Energy-based Models in Earth Observation: from Generation to Semi-supervised Learning

Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Sébastien Lefèvre

Energy-based Models in Earth Observation: from Generation to Semi-supervised Learning

Javiera Castillo-Navarro, *Student Member, IEEE,* Bertrand Le Saux, *Member, IEEE,* Alexandre Boulch, *Member, IEEE,* and Sébastien Lefèvre, *Senior Member, IEEE*

***Abstract*—Deep learning, together with the availability of large amounts of data, have transformed the way we process Earth observation tasks, like land cover mapping or image registration. Yet, today new models are needed to push further the revolution and enable new possibilities. This work focuses on a recent framework for generative modeling and explore its applicability to Earth observation images. The framework learns an energy-based model to estimate the underlying joint-distribution of the data and the categories, obtaining a neural network that is able to classify and synthesize images. On these two tasks, we show that energy-based models reach comparable or better performances than convolutional networks on various public EO datasets, and that they are naturally adapted to semi-supervised settings, with very few labeled data. Moreover, models of this kind allow us to address high-potential applications such as out-of-distribution analysis and land cover mapping with confidence estimation.**

***Index Terms*—Deep Learning, Energy-based Models, Generative Models, Semi-supervised Learning.**

I. Introduction

EARTH observation (EO) data analysis has become an essential component for global phenomena understanding. In the past years, the large amount of data, available thanks to recent sensors, have made possible the use of deep learning for Earth observation in fields as various as ecology, urban mapping, meteorology or natural disaster response, and will certainly be crucial on the battle against climate change.

However, most of the recent learning-based approaches rely heavily on labeled data. Data annotation for supervised learning remains a challenge, being time consuming and often requiring expert application knowledge. As a consequence, current available EO datasets are partial and provide biased samples of the global Earth land cover, since there is no efficient way to deliver humanly annotated labels for the immensity of EO data available.

On the other hand, the open remote sensing data streams such as Copernicus, provide massive and open imagery data. As of today, most of these data are not used for learning purposes, essentially because proper labels are not procured.

In consequence, EO data analysis should not be confined to supervised learning methods. On the contrary, we should explore less label-dependent approaches to leverage the diversity of (unlabeled) data that is available. One way to exploit data

Fig. 1. Class-conditional samples generated by Joint Energy-based Model (JEM) trained on the EuroSAT dataset. First two columns contain real EuroSAT samples. Third and fourth columns present JEM-generated samples trained on all training samples. Last two columns show samples generated following a semi-supervised learning strategy, with 100 labeled samples per class.

J. Castillo-Navarro is with ONERA, Université Paris-Saclay, Palaiseau, France and with Université Bretagne Sud, IRISA, Vannes, France. e-mail:javiera.castillo_navarro@onera.fr.

B. Le Saux is with European Space Agency, ESRIN, Frascati, Italy.

A. Boulch is with valeo.ai, Paris, France.

S. Lefèvre is with Université Bretagne Sud, IRISA, Vannes, France.

without the necessity of labels is using generative models [1], which aim to model the data distribution. The objective is to get a deeper understanding of data and their intrinsic features. Some applications of generative models are super-resolution, image desoising or image generation. Moreover, they can be adapted to integrate some label information into the learning process and perform semi-supervised learning.

Semi-supervised learning [2] refers to methods that leverage unlabeled data together with (few) labeled samples to learn a given task. In the last years, the interest for the development of semi-supervised methods has been rising, because they are essential for applications where –as for remote sensing– labeled data are hard or costly to obtain. Thus, the development of semi-supervised algorithms is one way to exploit unlabeled data on the learning process and it is one of the main challenges of Earth observation to take part of this very complete, global scale data [3].

This work establishes the potential of joint energy-based models for supervised and semi-supervised learning in EO images. Particularly it shows the interesting properties of EBMs, including a good model calibration which is crucial for prediction confidence estimation and the ability to perform out-of-distribution detection, thanks to the data distribution estimation. Our main contributions are:

- **First energy-based models for Earth Observation**, to the best of our knowledge[1], demonstrating the relevance of such models and paving the way for their future dissemination in the field.
- **Robust classification performances** of supervised and semi-supervised models.
- **Diversified and high-quality data generation** from the learned data distribution.
- **Model calibration** improved with respect to non-EBM models.
- **Out-of-distribution dataset** analysis assessed on several public EO datasets, including confidence estimation for land cover mapping use cases.

This paper is organized as follows: Sec. II discusses some related work, Sec. III recalls energy-based models, in particular, joint energy-based models and their semi-supervised extension. We then report experimental results for several applications in Sec. IV, including semi-supervision, calibration and out-of-distribution analysis. Limitations of the proposed method are discussed in Sec. V. Finally, conclusions and future works are presented in Sec. VI.

II. Related Work

A. Deep Learning in Earth Observation

Processing of EO data has greatly benefited from deep learning techniques in the last decade [5]. Indeed, they currently represent the state-of-the-art in the field: classification, semantic segmentation, change detection, building detection, to name a few, are nowadays tackled using neural networks. After seminal works for road detection [6], generic multi-class segmentation and classification were soon tackled with Convolutional Neural Networks (CNNs) and Fully Convolutional Networks (FCNs) [7], [8], [9], [10], [11], until latest developments which result in global cover maps of a continent or the entire planet [12].

Nowadays, deep learning research in remote sensing has evolved and includes more specific applications, such as interactive learning [13], visual question answering [14], domain adaptation [15], multimodal approaches [16] and semi-supervised learning [3].

B. Deep Generative Models

Generative models [1] comprise a family of techniques which aim to learn the instrinsic data distribution. Their ultimate goal is essentially to get a deeper understanding of data, by learning automatically the natural features of a dataset, its categories or dimensions. They are also useful for many real-life applications like super-resolution, image denoising, inpainting or neural network pre-training.

Current research on deep generative models can be grouped in different categories: Variational Auto-Encoders (VAEs) [17], Generative Adversarial Networks (GANs) [18], Autoregressive models [19], Normalizing flows [20], and Energy-based models [21]. These categories differ on the way they estimate data distribution. Some of them estimate directly the likelihood function or a proxy of it, while others approximate the distribution in an implicit way. This has a direct impact on the trade-off to make between execution time, architecture to use and the objective function to optimize. Usually, learning the distribution implicitly comes with the advantage of getting more realistic and sharper generated images, while the explicit expression of the likelihood function allows for other applications, like out-of-distribution detection.

In the remote sensing community, generative models and more particularly, Generative Adversarial Networks, have been used for different purposes [22], [23], [24], including scene classification [25], [26], [27], yet the flaws of these models are well known: difficulty or instability during training or mode collapse. Moreover, GANs estimate the data distribution implicitly, limiting their applications to synthetic data generation. Other generative models –that estimate the data distribution (or a proxy) in an explicit way– seem more appealing because of their wide range of applications besides image generation, however they have not been investigated in the EO field yet.

C. Energy-based Models

Inspired by statistical physics, Energy-Based Models (EBMs) [21], [28] specify probability densities up to an unknown normalizing constant. This family of models captures dependencies between variables only through a scalar function, known as the *energy function*, and does not place any restriction on the tractability of the normalizing constant. Therefore, they are easy to parameterize and can model a very wide family of probability distributions.

Recent works have shown that combined with the expressive power of deep neural networks, EBMs can model data distributions with impressive results in a wide range of applications,

[1]Please note that a preliminary version of our study, limited to supervised learning on EuroSAT data, has been published in [4].

including image generation, simultaneous generation and classification, class-incremental classification, out-of-distribution detection, etc [29], [30]. However, combination of EBMs and deep learning have been scarcely used in remote sensing, except for [31], and EBMs have never been used for joint classification and generation nor (semi-)supervised learning in this context.

D. Semi-supervised Learning

Semi-supervised learning [2] refers to methods that leverage unlabeled data together with (few) labeled samples to learn a task. The key idea behind semi-supervised learning is to learn a representation function (that maps a data point to its target) from labeled data as in the supervised approach, but using the available unlabeled data to leverage information about structure of these data to help the learning process.

In the last years, the interest for the development of semi-supervised methods has been rising, because they are essential for applications where labeled data are hard or costly to obtain. This is the case of remote sensing imagery, since there is no efficient way to provide humanly annotated labels for the immensity of EO data available [32], [33], [3].

Semi-supervised methods in deep learning developed to date exploit, in general, two principles: the first one is pseudo-labeling [34], where a model is initially trained on the labeled data and is used to make predictions on the unlabeled data, then it selects the examples where the prediction is confident and considers it as a pseudo-label to expand the labeled training set; the second one is consistency regularization that enforces the idea that realistic perturbations of data points should not significantly change the output of the predictor [35], [36], [37], [38].

Current state-of-the-art semi-supervised methods for classification, like MixMatch [39] or FixMatch [40], usually combine these ideas, achieving impressive results. However, they rely heavily on data augmentation, which works well on the image domain, but can be hard to adapt to other use-cases.

On the contrary, the Joint Energy-based Model (JEM) [30] relies on the capacity of generative models to estimate the underlying data distribution and can be easily extended to perform semi-supervised classification [41].

E. Semi-supervised learning in Earth observation

Semi-supervised learning techniques are especially appealing for the remote sensing community, since EO data are naturally well-suited in this context. Indeed, labeled data are hard to obtain, while raw (unlabeled) data are constantly gathered through satellite or aerial missions. Thus, semi-supervised methods are a feasible solution to improve the classification performances and the generalization capacities of our models.

In the last decades, several semi-supervised methods have been proposed for Earth observation data applications. Before the deep learning outbreak, different approaches have been explored, including graph-based methods to integrate unlabeled data into the learning process [42], [43]; use unlabeled examples to achieve manifold alignment of data coming from different modalities [44]; and factor analysis for hyperspectral image classification [45]. More recently, deep semi-supervised learning techniques have emerged, but most of them rely on self-training and pseudo-labeling [32], combining them with other techniques to build more robust models such as the use of an ensemble of CNNs to assign pseudo-labels and prevent error propagation [46]; using cross-modal data [33]; sample selection schemes to train transferable deep models for land use classification [47], [48]; or using stacked auto-encoders and soft-label propagation to tackle the building detection problem [49].

Other strategies that use semi-supervised learning in remote sensing applications include: a center-based discriminative adversarial learning framework for cross-domain land cover classification of aerial images [50]; integrating CNNs and active learning to better use unlabeled samples for hyperspectral image classification [51]; the use of a semi-supervised shallow network, self-organizing map framework, to classify and estimate physical parameters from multispectral and hyperspectral images [52]; using multi-attention and an adaptive kernel for semi-supervised classification of multispectral images [53]; and using multi-task learning regularization to leverage unlabeled data through unsupervised auxiliary tasks [3].

Fewer are the works that exploit generative models to leverage unlabeled samples for training, GANs have been used to extract features from hyperspectral images for semi-supervised classification [54], or jointly with gated attention and a discriminative network for scene classification of aerial images [55]. A modified GAN, with a classifier as discriminator, has been developed to tackle the multispectral scene classification problem [56]. Ours is the first work that focuses on EBMs to model the joint distribution of images and labels and that can be used to address generation, semi-supervised classification, out-of-distribution detection and confidence estimation.

III. Energy-based Models and Semi-supervised Learning

A. Energy-based Models

Energy-based models capture dependencies between variables, $\mathbf{x} \in \mathcal{X}$, through a scalar function $E : \mathcal{X} \to \mathbb{R}$, known as the *energy function*. Learning an EBM consists in finding an energy function that associates low energy values to realistic configurations of x, and higher energy values to unrealistic ones. Then, the energy can be considered as a measure of compatibility of different configurations of variables.

EBMs can be interpreted as normalized probabilistic models using the Gibbs distribution, which expresses the density $p(\mathbf{x})$ as:

$$p(\mathbf{x}) = \frac{\exp(-E(\mathbf{x}))}{Z}, \tag{1}$$

where $Z = \int_{\mathcal{X}} e^{-E(\mathbf{x})}$ is a normalization constant.

Training EBMs comes with the advantage that the energy function parameterizes all the information about inputs. This alleviates the burden of computing or estimating the normalization constant Z, which is often intractable. Therefore,

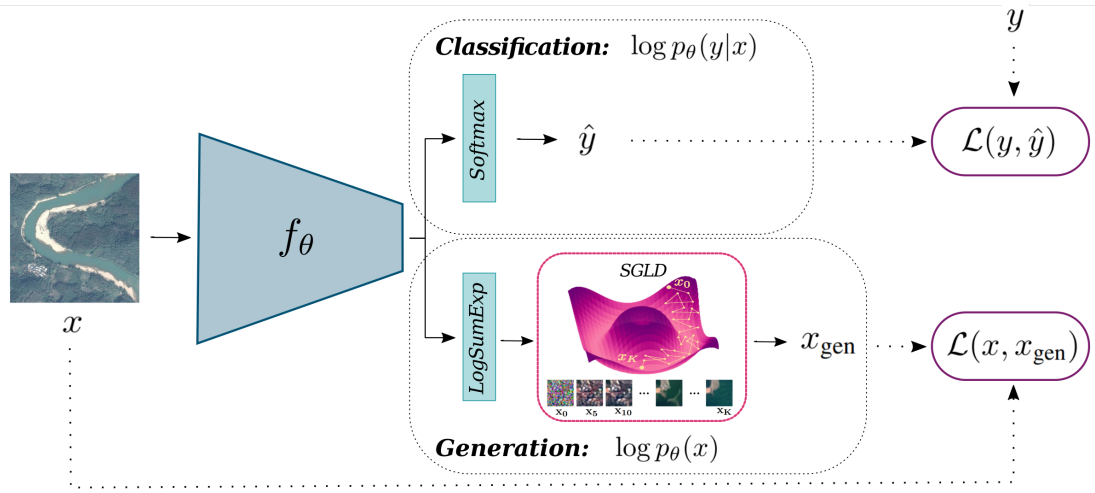

Fig. 2. JEM overview. In a nutshell, an input image x passes through a neural network f_θ. Then, the pipeline splits into two modules: (i) a classification module that applies a softmax function to $f_\theta(\mathbf{x})$ to obtain class scores and computes the classification loss (cross-entropy), and (ii) a generation module that computes the energy E_θ from Eq. (6) (LogSumExp), then runs a finite Stochastic Gradient Langevin Dynamics (SGLD) chain (Eq. (8)), drawing samples from $p_\theta(\mathbf{x})$ and uses them to compute the log-likelihood loss. The sum of both loss terms (Eq. (7)) is then optimized by backpropagation.

EBMs provide much more flexibility in the design –and thus the expressiveness– of learning models.

In this regard, EBMs have recently benefited from the expressive power of deep neural networks to model complex energy functions, with impressive results in generation, hybrid generation-classification and other applications [30], [29].

The standard way of learning EBMs with deep learning today is by maximum likelihood training. Let p_θ be the probability density of an EBM, whose energy function, E_θ, is parameterized by a neural network of parameters θ. The density of the model, $p_\theta(\mathbf{x})$, can be fit to the distribution of data, $p_{\text{data}}(\mathbf{x})$, by maximizing the expected log-likelihood function over the data distribution:

$$\begin{aligned}\mathcal{L}_{\text{ML}}(\theta) :=& \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log p_\theta(\mathbf{x})] \\ =& \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[-E_\theta(\mathbf{x})] - \log Z_\theta\end{aligned} \tag{2}$$

The gradient of the log-likelihood can be expressed as:

$$\nabla_\theta \mathcal{L}_{\text{ML}}(\theta) = \mathbb{E}_{p_\theta(\tilde{\mathbf{x}})}[\nabla_\theta E_\theta(\tilde{\mathbf{x}})] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x})] \tag{3}$$

To compute the gradient expressed in Eq. (3), one needs to be able to sample from the model distribution p_θ, which is not possible. Current approaches approximate p_θ using MCMC methods, like Langevin dynamics [57]. This allows us to approximately optimize the log-likelihood objective and generate samples from the model.

B. Joint Classification and Generation

Joint energy-based models (JEM) [30] have been recently presented to extend a standard classification neural network into an hybrid discriminative-generative model, by simply re-interpreting the outputs of the classifier.

Let $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$ be a classification neural network, parameterized by θ, with K the number of classes and D the input's dimension. The fundamental idea of JEM is to express the joint distribution of images ($\mathbf{x}$) and labels (y) as a joint energy-based model:

$$p_\theta(\mathbf{x}, y) = \frac{\exp(f_\theta(\mathbf{x})[y])}{Z_\theta}, \tag{4}$$

where the joint-energy function is parameterized by the neural network: $E_\theta(\mathbf{x}, y) = -f_\theta(\mathbf{x})[y]$. $f_\theta(\mathbf{x})[y]$ is the y-th entry of $f_\theta(\mathbf{x})$ and Z_θ the normalizing constant of the model.

By marginalizing Eq. (4) above, we obtain the distribution $p_\theta(\mathbf{x})$ expressed as:

$$p_\theta(\mathbf{x}) = \sum_{y=1}^{K} p_\theta(\mathbf{x}, y) = \frac{\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y])}{Z_\theta}. \tag{5}$$

From Eq. (5), one may observe that the distribution $p_\theta(\mathbf{x})$ is also an energy-based model, with the energy given by:

$$E_\theta(\mathbf{x}) = -\log\left(\sum_{y=1}^{K} \exp(f_\theta(\mathbf{x})[y])\right). \tag{6}$$

The JEM model is then trained to maximize the joint log-likelihood, $\log p_\theta(\mathbf{x}, y)$, which can be factorized as:

$$\log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y|\mathbf{x}) \tag{7}$$

As shown below, Eq. (7) is the key to obtain a joint discriminative-generative model.

a) Generation: The first term in Eq. (7), $\log p_\theta(\mathbf{x})$, corresponds to the generative part of the model. It is trained as an energy-based model by approximating the gradient $\nabla_\theta \mathcal{L}_{\text{ML}}(\theta)$ (Eq. (3)) using a sampler based on Stochastic Gradient Langevin Dynamics (SGLD) [29] and thus, generates samples following:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \tfrac{\alpha}{2}\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_i) + \varepsilon, \quad \mathbf{x}_0 \sim p_0(\mathbf{x}), \tag{8}$$

with $\varepsilon \sim \mathcal{N}(0, \alpha)$ and $p_0(\mathbf{x})$ usually a uniform distribution, and α a step-size following a polynomial decaying.

b) Classification: The second term is related to $p_\theta(y|\mathbf{x})$, which written as $p_\theta(y|\mathbf{x}) = p_\theta(\mathbf{x}, y) \ / \ p_\theta(\mathbf{x})$ matches the softmax output of a usual classifier. Thus it can be simply optimized using the cross-entropy loss, as a standard classification neural network.

Figure 2 illustrates how JEM works in practice. An input image x passes through a neural network f_θ, which outputs $f_\theta(\mathbf{x}) \in \mathbb{R}^K$. Then, the pipeline splits into two modules: (i) a classification module (Fig. 2 top) that applies a softmax function to $f_\theta(\mathbf{x})$ to obtain class scores and computes the classification loss (cross-entropy), and (ii) a generation module (Fig. 2 bottom) that computes the energy E_θ from Eq. (6) (LogSumExp), then runs a finite SGLD chain (Eq. (8)), drawing samples from $p_\theta(\mathbf{x})$ and uses them to compute the log-likelihood loss. The sum of both loss terms (Eq. (7)) is then optimized by backpropagation.

C. Semi-supervised Learning with JEM

Moreover, JEM, as described above, also allows to extend a conventional classifier to semi-supervised learning in a very natural way [41].

Indeed, if labels are available, one can optimize the main objective $\log p_\theta(\mathbf{x}, y)$ as in Eq. (7), otherwise one may simply marginalize it out and optimize $\log p_\theta(\mathbf{x})$ only. In practice, following the scheme in Fig. 2, this means that for labeled samples the network is updated as decribed above (Sec. III-B), but unlabeled samples only go through the generation module (bottom section Fig. 2) to update the network.

We have recalled here the main concepts of JEM, a recent energy-based model for joint generation and classification of images. However, to the best of our knowledge, the relevance of such energy-based models to deal with EO data has not been studied yet. We report in the next section some experiments we conducted with JEM to address various applications of high interest in remote sensing.

IV. Experiments

Since JEM is a multifaceted model, in this section we explore its capacities in various tasks, including: classification, generation, semi-supervised classification, out-of-distribution detection and land cover mapping. In Table I, we compare JEM to other models that perform well on each task, however none of them is as versatile as JEM, being limited to one or two tasks to perform simultaneously.

We perform experiments training our models on several publicly available EO datasets for scene classification: the EuroSAT Dataset [58], the So2Sat LCZ42 Dataset [59], the Aerial Image Dataset [60] and the UCMerced Dataset [61].

The **EuroSAT Dataset** comprises patches from Sentinel-2 images over 34 countries in Europe. Each patch is labeled with one of 10 land cover/land use classes (e.g. industrial, residential, highway, pasture, forest, etc.). Classes are well-balanced, with 2,000 to 3,000 examples per class, 80% of which are used for training. We use the EuroSAT RGB version.

The **So2Sat LCZ42 Dataset** is composed of Sentinel-1 and Sentinel-2 image patches over 42 locations over the globe. Patches are labeled according to the Local Climate Zones scheme (LCZ), with 17 categories. It is worth to mention that the training set and testing set are geographically independent, containing images from different locations. This makes this dataset particularly difficult, because models need to be sufficiently robust to generalize well on the test data. For our experiments, we only make use of the RGB Sentinel-2 bands (B04, B03, B02), as in EuroSAT.

Model	Classification	Generation	Semi-supervision	OOD detection
Wide-ResNet	✓	✗	✗	✗
VAE	✗	✓	✗	✗
GAN	✗	✓	✗	✗
BerundaNet	✓	✗	✓	✗
FixMatch	✓	✗	✓	✗
JEM	✓	✓	✓	✓

TABLE I
MODELS COMPARISON. JEM IS THE ONLY MODEL ABLE TO PERFORM ALL THESE TASKS SIMULTANEOUSLY.

The **Aerial Image Dataset (AID)** consists of 10,000 optical aerial images from different countries around the world, labeled within 30 scene classes. Original RGB tiles are of size 600px $\times$ 600px. Due to the computing time of JEM, we have resized them to 64px $\times$ 64px during training.

The **UCMerced Dataset** is a small-size dataset and has been widely used for the evaluation of aerial scene classification. It contains 2,100 aerial ortho-images from different regions of USA. Each image is labeled with one of the 21 land use classes. Original 256px $\times$ 256px tiles have been resized to 64px $\times$ 64px for JEM training.

For evaluation in out-of-distribution detection and other tasks of interest, we use in addition several public EO datasets: ISPRS Potsdam [62], OSCD dataset [63], DFC2017 [64] and BigEarthNet [65].

Implementation details. Following [30], we perform our experiments using a Wide-ResNet-28-10 architecture [66], with no batch normalization. We train our networks with the Adam optimizer [67], during 200 epochs, following the JEM training scheme.

Moreover, we adopt a hold-out evaluation method, defining a training and a test set (80% and 20% of data, respectively, for all datasets, except So2Sat LCZ42 where train and test partitions are already defined). Additionally, 10% of the training set was used as validation partition during training. This is especially important when training on very few labeled data to adopt an early stopping strategy and avoid overfitting.

Pytorch [68] is used for all implementations.

A. Joint Classification and Generation with JEM

In this section we show that this new training paradigm allows to get an hybrid model, with competitive performances in both tasks, classification and generation.

Wide-ResNet is trained as a usual classifier (with cross-entropy loss), while JEM is trained as described in Section III-B. We compare the generative performance with a standard VAE [69]. Results on the EuroSAT dataset are summarized in Table II.

Type	Model	Classification Accuracy (↑)	Generation FID (↓)	Generation KID (↓)
Discriminative	Wide-ResNet	**97.56**± 0.52 %	✗	✗
Hybrid	JEM	97.42± 0.19 %	**122.1**	**0.06**
Generative	VAE	✗	215.4	0.14

TABLE II
CLASSIFICATION AND GENERATION SCORES OF MODELS TRAINED ON EUROSAT. COMPARISON OF JEM WITH RESPECT TO A PURELY SUPERVISED MODEL (WIDE-RESNET-28-10) AND A PURELY GENERATIVE MODEL (VAE). NOTE THAT JEM IS THE ONLY MODEL THAT CAN PROVIDE BOTH CLASSIFICATION AND GENERATION SCORES. BEST SCORES IN BOLD.

Given uncertainty measured by standard deviation, JEM results reach the same level of performances as classification-only Wide-ResNet and previous reported classification results on EuroSAT, namely ResNet-50 and GoogLeNet with 98.6% and 98.2% of overall accuracy respectively [58]. The small difference observed might be explained by the intrinsic regularization of the multi-task JEM model. Furthermore, [58] does not specify a training and test partition, which might also explain the discrepancy with our results. In terms of generation, we rely on the Fréchet Inception Distance (FID) [70] and the Kernel Inception Distance (KID) [71] to evaluate the quality of generated samples. According to these metrics, JEM generated samples are superior to VAE samples.

Fig. 1 shows some class-conditional examples generated by the network after being trained on the EuroSAT dataset, with different settings. Each row represents a class in the dataset. First two columns show real samples from the dataset, third and fourth columns present JEM-generated samples trained on the whole EuroSAT dataset and last two columns display JEM-generated samples with the model trained in a semi-supervised manner with 100 labeled samples per class (and the rest of the dataset as unlabeled data, more details in Sec. IV-B). From these examples, we observe that JEM captures the data distribution properly, since generated samples are extremely similar to real EuroSAT samples regardless of the fraction of annotated examples available for training. Moreover, the model is capable to produce samples for every class on the dataset, with a large variety of images per class.

However, some classes remain difficult to apprehend, e.g. forests or sea and lakes. This might be due to the lack of texture on these images. Industrial buildings (first row in Fig. 1) would require finer and more rectangular outlines to correctly match industrial buildings in the EuroSAT dataset. On the other hand, JEM is able to handle impressively images of highways, rivers and different types of fields. Indeed, generated samples of these classes are remarkably similar to real images. This shows that the model is able to learn the true distribution behind the dataset and leads to compelling applications. Synthetic examples generated from the learned data distribution may be used for simulation or even for training new models.

B. Semi-supervised classification with JEM

In this section we perform semi-supervised classification and show the potential of JEM in extreme settings when very few labeled samples are available.

We train the JEM model with a small subset of labeled examples and the entire dataset as unlabeled data, following the approach described in Section III-C. We vary the number of labeled samples per class on which the model is trained, and compare our results with three baselines: the fully supervised Wide-ResNet, BerundaNet [3], a semi-supervised method based on multi-task learning, and FixMatch [40], which is currently the state-of-the-art algorithm for semi-supervised classification in computer vision. Wide-ResNet, as a supervised method, is trained on labeled data only, while BerundaNet, FixMatch and JEM are trained similarly on the whole dataset, using labels when available. Table III summarizes our results on the EuroSAT dataset.

First row in Table III presents completely supervised results on the entire training set, as an upper bound for the semi-supervised strategies. We observe that all methods are on par in terms of performance, FixMatch being slightly better. We observe from the following rows that FixMatch, being especially designed to tackle the semi-supervised classification problem, is superior to all methods and performs remarkably well, even in extreme situations when very few labeled data is available. In the case of BerundaNet and JEM, there is a point where they perform considerably better than Wide-ResNet. In the case of JEM, there is no significant difference with respect to the Wide-ResNet performance in the 5% and 100% labeled samples per class regime, however the advantage of JEM becomes tangible as soon as the model is trained with 1% of labeled samples or less, with a performance gap varying from 6.2% to 10.7% of accuracy. Moreover, in the semi-supervised setting, JEM is always superior to BerundaNet. This difference might be explained by the way these methods leverage unlabeled samples during training. Indeed, BerundaNet uses them to compute a regularization term through a secondary task (reconstruction), while JEM uses unlabeled samples to estimate their underlying distribution, which might contain valuable information for classification.

These results show that: first, the energy function can be learned from unlabeled data as well as labeled data; and second, if the image distribution $p_\theta(\mathbf{x})$ is well estimated, it is easier then to estimate the conditional distribution $p_\theta(y|\mathbf{x})$ from a small set of annotated training samples.

Furthermore, even if FixMatch has undeniably better performances in the semi-supervised settings, it is worth to notice that it was especially designed to perform semi-supervised classification, and cannot perform other tasks like OOD detection nor generation (see Table I). On the contrary, JEM is a versatile model that can perform several tasks simultaneously. Moreover, it could be optimized to achieve better results on semi-supervised learning, for instance, by integrating FixMatch features such as massive data augmentation strategies and consistency regularization.

Additionally, we compare JEM and Wide-ResNet on two well-known benchmarks for scene classification, AID and UCMerced. Results are summarized on Table IV. They confirm what we observed previously on EuroSAT: in the supervised setting, when all labels are available for the training data, Wide-ResNet is slightly superior to JEM, because of

Labeled samples/class	% of labels	Wide-ResNet	BerundaNet [3]	FixMatch [40]	JEM
2000 on avg.	100%	97.56 ±0.52	96.90 ±0.67	**98.81** ±0.06	97.42 ±0.19
100	∼ 5%	86.36 ±0.26	74.78 ±2.01	**97.83** ±0.12	86.23 ±0.80
20	∼ 1%	62.93 ±1.01	54.25 ±2.41	**95.78** ±0.99	69.11 ±1.18
10	∼ 0.5%	52.33 ±1.59	46.84 ±1.83	**94.95** ±1.12	61.60 ±1.49
5	∼ 0.25%	43.83 ±3.18	39.80 ±1.51	**94.45** ±1.29	54.79 ±3.55
1	∼ 0.05%	28.02 ±0.97	32.77 ±1.05	**67.46** ±4.67	36.86 ±1.11

TABLE III

CLASSIFICATION RESULTS ON EUROSAT (ACCURACY [%] ↑). COMPARISON WITH A PURELY SUPERVISED METHOD (WIDE-RESNET), A MULTI-TASK SEMI-SUPERVISED NETWORK (BERUNDANET) AND A PURELY SEMI-SUPERVISED METHOD (FIXMATCH), TRAINED ON THE SAME NUMBER OF LABELED SAMPLES. GREY CELLS INDICATE MODEL LEVERAGING UNLABELED DATA. BEST SCORES IN BOLD, SECOND BEST UNDERLINED.

Dataset	Labeled samples/class	% of labels	Wide-ResNet	JEM
So2Sat	∼ 20000	100%	50.93 ± 0.16	**54.60** ±0.35
	1000	∼ 5%	44.17 ±0.40	**48.59** ±0.58
	200	∼ 1%	35.45 ±0.17	**42.43** ±0.47
	100	∼ 0.5%	30.90 ±0.35	**38.71** ±0.64
AID	∼ 300	100%	**78.71** ±0.08	74.11 ±0.24
	20	∼ 7%	41.07 ±1.87	**50.23** ±0.69
	13	∼ 5%	34.46 ±0.59	**44.49** ±0.65
	3	∼ 1%	17.38 ±0.32	**25.68** ±0.65
	1	∼ 0.5%	9.98 ±0.36	**16.21** ±0.58
UCMerced	80	100%	81.71 ±0.72	80.49 ±1.67
	10	∼ 12.5%	45.41 ±0.43	**48.91** ±0.42
	4	∼ 5%	26.99 ±1.24	**34.16** ±1.78
	1	∼ 1%	14.34 ±1.88	**24.31** ±1.87

TABLE IV

CLASSIFICATION RESULTS ON DIFFERENT EO DATASETS (ACCURACY [%] ↑). GREY CELLS INDICATE MODEL LEVERAGING UNLABELED DATA. BEST SCORES IN BOLD.

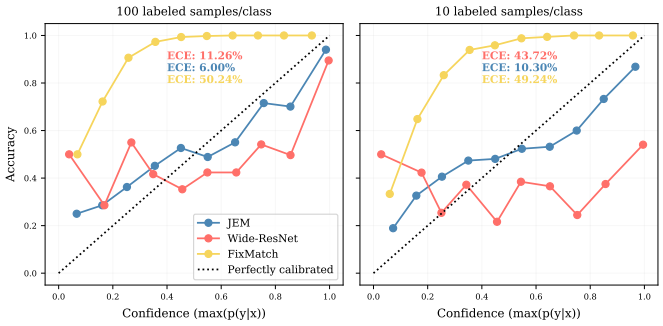

Fig. 3. Calibration curves for supervised Wide-ResNet and semi-supervised JEM and FixMatch trained on EuroSAT dataset. Left: trained on 100 labeled samples per class. Right: trained on 10 labeled samples per class. ECE: Expected Calibration Error (↓).

the intrinsic regularization of the latter. However, when few labels are available, JEM has considerably better classification performance.

Finally, we also perform experiments on the more realistic, large-scale So2Sat dataset. Table IV summarizes the results. Not only they confirm the tendency observed on EuroSAT data, but the superiority of the JEM model over Wide-ResNet is even more consistent, including the supervised setting. This is explained by the existing domain gap between training data and testing data in So2Sat, due to different geographic locations. Indeed, standard discriminative classifiers, like Wide-ResNet, are prone to lack robustness to distribution shifts. However, learning the underlying distribution of the data by a generative model such as JEM helps to overcome this issue and sets a starting point to bridge the performance gap when dealing with domain shifts.

Model Calibration

Beyond classification scores, an important and desirable feature of models is the calibration. A model is said to be calibrated if its output confidence, usually measured as $\max_y p(y|\mathbf{x})$), coincides with its expected accuracy. Therefore, a calibrated model is more informative, being able to provide the uncertainty associated to a prediction.

We thus evaluate and compare the calibration of our supervised (Wide-ResNet) and semi-supervised models (FixMatch and JEM). In particular, we study the 100-labeled-samples-per-class and the 10-labeled-samples-per-class settings. Figure 3 shows the calibration curves for both experiments. A perfectly calibrated classifier should match the straight line $y = x$. We can observe that, in both settings, JEM is the model with best calibration, FixMatch being very underconfident and Wide-ResNet being overconfident.

We quantitatively verify this by computing a usual metric for calibration: the Expected Calibration Error [72] (ECE) score, for both settings. The obtained ECE scores are 11.26%, 50.24% and 6.00% for Wide-ResNet, FixMatch and JEM, respectively, in the case of 100-labeled-samples-per-class; and 43.73%, 49.24% and 10.22% in the extreme setting of 10-labeled-samples-per-class. Since a perfect ECE is equal to zero, these scores confirm that the semi-supervised JEM model is better calibrated, the difference being flagrant in extreme conditions (very few labels). FixMatch exhibits very poor calibration properties. Therefore, unlabeled data regularization, by learning the data distribution, comes with the advantage of allowing for more informative predictions.

C. Out-of-distribution Analysis

Out-of-distribution detection (OOD) refers to the task of recognizing significantly different or anomalous examples, with respect to the ones seen during training. Asserting the capacity of a model to correctly classify a sample from a new domain is a very important and desirable feature, especially in applications which involve real-world decisions.

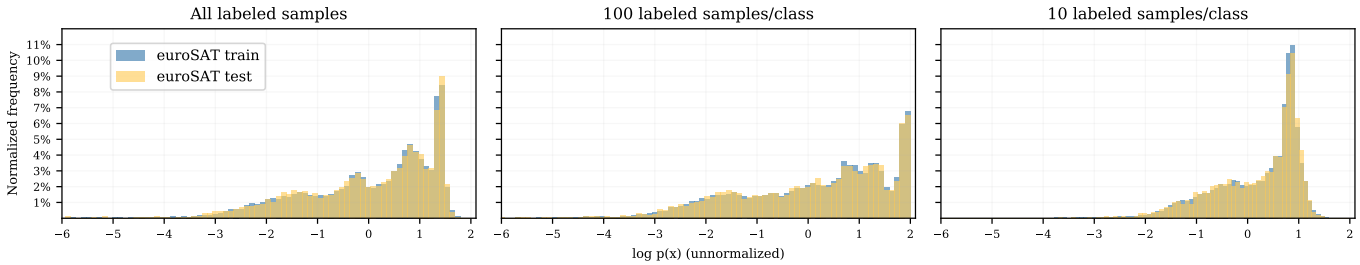

Fig. 4. JEM log-likelihood (unnormalized) histograms for EuroSAT dataset. Stability of the estimated energy function. Supervised vs. Semi-supervised with 100 labeled samples and 10 labeled samples per class comparison. We observe that the values of the unnormalized log-likelihood are comparable, regardless the amount of labeled data available during training.

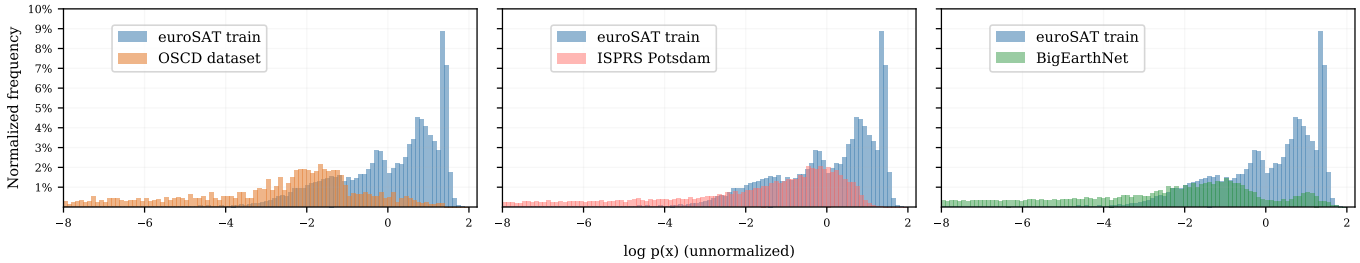

Fig. 5. Out-of-Distribution detection on different public EO datasets. Unnormalized log-likelihood values computed through the supervised model.

In this section, we assess the capacity of our model to assess global out-of-distribution analysis, i.e. if an entire dataset can be considered *in-distribution* with respect to the learned distribution. In this regard, we compare the histograms of the unnormalized log-likelihood (i.e. $-E(\mathbf{x})$) values of the EuroSAT training set with the obtained histograms for different public datasets.

1) Supervised vs. Semi-supervised Energy Function: Figure 4 presents the unnormalized log-likelihood histograms for the EuroSAT dataset in the fully-supervised JEM setting (left) and two semi-supervised settings: training with 100 labeled samples per class (center) and with 10 labeled samples per class (right). In all cases, the histogram profiles of the training and test partition match perfectly, which means that, as expected, there is no shift of the estimated distribution from EuroSAT train to EuroSAT test.

Moreover, we observe that the log-likelihood distribution estimated by the models is very similar, showing that the energy is not linked to the labels, but to the data.

2) Comparing Datasets: On the other hand, Figure 5 shows the unnormalized log-likelihood histograms of 3 public EO datasets: OSCD dataset, ISPRS Potsdam and BigEarthNet, obtained after training the model on the whole EuroSAT training set.

We observe that for these datasets, the histogram profile does not exactly match the one of the EuroSAT training data. Actually, values of the unnormalized $\log p(\mathbf{x})$ can be extremely small, which can be interpreted as the samples from these datasets are not likely to come from the distribution learnt from EuroSAT. We can confirm this observation by computing the Kullback-Leibler (KL) divergence with respect to the distribution of the EuroSAT train histogram. Indeed, while KL value for EuroSAT test data is 0.27; the other datasets KL values are 28.2, 25.6 and 26.3 for Potsdam, OSCD and BigEarthNet, respectively. In view of this, more information would be needed for the model to correctly represent those datasets that differ on location, resolution or appearance.

Finally, it is interesting to notice that the distribution that differs the most is ISPRS Potsdam, the only dataset with a different resolution. This might imply that resolution is an important factor for domain adaptation.

D. Application to Land Cover Mapping

Land cover mapping is an interesting application of JEM on new unseen domains as detailed in the following sections.

1) Patch-wise classification: We apply our EuroSAT-trained models –including Wide-ResNet, supervised and semi-supervised JEM– to unseen OSCD tiles. To do so, the tiles are split into 64×64 patches which go through the already trained network to obtain the corresponding class per patch, leading to a patch-wise classification map.

We observe in Figure 6 the results on two locations from OSCD: Beirut and Rio de Janeiro. The maps produced by the classifier are, in general, globally correct and retrieve various densities of urban and green areas. As expected, the quality of predictions deteriorates as the number of labeled samples decreases. Indeed, supervised Wide-ResNet and JEM predictions are both plausible land cover maps for these locations. The map of JEM semisup-100 is still trustworthy, while 10 labeled samples per class seem not enough to train an accurate model.

Similarly, we apply the So2Sat LZC42-trained models to the unseen tile of Rome from DFC2017. Since So2Sat LZC42 is composed of 32×32 images, the Rome tile is also split in 32×32 patches to pass through the network. Figure 7 presents

our patch-wise classification maps. As before, the maps are reasonable, JEM being more accurate than Wide-ResNet to recognize low plants, where Wide-ResNet over estimate heavy industry.

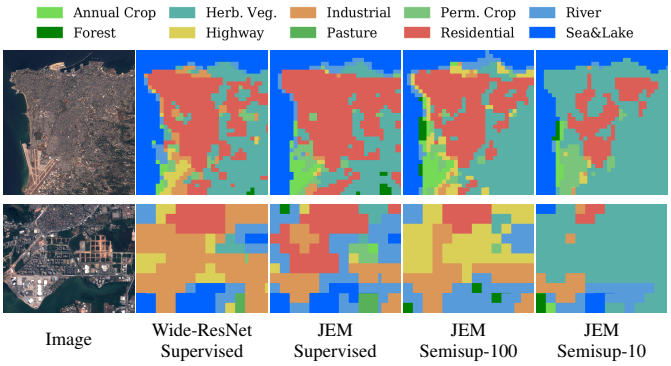

Fig. 6. Semantic maps on never-seen OSCD cities. Top: Beirut. Bottom: Rio de Janeiro. Supervised indicates models trained on the entire EuroSAT dataset. Semisup-x is JEM trained with a semi-supervised strategy with x labeled samples per class. No semantic segmentation ground truth is provided with this dataset.

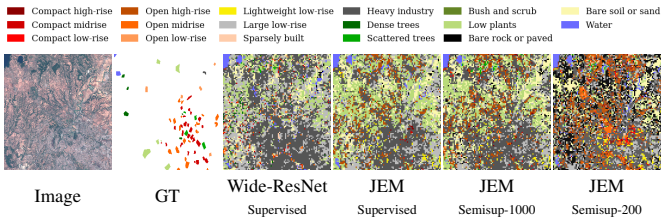

Fig. 7. Semantic maps on never-seen DFC 2017 tile of Rome. Supervised indicates models trained on the entire So2Sat dataset. Semisup-x is JEM trained with a semi-supervised strategy with x labeled samples per class.

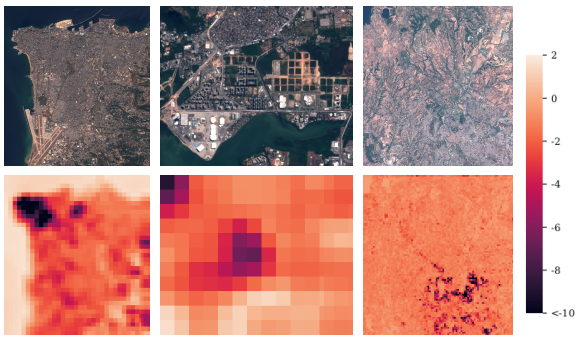

Fig. 8. Confidence maps obtained by JEM on never-seen OSCD and DFC2017 tiles. Confidence is measured as the unnormalized $\log p(\mathbf{x})$), From left to right: Beirut, Rio de Janeiro and Rome.

2) Confidence maps: The major advantage of JEM over a standard classifier such as Wide-ResNet is its capacity to estimate the underlying data distribution through the energy function. We can use the unnormalized log-likelihood value as a proxy for the confidence of the model's prediction. Indeed, if the model assigns a high value of log-likelihood to an image it could be considered as *in-distribution*, and thus the model's prediction should be pertinent. Conversely, if the model's log-likelihood on a sample is low, we could consider it as *out-of-distribution* and be more cautious with respect to its prediction.

Figure 8 shows the confidence maps obtained by the supervised JEM over the OSCD tiles (trained on EuroSAT) and over Rome tile from DFC2017 (trained on So2Sat). We observe that the confidence of the model varies accross the patches. Indeed, on OSCD, the model is more confident on scenes representing water or fields, while it is considerably less confident in residential and industrial areas, which are more likely to be different from training European cities from the EuroSAT dataset. In the case of Rome, the model is less confident in general, and in particular on the compact zones (according to the ground-truth in Fig. 7).

V. Limitations

Training Energy-based Models by maximum likelihood can be very challenging. Indeed, the gradient estimators used to estimate log-likelihood are considerably unstable and prone to diverging during training, this is why hyperparameters must be chosen carefully. Moreover, MCMC-like iterative sampling increases training time linearly with the image size. This may be prohibitive when dealing with large images, which is likely the case in remote sensing applications. This is why we decided to resize AID and UCMerced images for our experiments in Sec. IV-B.

Despite these limitations, we strongly believe that the remote sensing community might deeply benefit from the multiple applications of EBMs, that we tried to bring forward in this article. We believe that there is still much progress to make to improve and optimize EBMs' training, just as the community has achieved great progress on GANs' training in only a few years.

VI. Conclusions

We have considered a recent framework to train neural networks to jointly perform classification and generation of images and applied it to remote sensing data. By re-interpreting the outputs of a classification neural network, the Joint Energy-based Model (JEM) expresses the joint distribution of image-label pairs as an energy-based model. In practice, it allows us to train a robust classifier and estimate the underlying distribution of data, simultaneously. Moreover, this hybrid model is well suited and extends naturally to perform semi-supervised learning.

This seminal application of JEM to EO data led to several important conclusions. First, in small-scale datasets like EuroSAT, we observe that JEM is a strong classifier with performance on par to state-of-the-art methods. More interestingly, in the semi-supervised setting when very few labeled examples are available, JEM is superior to a standard supervised network, both in terms of classification scores and robustness (i.e. better calibrated). Second, with more realistic, large-scale datasets like So2Sat, JEM exhibits outstanding generalization properties, with better performance than usual classifiers in the supervised and semi-supervised settings. However, future work could focus on the integration in JEM of FixMatch mechanisms especially designed for semi-supervised learning, namely data augmentation techniques, pseudo-labeling or consistency regularization strategies. The challenge lies in

realistically augmenting the data, and the distribution estimate given by JEM could be an asset here.

Finally, we have also demonstrated that JEM is able to correctly estimate the data distribution, allowing us to generate faithful and diverse images. Estimating the data distribution enables the model to detect out-of-distribution samples and thus to decide if it can be reliably used in a new domain. This gives JEM the ability to classify unseen zones with a confidence map based on the log-likelihood estimated by the model.

In summary, we have shown through our experiments several appealing applications in remote sensing for this kind of hybrid discriminative-generative model, such as semi-supervised learning, out-of-distribution detection or the generation of synthetic realistic new data. It is a starting point to pave the way to tomorrow's real-life applications.

ACKNOWLEDGMENT

Javiera Castillo-Navarro's work is partially funded by a grant from CNES (Centre National d'Études Spatiales).

REFERENCES

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive Models," *arXiv preprint arXiv:2103.04922*, 2021.

[2] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*. The MIT Press, 2006.

[3] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in Earth observation: the MiniFrance suite, dataset analysis and multi-task network study," *Machine Learning*, pp. 1–36, 2021.

[4] J. Castillo-Navarro, B. Le Saux, A. Boulch, and S. Lefèvre, "Classification and generation of Earth observation images using a joint energy-based model," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2021.

[5] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[6] V. Mnih and G. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[7] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

[8] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest–Part A: 2-D Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, 2016.

[9] N. Audebert, B. Le Saux, and S. Lefevre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[10] N. Rey, M. Volpi, S. Joost, and D. Tuia, "Detecting animals in African Savanna with UAVs and the crowds," *Remote Sensing of Environment*, vol. 200, pp. 341–351, 2017.

[11] N. Audebert, A. Boulch, H. Randrianarivo, B. Le Saux, M. Ferecatu, S. Lefèvre, and R. Marlet, "Deep learning for urban remote sensing," in *2017 Joint Urban Remote Sensing Event (JURSE)*, 2017.

[12] M. Demuzere, B. Bechtel, A. Middel, and G. Mills, "Mapping Europe into Local Climate Zones," *PLOS ONE*, vol. 14, no. 4, pp. 1–27, 04 2019.

[13] G. Lenczner, A. Chan-Hon-Tong, N. Luminari, B. Le Saux, and G. Le Besnerais, "Interactive learning for semantic segmentation in Earth observation," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) MACLEAN Workshop*, 2020.

[14] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[15] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "SEMI2I: Semantically consistent image-to-image translation for domain adaptation of remote sensing data," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020.

[16] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[17] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learnining*, vol. 12, no. 4, pp. 307–392, 2019.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[19] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[20] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2021.

[21] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, 2006.

[22] N. Merkle, S. Auer, R. Müller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1811–1820, 2018.

[23] N. Audebert, B. Le Saux, and S. Lefèvre, "Generative adversarial networks for realistic synthesis of hyperspectral samples," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[24] L. Courtrai, M.-T. Pham, and S. Lefèvre, "Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks," *Remote Sensing*, vol. 12, no. 19, p. 3152, 2020.

[25] S. Xu, X. Mu, D. Chai, and X. Zhang, "Remote sensing image scene classification based on generative adversarial networks," *Remote Sensing Letters*, vol. 9, no. 7, pp. 617–626, 2018.

[26] D. Ma, P. Tang, and L. Zhao, "SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1046–1050, 2019.

[27] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 519–531, 2019.

[28] Y. Song and D. P. Kingma, "How to train your energy-based models," *arXiv preprint arXiv:2101.03288*, 2021.

[29] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3608–3618.

[30] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[31] L. Mou, X. Zhu, M. Vakalopoulou, K. Karantzalos, N. Paragios, B. Le Saux, G. Moser, and D. Tuia, "Multitemporal very high resolution from space: Outcome of the 2016 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3435–3447, 2017.

[32] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 23–43, 2018.

[33] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 12–23, 2020.

[34] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of the*

International Conference on Machine Learning (ICML), Workshop on Challenges in Representation Learning, vol. 3, no. 2, 2013.

[35] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[36] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[37] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[38] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[39] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[41] S. Zhao, J.-H. Jacobsen, and W. Grathwohl, "Joint energy-based models for semi-supervised classification," in *Proccedings of the International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[42] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044–3054, 2007.

[43] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 35–49, 2019.

[44] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7708–7720, 2014.

[45] B. Zhao, J. R. Sveinsson, M. O. Ulfarsson, and J. Chanussot, "Semi-supervised mixtures of factor analyzers and deep mixtures of factor analyzers dimensionality reduction algorithms for hyperspectral images classification," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2019.

[46] R. Fan, R. Feng, L. Wang, J. Yan, and X. Zhang, "Semi-MCNN: A semisupervised multi-CNN ensemble learning method for urban land cover classification using submeter HRRS images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4973–4987, 2020.

[47] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[48] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Weakly supervised change detection using guided anisotropic diffusion," *Machine Learning*, 2021.

[49] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sensing*, vol. 13, no. 3, p. 371, 2021.

[50] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 155, pp. 72–89, 2019.

[51] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4604–4616, 2020.

[52] F. M. Riese, S. Keller, and S. Hinz, "Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data," *Remote Sensing*, vol. 12, no. 1, p. 7, 2020.

[53] K. Zhang and H. Yang, "Semi-supervised multi-spectral land cover classification with multi-attention and adaptive kernel," in *Proceedings of the International Conference on Image Processing (ICIP)*. IEEE, 2020.

[54] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 212–216, 2017.

[55] D. Guo, Y. Xia, and X. Luo, "GAN-based semisupervised scene classification of remote sensing image," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[56] S. Roy, E. Sangineto, N. Sebe, and B. Demir, "Semantic-fusion GANs for semi-supervised satellite image classification," in *Proceedings of the International Conference on Image Processing (ICIP)*. IEEE, 2018.

[57] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proccedings of the International Conference on Machine Learning (ICML)*, 2011.

[58] P. Helber, B. Bischke, A. Dengel, and 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

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

[60] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[61] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[62] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The ISPRS benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, p. 3, 2012.

[63] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[64] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, C. Debes, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1363–1377, 2018.

[65] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[66] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic pptimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[69] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[71] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[72] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Javiera Castillo-Navarro received her M.Eng. in 2017 from École CentraleSupélec, France, and from Universidad de Chile, Chile, her M.Sc. degree from Université Paris-Saclay, France. Since 2019, she is pursuing her Ph.D. degree from the Université Bretagne Sud, France, in collaboration with the Office National d'Études et Recherches Aérospatiales, Université Paris-Saclay, Palaiseau, France. Her research at-present focuses on deep learning for scene understanding and Earth observation applications.

Bertrand Le Saux (Member, IEEE) received the Ms.Eng. and M.Sc. degrees from INP, Grenoble, France, in 1999, the Ph.D. degree from the University of Versailles/Inria, Versailles, France, in 2003, and the Dr. Habil. degree from the University of Paris-Saclay, Saclay, France, in 2019. He is a Senior Scientist with the European Space Agency/European Space Research Institute Φ-lab in Frascati, Italy. His research interest aims at visual understanding of the environment by data-driven techniques including Artificial Intelligence and (Quantum) Machine Learning. He is interested in tackling practical problems that arise in Earth observation, to bring solutions to current environment and population challenges. Dr. Le Saux is an Associate Editor of the Geoscience and Remote Sensing Letters. He was Co-Chair (2015–2017) and chair (2017–2019) for the IEEE GRSS Technical Committee on Image Analysis and Data Fusion.

Alexandre Boulch received the Ms.Eng. from Ecole Polytechnique, Paris, France, M.Sc. degrees from ENS, Cachan, France, in 2011 and the Ph.D. degree from the Eastern Paris Federal University, Marne-la-Valle, France, in 2014. He is a Research Scientist at valeo.ai in Paris, France, a research oriented team for assisted and autonomous driving applications. His research interests at complex scene understanding with multiple data source from images to point cloud for both Earth observation and autonomous driving. Efficient machine learning and learning with partially annotated or un-anotated is at the core of his research topics

Sébastien Lefèvre (Senior Member, IEEE) received the MS. Eng and M.Sc. degrees from TU Compiègne, France in 1999, the Ph.D. degree from University of Tours, France, in 2002, and the Dr. Habil. degree from University of Strasbourg, France in 2009. He is Full Professor in Computer Science in University of South Brittany (Université Bretagne Sud) since 2010, where he conducts his researches within the Institute for Research in Computer Science and Random Systems (IRISA). He founded and led until 2021 the OBELIX team dedicated to image analysis and machine learning for remote sensing and Earth Observation (http://www.irisa.fr/obelix). His current research interests are mainly related to hierarchical image analysis and deep learning applied to remote sensing of environment. Prof. Lefèvre is an Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing. He was co-chair of GEOBIA 2016 and JURSE 2019.