellite images. Our proposed deep multitask architecture was trained on a combination of three tasks (plume segmentation, fuel-type classification, and power generation regression). Experiments confirmed that the multitask model outperforms the single-task baselines, for all tasks. To ensure physically realistic outputs, we introduced physical constraints in the regression and segmentation loss functions, taking into account the plume condensation phenomenon, which improved the predictions: our model is thus able to predict power generation rates from individual satellite images with $R^2 = 0.78$ or within 157 MW (MAE) and CO₂ emission rates within 166 t/h (MAE).

As part of our future work, we would like to investigate the use of self-supervised techniques to pretrain the backbone of our model using unlabeled satellite data, which are widely available, before fine-tuning the task-specific heads on the limited available labeled data. These techniques have proven to be very effective in many remote-sensing applications [37], [38].

REFERENCES

- [1] J. M. Reilly, H. D. Jacoby, and R. G. Prinn, *Multi-Gas Contributors to Global Climate Change*. The Center for Climate and Energy Solutions, Pew Center on Global Climate Change, 2003.
- [2] OE Statistics. (2015). *What are Europe's Biggest Sources of Carbon Emissions?* [Online]. Available: <https://www.weforum.org/agenda/2015/11/what-are-europes-biggest-sources-of-carbon-emissions/>
- [3] Entsoe. (2021). *Actual Generation per Generation Unit Table*. [Online]. Available: <https://transparency.entsoe.eu/generation/r2/>
- [4] M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.
- [5] J. Baxter, "A Bayesian/information theoretic model of learning to learn via multiple task sampling," *Mach. Learn.*, vol. 28, no. 1, pp. 7–39, Jul. 1997.
- [6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
- [7] T. Yin, L. Byers, L. Valeri, and J. Friedrich, "Estimating power plant generation in the global power plant database," World Resour. Inst., Washington, DC, USA, Tech. Note, 2020. [Online]. Available: <https://wri.org/upload/estimating-power-plant-generation-in-the-global-power-plant-database>
- [8] F. Liu et al., "A methodology to constrain carbon dioxide emissions from coal-fired power plants using satellite observations of co-emitted nitrogen dioxide," *Atmos. Chem. Phys.*, vol. 20, no. 1, pp. 99–116, Jan. 2020.
- [9] Y. Hu and Y. Shi, "Estimating CO₂ emissions from large scale coal-fired power plants using OCO-2 observations and emission inventories," *Atmosphere*, vol. 12, no. 7, p. 811, Jun. 2021.
- [10] M. Gray, L. Watson, S. Ljungvaldh, and E. Morris, "Using satellite imagery to estimate the utilisation of fossil fuel power plants," Carbon Tracker, London, U.K., Tech. Rep., 2020.

- [11] H. D. Couture et al., "Towards tracking the emissions of every power plant on the planet," in *Proc. Tackling Climate Change With Mach. Learn. Workshop NeurIPS*, vol. 3, 2020.
- [12] L. Scheibenreif, M. Mommert, and D. Borth, "Toward global estimation of ground-level NO₂ pollution with deep learning and remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705914.
- [13] Y. Wang, W. Ding, R. Zhang, and H. Li, "Boundary-aware multitask learning for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 951–963, 2021.
- [14] X. Zhang et al., "Multitask learning for ship detection from synthetic aperture radar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8048–8062, 2021.
- [15] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multitask learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1480–1484.
- [16] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.
- [17] M. Mommert, M. Sigel, M. Neuhausler, L. Scheibenreif, and D. Borth, "Characterization of industrial smoke plumes from remote sensing data," in *Proc. Tackling Climate Change With Mach. Learn. Workshop NeurIPS*, 2020.
- [18] M. Mommert, L. Scheibenreif, J. Hanna, and D. Borth, "Power plant classification from remote imaging with deep learning," 2021, *arXiv:2107.10894*.
- [19] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *Proc. ECCV*, 2018, pp. 282–299.
- [20] D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. Zhu, and G. Camps-Valls, "Toward a collective agenda on AI for Earth science data analysis," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 88–104, Jun. 2021.
- [21] S. Seo, S. Arik, J. Yoon, X. Zhang, K. Sohn, and T. Pfister, "Controlling neural networks with rule representations," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds. Curran, 2021, pp. 11196–11207. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/5cd5058bca53951ffa7801bcd421651-Paper.pdf
- [22] A. Karpatne et al., "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2318–2331, Oct. 2017.
- [23] A. Daw, A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," 2017, *arXiv:1710.11431*.
- [24] E. de Bézenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: Incorporating prior scientific knowledge," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124009.
- [25] H. Shen and L. Zhang, "Mechanism-learning coupling paradigms for parameter inversion and simulation in Earth surface systems," *Sci. China Earth Sci.*, vol. 66, no. 3, pp. 568–582, Mar. 2023.
- [26] K. Kanellopoulos, M. De Felice, I. H. Gonzalez, A. B. Dumitriu, and A. Uihlein, *The Joint Research Centre Power Plant Database (JRC-PPDB)*, EUR 29806 EN, JRC117303, Publications Office of the European Union, Luxembourg, 2019, doi: [10.2760/5281](https://doi.org/10.2760/5281).
- [27] M. Tkachenko, M. Malyuk, N. Shevchenko, A. Holmanyuk, and N. Liubimov. (2021). *Label Studio: Data Labeling Software*. [Online]. Available: <https://github.com/heartexlabs/label-studio>
- [28] H. Hersbach et al., "ERA5 hourly data on pressure levels from 1979 to present. Copernicus climate change service (C3S) climate data store (CDS)," Tech. Rep., 2020, doi: [10.24381/cds.bd0915c6](https://doi.org/10.24381/cds.bd0915c6).
- [29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, doi: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [30] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. ECCV*, 2018.
- [31] A. Kirillov, K. He, R. Girshick, and P. Dollár. (2015). *A Unified Architecture for Instance and Semantic Segmentation*. [Online]. Available: <http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf>
- [32] K. Gutmann, J. Huscher, D. Urbaniak, A. White, and C. Schaible, "Europe's dirty 30: How the EU's coal-fired power plants are undermining its climate efforts," Climate Action Netw. (CAN) Eur., Undermining Environ. Alliance (HEAL), WWF Eur. Policy Office, Eur. Environ. Bur. (EEB), Climate Alliance Germany, Brussels, Belgium, Tech. Rep., 2014.
- [33] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
- [34] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.
- [35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [36] E-PRTR. (2021). *European Pollutant Release and Transfer Register*. [Online]. Available: <https://www.eea.europa.eu/data-and-maps/data/industrial-reporting-under-the-industrial>
- [37] L. Scheibenreif, M. Mommert, and D. Borth, "Contrastive self-supervised data fusion for satellite imagery," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 705–711, 2022, doi: [10.5194/isprs-annals-V-3-2022-705-2022](https://doi.org/10.5194/isprs-annals-V-3-2022-705-2022).
- [38] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1421–1430.



Joëlle Hanna (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronics engineering and the M.Sc. degree in electrical and electronics engineering with a focus on data science from École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2018 and 2020, respectively. She is currently pursuing the Ph.D. degree in computer science with the Artificial Intelligence and Machine Learning Chair, University of St. Gallen, St. Gallen, Switzerland.

Her research interests include deep-learning approaches for remote-sensing data.



Damian Borth (Member, IEEE) received the Ph.D. degree from the University of Kaiserslautern, Kaiserslautern, Germany, in 2014.

During the Ph.D., he worked as a Visiting Researcher with the Digital Video and Multimedia Laboratory, Columbia University, New York, NY, USA. He is the Director of the Institute of Computer Science, University of St. Gallen, St. Gallen, Switzerland, where he holds the Chair as a Full Professor in artificial intelligence and machine learning (AIML). Before that, he was the Founding Director of the Deep Learning Competence Center, German Research Center for Artificial Intelligence (DFKI), where he was the Principal Investigator of the NVIDIA AI Laboratory Kaiserslautern. He was a Post-Doctoral Researcher with the University of California at Berkeley, Berkeley, CA, USA, and the International Computer Science Institute (ICSI), Berkeley, where he was involved in deep-learning projects at the Lawrence Livermore National Laboratory.



Michael Mommert (Member, IEEE) received the diploma degree in physics from the University of Heidelberg, Heidelberg, Germany, in 2009, and the Ph.D. degree in earth sciences from Free University Berlin, Berlin, Germany, in 2013.

He is currently an Assistant Professor of computer vision at the School of Computer Science, University of St. Gallen, St. Gallen, Switzerland. His interests include the intersection of computer vision and remote sensing with an emphasis on data-efficient deep-learning methods. Before joining the University of St. Gallen in 2020, he conducted research in solar system astronomy at Lowell Observatory and Northern Arizona University, Flagstaff, AZ, USA.