Environmental Data Science (2025), 4: e22, 1–34
doi:10.1017/eds.2025.13

METHODS PAPER

Space-scale exploration of the poor reliability of deep learning models: the case of the remote sensing of rooftop photovoltaic systems

Gabriel Kasmi[1,2], Laurent Dubus[2,3], Yves-Marie Saint-Drenan[1] and Philippe Blanc[1]

[1]MINES Paris, Université PSL Centre Observation Impacts Energie (O.I.E.), 06904 Sophia-Antipolis, France
[2]RTE France, Direction de la Recherche et du Développement, 92973 Paris La Défense, France
[3]WEMC (World Energy & Meteorology Council), Norwich NR4 7TJ, UK
Corresponding author: Gabriel Kasmi; Email: gabriel.kasmi@minesparis.psl.eu

Received: 29 July 2024; **Revised:** 18 January 2025; **Accepted:** 26 February 2025

Keywords: deep learning; distribution shifts; remote sensing; solar energy; wavelets

Abstract

Photovoltaic (PV) energy grows rapidly and is crucial for the decarbonization of electric systems. However, centralized registries recording the technical characteristics of rooftop PV systems are often missing, making it difficult to monitor this growth accurately. The lack of monitoring could threaten the integration of PV energy into the grid. To avoid this situation, remote sensing of rooftop PV systems using deep learning has emerged as a promising solution. However, existing techniques are not reliable enough to be used by public authorities or transmission system operators (TSOs) to construct up-to-date statistics on the rooftop PV fleet. The lack of reliability comes from deep learning models being sensitive to distribution shifts. This work comprehensively evaluates distribution shifts' effects on the classification accuracy of deep learning models trained to detect rooftop PV panels on overhead imagery. We construct a benchmark to isolate the sources of distribution shifts and introduce a novel methodology that leverages explainable artificial intelligence (XAI) and decomposition of the input image and model's decision regarding scales to understand how distribution shifts affect deep learning models. Finally, based on our analysis, we introduce a data augmentation technique designed to improve the robustness of deep learning classifiers under varying acquisition conditions. Our proposed approach outperforms competing methods and can close the gap with more demanding unsupervised domain adaptation methods. We discuss practical recommendations for mapping PV systems using overhead imagery and deep learning models.

Impact Statement

This paper analyzes the effects of distribution shifts on deep learning models trained to detect rooftop photovoltaic (PV) systems on aerial imagery by combining explainable artificial intelligence methods. It then proposes practical solutions based on this analysis to enhance the robustness of these models, thereby improving their reliability and facilitating the use of remote sensing techniques to support the insertion of rooftop PV systems into the grid. The methodology laid out in this work can be replicated for other case studies.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

1. Introduction

Photovoltaic (PV) energy grows rapidly and is crucial for the decarbonization of electric systems (Haegel et al., 2017). The rapid growth of rooftop PV systems makes it challenging to of estimate the global PV installed capacity, as centralized data is often lacking (Hu et al., 2022; Kasmi et al., 2022). Remote sensing of rooftop PV systems using orthoimagery and deep learning models is a blooming solution for mapping rooftop PV installations. Deep learning-based pipelines have become the standard method for remote sensing PV systems, with works like DeepSolar (Yu et al., 2018) paving the way for country-wide mapping of PV systems using deep learning and airborne or spaceborne orthoimagery. Recently, methods for mapping rooftop PV systems in many regions, especially in Europe, have been proposed (Frimane et al., 2023; Kasmi et al., 2022; Kausika et al., 2021; Lindahl et al., 2023; Mayer et al., 2020; Rausch et al., 2020; Zech and Ranalli, 2020). Some of these works (Kasmi et al., 2022; Mayer et al., 2022) introduced methods to estimate the technical characteristics of the PV systems (individual localization, orientation, PV installed capacity). The identification of rooftop PV systems improves their integration into the electric grid by improving the ability of transmission system operators (TSOs) to more accurately estimate their power production in real-time (Kasmi et al., 2024) but can also promote their future expansion, as this data helps understanding the drivers behind rooftop PV adoption (Alipour et al., 2020; Colas and Saulnier, 2024; Graziano and Gillingham, 2015; Wang et al., 2022).

However, deep learning-based detection methods are sensitive to so-called distribution shifts, i.e., differences between the training and testing data (Koh et al., 2021). This sensitivity manifests by unpredictable and sharp accuracy drops when the model is deployed on unseen images. It limits their practical usability as a trained model cannot be deployed without retraining to carry out registry updates. Besides, the unpredictability of the model’s behavior limits its reliability as it casts doubt on *what* it perceives as a PV panel (De Jong et al., 2020; Hu et al., 2022). In this work, we define the reliability of a model as its ability to rely on relevant features to identify PV systems, i.e., to be “right for the right reasons” and to be simultaneously able to rely on robust features (Kasmi, 2024; Ross et al., 2017). Steps towards improving the quality of registries (i.e., tables recording the location and some technical information on PV systems) of rooftop PV systems constructed using deep learning algorithms have been taken, with Hu et al., 2022 and Kasmi et al., 2022 discussing the practical evaluation of the mapping algorithms or Li et al., 2021 identifying the minimum resolution to detect rooftop PV systems from orthoimagery (whether this minimum resolution is the native image resolution or the resolution obtained after increasing the input image resolution with methods such as those proposed by Ho et al., 2022). To date, Wang et al. (2017) is the only work that studied the poor generalizability of PV mapping algorithms, though it was limited to two cities and one image dataset. More recently, Pena Pereira et al. (2024) analyzed how PV system typologies and backgrounds affect performance, recommending input patch size adjustments and data augmentation to improve detection accuracy. Despite these advances, further work is needed to understand what the model identifies as a PV panel during training and how distribution shifts —arising from variations in PV systems, backgrounds, or acquisition conditions—impact performance.

This work aims to improve the reliability of deep learning models deployed in real-world settings prone to distribution shifts, taking the remote sensing of rooftop PV systems as a case study. We introduce a novel methodology to understand and address the sensitivity to distribution shifts based on empirical experiments and throughout analysis of the model’s decision using explainable AI (XAI) methods. Empirical evaluation and XAI methods enable us to identify the most important sources of distribution shifts and grasp why they occur. We evaluate a wide range of popular domain adaptation techniques (i.e., methods that aim at reducing the sensitivity to distribution shifts of deep learning algorithms) and introduce a novel data augmentation method. This method, based on our empirical findings, aims at effectively and reliably reducing the sensitivity to distribution shifts of deep learning models trained to detect PV systems from orthoimagery. We discuss practical takeaways regarding the choice of training data and domain adaptation methods for remote sensing PV systems. Since the sensitivity to distribution shifts is a recurring issue with the real-world deployment of deep learning systems (Koh et al., 2021), we discuss the requirements for applying our methodology to alternative use cases.

The code for replicating the results of this paper can be found at https://github.com/gabrielkasmi/robust_pv_mapping, and model weights can be found at https://zenodo.org/records/14673918.

2. Related works

2.1. Remote sensing of rooftop photovoltaic installations

The remote sensing of rooftop PV systems is now a well-established field with early works dating back to Golovko et al., 2018; Malof et al., 2015, Malof et al., 2016; Yuan et al., 2016. The DeepSolar project (Yu et al., 2018) marked a significant milestone by mapping distributed and utility-scale installations over the continental United States using state-of-the-art deep learning models. Many works built on DeepSolar to map regions or countries, especially in Europe, covering areas such as North-Rhine Westphalia (Mayer et al., 2020), Switzerland (Casanova et al., 2021), Oldenburg in Germany (Zech and Ranalli, 2020), parts of Sweden (Frimane et al., 2023; Lindahl et al., 2023), Northern Italy (Arnaudo et al., 2023), the Netherlands (Kausika et al., 2021) or the surroundings of Berkeley in California (Parhar et al., 2021), Connecticut (Malof et al., 2019) or the surroundings of Sfax, in Tunisia (Bouaziz et al., 2024). Several works even included GIS data to construct registries of PV installations (Kasmi et al., 2022; Kausika et al., 2021; Mayer et al., 2022; Rausch et al., 2020). In the current context of rapid rooftop PV growth (Haegel et al., 2017; RTE France, 2022), remote sensing of rooftop PV installations using deep learning and orthoimagery offers the potential to address the lack of systematic registration of small-scale PV installations (Kasmi et al., 2022; Kausika, 2022).

However, current methods cannot be transposed from one region to another without incurring accuracy drops, thus limiting their practical usability (Hu et al., 2022), as the aim of these models is to be regularly deployed on new images to construct and maintain official registries of PV systems (De Jong et al., 2020). The unpredictability of the accuracy drops also casts doubt regarding the reliability of these methods in such applied settings. To address this issue, Kasmi et al., 2022 recently introduced a method aiming at indirectly assessing the accuracy of the detections by automatically comparing the registry generated by deep learning algorithms to reference data, which is often aggregated at the city scale. While this work enabled the quantification of the drop in accuracy encountered during deployment, no cues as to why the accuracy varied during deployment were discussed. Kasmi et al. (2023a) introduced a benchmark to disentangle the sources of distribution shifts occurring with orthoimages of PV systems and outlined some promising directions to improve the reliability of deep learning algorithms. This work builds on and deepens the analysis of Kasmi et al., 2023a to propose a methodology for identifying the main sources of distribution shifts when dealing with the remote sensing of rooftop PV systems, understanding how these shifts affect deep learning models and extensively discussing how explainable AI techniques and domain adaptation methodologies can help mitigate the sensitivity to distribution shifts while improving the end user's trust towards deep learning black-boxes.

2.2. Distribution shifts and domain adaptation

Definition. Distribution shifts, i.e., the sensitivity to the fact that "*the training distribution differs from the test distribution*" (Koh et al., 2021) are ubiquitous in machine learning (Torralba and Efros, 2011). The sensitivity to distribution shifts causes unpredictable performance drops, which can have dire consequences as models are deployed in safety-critical settings such as autonomous driving (Sun et al., 2022b), medical diagnoses (Pooch et al., 2020) or finance (Thimonier et al., 2024). Distribution shifts formally consist of a break in the assumption that the training and testing (or deployment) data are independently and identically distributed (i.i.d., Zhou et al., 2023). This assumption is central when training models to minimize the empirical risk, as the underlying assumption is that the empirical risk is a good approximation of the true risk of the estimator. This assumption is true only if the data is i.i.d.; otherwise, the empirical risk no longer represents the true risk. It corresponds to *epistemic* uncertainty (Gal, 2016) as the model is exposed to data outside its prior training experience.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Distribution shifts in remote sensing. Due to its nature, remote sensing data often breaks the i.i.d. assumption (Tuia et al., 2024). For instance, the raw imagery consists of large image tiles cut into smaller thumbnails before being passed to the model. Therefore, the thumbnails exhibit a strong spatial correlation. Tuia et al. (2016) identified two primary sources of shifts in the input data to which models are sensitive: variations in the geographical scenery and varying acquisition conditions. Following Murray et al., 2019, we can add a third one: the ground sampling distance (GSD).

The acquisition conditions encompass the conversion of a scene into a digital image and include all sources of variability in the input images caused by different sensors, exposure, attitude and altitude during acquisition, and atmospheric conditions. The ground sampling distance (GSD) is the upper bound to the image's effective resolution. The effective resolution considers the distortions induced by the angle of incidence of the sensor (e.g., RGB camera). The lower the ground sampling distance, the more detailed the image. In practice, the effective resolution is limited by the GSD and the image quality (noise, optical transfer function, and intrinsic geometric consistency). In this article, with a slight abuse of wording, we will use the terms "GSD" and "resolution" interchangeably. The GSD corresponds to the distance between two consecutive pixels measured on the ground and is expressed in meters per pixel. The resolution measures the number of pixels per unit of length (e.g., inches or centimeters), and the image size describes its dimension. For consistency with the related literature, we will use the term "resolution" when broadly referring to the GSD. However, we will explicitly use "GSD" in specific contexts, such as when expressing it with its unit, as it makes no sense to refer to a "resolution of 0.2 m/pixel".

So far, the only work that investigated the poor reliability of deep learning models applied to the remote sensing of PV panels is Wang et al., 2017. The authors argued that the generalization ability from one city to another depends on how "hard" to recognize the PV panels are. However, no proper definition of the "hardness" to recognize PV panels or a proper disentanglement of the effect of each source of variability was carried out, and there was no prescription regarding model training or data preprocessing. More recently, Li et al., 2021 and Pena Pereira et al., 2024 studied the practical implications of having different resolutions or PV panel instances on the model's performance. These studies focus on the observable impacts of factors such as system heterogeneity, ground sampling distance, or image resolution on performance but overlook the underlying mechanisms driving the performance degradation. Following Kasmi et al., 2023a, we consider that improving the reliability of PV mapping models requires a deeper understanding of the underlying reasons for their sensitivity to these distribution shifts. Finally, to the best of our knowledge, Kasmi et al., 2023a is the only work to have implemented some domain adaptation techniques in the context of mapping rooftop PV systems.

Distribution shifts and domain adaptation. Domain adaptation is the go-for approach to address the sensitivity of machine learning models to distribution shifts (Ben-David et al., 2006). The different distributions are referred to as the source domain S, on which the model is initially trained, and the target domain T, on which the model is deployed. Different approaches can be distinguished depending on the number of source and target domains or the availability of labeled data. In its most constrained setting, we assume that we have access to labeled data from the source domain and, at most, unlabeled data from the target domain. This setting is sometimes referred to as *unsupervised* domain adaptation. We refer the reader to surveys such as Csurka, 2017; Csurka et al., 2021; Guan and Liu, 2022; Tuia et al., 2016; Zhou et al., 2023 for an extensive discussion of the domain adaptation settings and techniques. The general idea of domain adaptation is to learn a representation of the data invariant across domains or, equivalently, insensitive to distribution shifts.

We can distinguish two broad approaches to domain adaptation: implicit and explicit regularization. Implicit regularization encourages the model to generalize across domains without imposing specific constraints on the loss function during the initial training. Data augmentations form a first class of implicit regularization techniques. By viewing multiple copies of the same image that have been altered, the model learns to be invariant to these alterations. The aim is that the model is no longer sensitive to a given set of perturbations of the input images. Popular data augmentation methods consist of defining a method to generate as many perturbed samples as possible while preserving the semantic content of the image. To this end, AugMix (Hendrycks et al., 2020) applies a random sequence with random weights of

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

perturbations to the input image. Similarly, Hendrycks et al., 2022 augmented an input image with fractal patterns, and Sun et al., 2022a perturbed the Fourier spectrum of the input image. Cubuk et al., 2019 used a reinforcement learning framework to search for an optimal augmentation policy, selecting the type, magnitude, and probability of transformations based on a target validation set, and Cubuk et al., 2020 simplified this framework to make it computationally less demanding. Another approach for implicit regularization is to modify the model's architecture by enforcing additional invariances, such as the invariance to various groups of translations, reflections, and rotations as done by Cohen and Welling, 2016.

On the other hand, explicit regularization techniques require access to several source domains or unlabeled samples from the target domain, making these approaches more demanding than the implicit ones. The most popular approach is CORrelation ALignment (CORAL), and its counterpart for deep models DeepCORAL (Sun and Saenko, 2016), which aligns the distributions or the representations across domains by aligning their second-order statistics. On the other hand, Ganin et al., 2016; Shen et al., 2018 or Tzeng et al., 2017 leveraged adversarial training to align the feature representations across domains. More recently, invariant risk minimization (Arjovsky et al., 2019) ensured that the model's representation is invariant across environments by ensuring the model's predictions remained the same across domains. This approach, however, required at least two source environments to compute invariant representations and struggled to scale to complex model architectures such as ResNets (Zhou et al., 2022).

Fundamentally, improving the robustness against distribution shifts is a long-tailed problem, meaning that unseen situations eventually arise, and not all situations can be accounted for (Recht et al., 2019; Torralba and Efros, 2011). Therefore, to improve the reliability of deep learning systems and not only their robustness, we need to be able to characterize the representation learned by the model and understand how it is affected by the distribution shifts. To this end, we propose to use explainable artificial intelligence (XAI) methods.

2.3. Explainable artificial intelligence (XAI)

Modern deep learning algorithms are often qualified as black boxes, meaning it is hard to grasp their inner workings fully. This black-box nature limits the applicability of machine learning in safety-critical settings (Achtibat et al., 2022). We can distinguish two main approaches for machine learning explainability: by-design interpretable models and post-hoc explainability (Parekh, 2023). Flora et al. (2022) note that there is no consensus yet in the literature regarding the use of the terms explainability and interpretability. Following Flora et al., 2022, we say that a model is interpretable if it is inherently or by design interpretable, and a model is explainable if we can compute a post-hoc explanation of its decision. By-design interpretability aims at constructing models that are transparent and self-explanatory (Sudjianto and Zhang, 2021), e.g., the decision boundaries of a decision tree. On the other hand, post-hoc explainability seeks to explain a model's decision by highlighting important features contributing to this decision without explicitly stating how these features affected the model. Methods such as class activation maps (CAMs, Zhang et al., 2017), which plot a heatmap of the important image regions for the classification of this image, fall into this category.

XAI methods for model debugging. One of the main motivations for XAI is to inspect the decision of models to assess whether they rely on relevant factors to make predictions. Several works highlighted biases in the decision process, such as the reliance on spurious features. Lapuschkin et al., 2019 leveraged the GradCAM (Selvaraju et al., 2017) to show how classifiers could rely on watermarks rather than relevant areas of the input image for horses classification, thus highlighting a so-called "Clever Hans" (Pfungst, 1911) effect. CAMs (Zhang et al., 2017) have also been used to understand the behavior of convolutional neural networks (CNNs) in medical imagery classification by Zhang et al., 2021. Another example of the usage of XAI tools to understand and debug a model was proposed by Dardouillet et al., 2023, who leveraged SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017) to understand a model deployed for oil slick pollution detection on the sea surface. Going one step further, Andeol et al., 2023 recently used conformal predictions to improve the trustworthiness of railway signal detections, a case study where one needs to be sure that the model makes predictions for adequate reasons. In this work,

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

we exploit the complementarities between post-hoc and by-design interpretable XAI methods to provide a thorough understanding of the sensitivity to distribution shifts of CNNs deployed for mapping PV systems from orthoimagery.

3. Data

To analyze the effect of distribution shifts on deep learning models in the context of the remote sensing of rooftop PV systems, we rely on the training dataset Base de données d'apprentissage profond pour les installations photovoltaiques (Database for deep learning applied to PV systems, BDAPPV, Kasmi et al., 2023c). BDAPPV contains nearly 50,000 annotated images of PV systems. A very interesting feature of our case study is that the database contains annotations for 28,000 unique PV systems in France and neighboring countries. The training images were also retrieved from two different sources: satellite images coming from the Google Earth Engine (hereafter referred to as "Google," Gorelick et al., 2017) and aerial images coming from the IGN (IGN, 2024), the French public operator for geographic information. We have annotations for about 28,000 Google images and 17,000 IGN images. Both providers overlap, meaning we have two annotations for about 7,000 individual PV systems. The dataset is nearly balanced. We refer the reader to Kasmi et al., 2023c for more details regarding the dataset's characteristics. Figure 1 presents some samples coming from BDAPPV. We refer the reader to Section 4.1 to understand how we used BDAPPV to disentangle the different sources of distribution shifts.

4. Methods

We aim to explain why convolutional neural networks (CNNs) applied to detect PV panels on orthoimages are sensitive to distribution shifts. We first construct a benchmark to isolate the effect of the three main instances of distribution shifts on orthoimagery highlighted by Tuia et al., 2016 and Murray et al., 2019 using the BDAPPV dataset (see Section 4.1 for more details). These instances include the variability in the geographic location, varying acquisition conditions, and the varying ground sampling distance (GSD).

After quantifying the respective impact on prediction accuracy—measured by the F1 score—we leverage two XAI approaches to understand *why* these shifts affect the performance. Our working

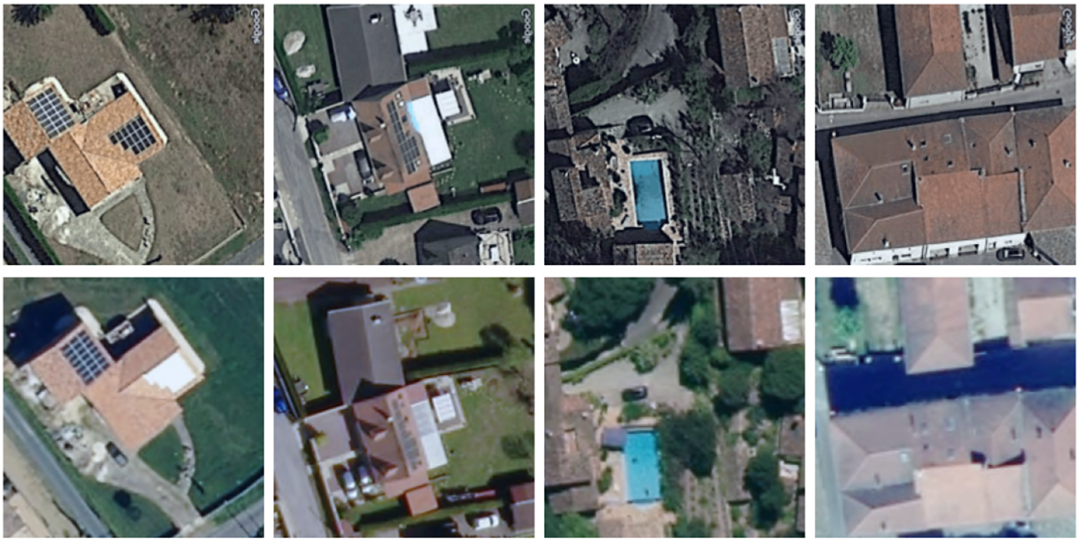

Figure 1. *Examples of images of the same PV panels but with different providers and acquisition dates (Up Google, down: IGN).*

hypothesis is that analyzing the model's prediction in terms of *scales* can help us understand why the model is sensitive to distribution shifts. Indeed, scales are located in space and, for each location, correspond to a dyadic partition of the frequency space. Therefore, given each location, we can identify which frequency ranges the model relies on. On the other hand, frequencies are unevenly affected by distribution shifts (for instance, high frequencies are more fragile, Chen et al., 2022). So, scales enable us to assess whether the model focuses on the PV panel to make a prediction *and* which frequencies it focuses on at this location. Using decomposition in terms of scale is particularly well suited in the case of remote sensing images since the scales, expressed in pixels on images, are indexed in meters and can thus point towards actual elements depicted in the images.

We combine two complementary approaches to explain the model's decision. Both of these approaches are grounded in the wavelet theory. On the one hand, we leverage a by-design interpretable model, the Scattering transform (Bruna and Mallat, 2013, introduced in Section 4.2.2). We compare the predictions of this model—which are intrinsically interpretable—with those of CNN to see when the predictions are the same and when they differ. On the other hand, we decompose the decision of the model using a post-hoc explainability method, the wavelet scale attribution method (WCAM, Kasmi et al., 2023b), which is a post-hoc explainability method to isolate the important scales in the predictions of our black-box CNN model.

Finally, based on our findings, we propose a data augmentation method to improve the robustness of CNNs, compare our approach with popular domain adaptation methods, and draw some conclusions regarding the choice of image data.

4.1. Disentangling the sources of distribution shifts on orthoimagery

BDAPPV features images of the same installations from two providers and records the approximate location of the PV installations. Using this information, we can define three test cases to disentangle the distribution shifts that occur with remote sensing data: the GSD, the acquisition conditions, and the geographical variability. Natively, our dataset disentangles the effect of the spatial shift, thanks to Google images being roughly geolocalized. Both the resolution and acquisition conditions vary when shifting from Google to IGN images. To disentangle the two sources of shifts, we downsampled the Google images to a GSD of 0.2 m/pixel to match the GSD of the IGN images. We chose not to upsample the IGN images to a GSD of 0.1 m/pixel as it would require adding information to the images and making additional assumptions regarding the method used to carry out the super-resolution task.

We train a ResNet-50 model (He et al., 2016) on Google images downsampled at 0.2 m/pixel of resolution and evaluate it on three datasets: a dataset with Google images at their native 0.1 m/pixel GSD ("Google 0.1 m/pixel"), the IGN images with a native 0.2 m/pixel GSD ("IGN") and Google images downsampled at 0.2 m/pixel located outside of France ("Google Spatial Shift"). We add the test set to record the test accuracy without distribution shift ("Google baseline"). We only do random crops, rotations, and ImageNet normalizations (i.e., with a mean of [0.485, 0.456, and 0.406] and a standard deviation of [0.229, 0.224, and 0.225]). Figure 2 plots examples of the different test images to disentangle the effects of distribution shifts. The baseline and IGN images represent the same panel at the same spatial resolution. The Google 0.1 m/pixel depicts the same scene but with the native resolution of Google images. Finally, the Spatial shift test set contains images from outside of France.

4.2. Space-scale decomposition of a model's decision process

4.2.1. Background: the wavelet transform of an image

Motivation and definition. We propose to analyze the decision process of an off-the-shelf CNN model through the lenses of space-scale or wavelet decomposition. Wavelets are a natural tool to decompose an image into scales while maintaining local analysis in space: they provide a single space-scale decomposition. As scales are indexed in terms of actual distances on the ground, we can directly identify the important objects contributing to a model's decision by studying the important scales. In appendix A, we provide further evidence of the limitation of "traditional" feature attribution methods for explaining the

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

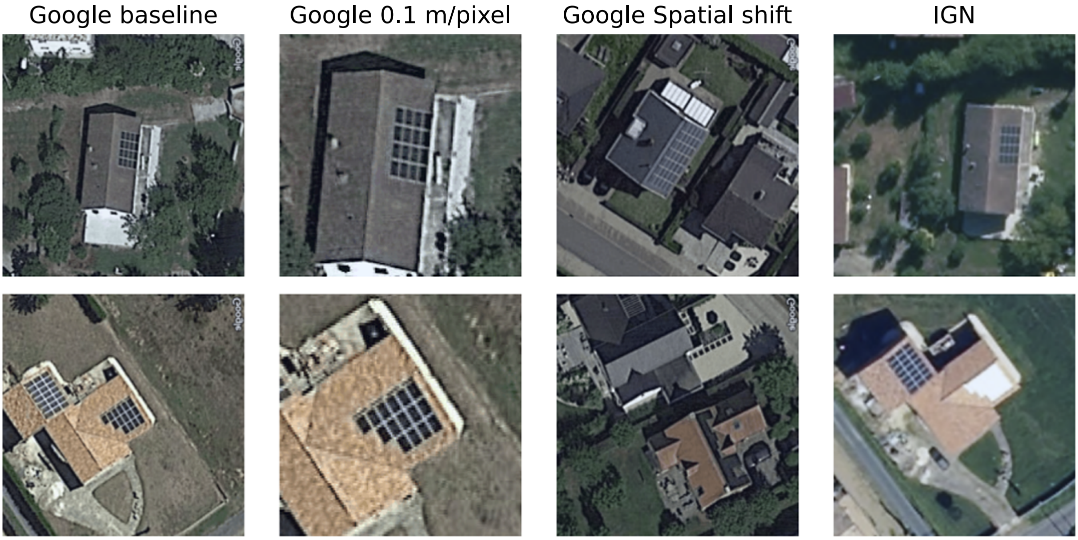

Figure 2. *Test images on which a model trained on Google images (downsampled to 0.2 m/pixel of GSD, "Google baseline") is evaluated. "Google 0.1 m/pixel" corresponds to the source Google image before downsampling and evaluates the effect of the varying image resolutions. "Google Spatial Shift" corresponds to Google images taken outside of France. "IGN" corresponds to images depicting the same installations as Google baseline but with a different provider.*

false detection of deep learning models in our use case. Figure 3 illustrates the objects that can be found at different scales of an orthoimage.

A wavelet is an integrable function $\psi \in L^2(\mathbb{R})$ with zero average, normalized, and centered around 0. Unlike a sinewave, a wavelet is localized in space and in the Fourier domain. This implies that dilatations of this wavelet enable to scrutinize different scales, while translations enable to scrutinize spatial location. In other words, scales correspond to different spatial frequency ranges or spectral domains.

To compute an image's (continuous) wavelet transform (CWT), one first defines a filter bank $\mathcal{D}$ from the original wavelet ψ with the scale factor s and the 2D translation in space u. We have

$$\mathcal{D} = \left\{ \psi_{s,u}(x) = \frac{1}{\sqrt{s}} \psi\left(\frac{x-u}{s} \right) \right\}_{u \in \mathbb{R}^2,\ s \geq 0}, \tag{1}$$

The computation of the wavelet transform of a function $f \in L^2(\mathbb{R})$ at location x and scale s is given by

$$\mathcal{W}(f)(x,s) = \int_{-\infty}^{+\infty} f(u) \frac{1}{\sqrt{s}} \psi^* \left(\frac{x-u}{s} \right) \mathrm{d}u, \tag{2}$$

which can be rewritten as a convolution (Mallat, 1999). Computing the multi-level decomposition of f requires applying Equation 2 J times, with $1 \leq s \leq J$. J denotes the number of levels of decomposition. For each scale, the translation in space u corresponds to the orientations at a given level.

Mallat (1989) showed that one could implement the multi-level dyadic decomposition of the discrete wavelet transform (DWT) by applying a high-pass filter H to the original signal f and subsampling by a factor of two to obtain the *detail* coefficients and applying a low-pass filter G and subsampling by a factor of two to obtain the *approximation* coefficients. Iterating on the approximation coefficients yields a multi-level transform where the j^{th} level extracts information at resolutions between 2^j and 2^{j-1} pixels. The detail coefficients can be decomposed into various rotations (usually horizontal, vertical, and diagonal) when dealing with 2D signals (e.g., images).

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

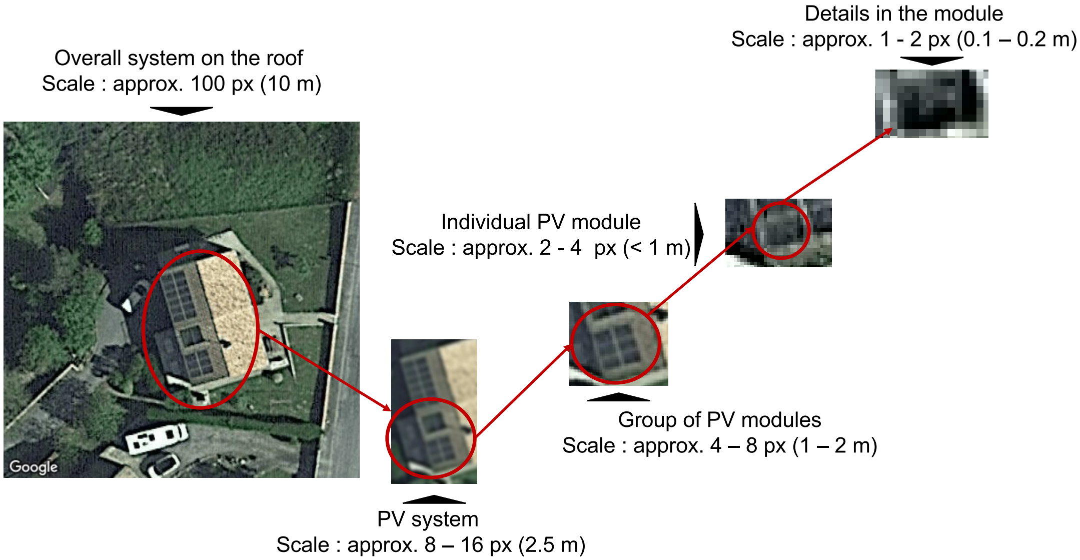
Overall system on the roof
Scale : approx. 100 px (10 m)
Google
PV system
Scale : approx. 8 – 16 px (2.5 m)
Group of PV modules
Scale : approx. 4 – 8 px (1 – 2 m)
Individual PV module
Scale : approx. 2 - 4 px (< 1 m)
Details in the module
Scale : approx. 1 - 2 px (0.1 – 0.2 m)

Figure 3. *Decomposition of a PV panel into scales.*

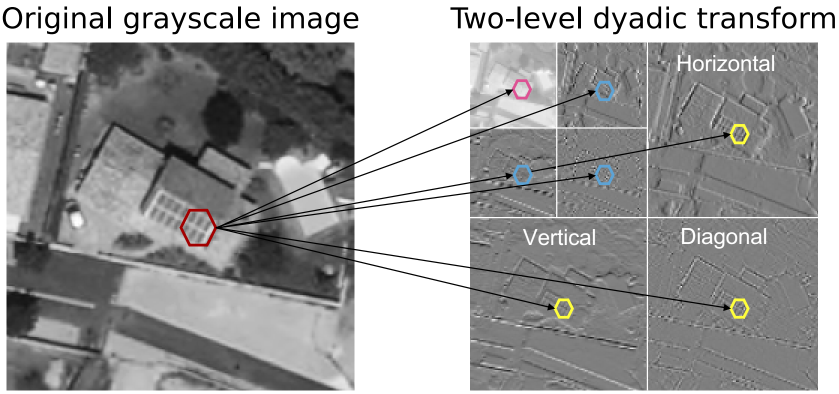

Figure 4. *Image and associated two-level dyadic wavelet transform with indications to interpret the wavelet transform of the image. "Horizontal," "diagonal," and "vertical" indicate the direction of the detail coefficients. The direction is the same at all levels.*

Interpreting the wavelet transform of an image. Figure 4 illustrates how to interpret the (two-level) wavelet transform of an image. Reading is the same for any multi-level decomposition. The right image plots the two-level dyadic decomposition of the original image on the left. Following this transform, the localization on the image highlighted by the red polygon can be decomposed into six detail components (marked yellow and blue) and one approximation component (marked pink). Each detail component has three directions: horizontal, vertical, and diagonal. The yellow components correspond to details at the 1–2 pixel scale, and the blue components to the details at the 2–4 pixel scale. For each location, the wavelet transform summarizes the information in the image at this scale and location.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

4.2.2. By design interpretable XAI method: the Scattering transform

The Scattering transform (Bruna and Mallat, 2013) is a deterministic feature extractor. CNNs and the Scattering transform share the same multi-level architecture, where the previous layer's output is passed onto the next after a nonlinearity is applied. The nonlinearities in a CNN are generally rectified linear units (ReLU), whereas in the Scattering transform, it is a modulus operation. Unlike CNNs, whose kernel coefficients are learned during training, the coefficients of the Scattering transform are fixed. Bruna and Mallat (2013) showed that the Scattering transform computes representations from an input image that share the same properties of translational invariance as the representations computed with a CNN. The advantage of the Scattering transform is that as filters are fixed, we can know precisely what information they extract from the input image. Figure 5 summarizes the feature extraction process of the Scattering transform.

The input image x is downsampled, and a wavelet filter ϕ is applied in J directions. The wavelet coefficients at that scale are retrieved (black arrows), and the image is passed onto the next layer (blue arrows). As the depth increases, the spatial extent covered by the filters decreases. At each spatial location, one takes the modulus of the wavelet transform to compute a scale-invariant representation that indicates the amount of "energy" in the image at this scale and localization.

The Scattering transform is parameterized by the number m of layers and the number J of orientations. We have a total of $mJ + m^2J(J-1)/2$ coefficients. At the end of the decomposition, the features, i.e., the scattering coefficients, are flattened into a single vector of size $mJ + m^2J(J-1)/2$. We can identify to which scale, location, and orientation on the input image this feature corresponds.

We implement three variants of the Scattering transform with depths m varying from one to three levels. Bruna and Mallat (2013) stated that first-order coefficients were insufficient to discriminate between two very different images but that coefficients of order $m = 2$ could. We consider $J = 8$ orientations. We stack the scattering coefficients into a vector of dimension $mJ + m^2J(J-1)/2$, akin

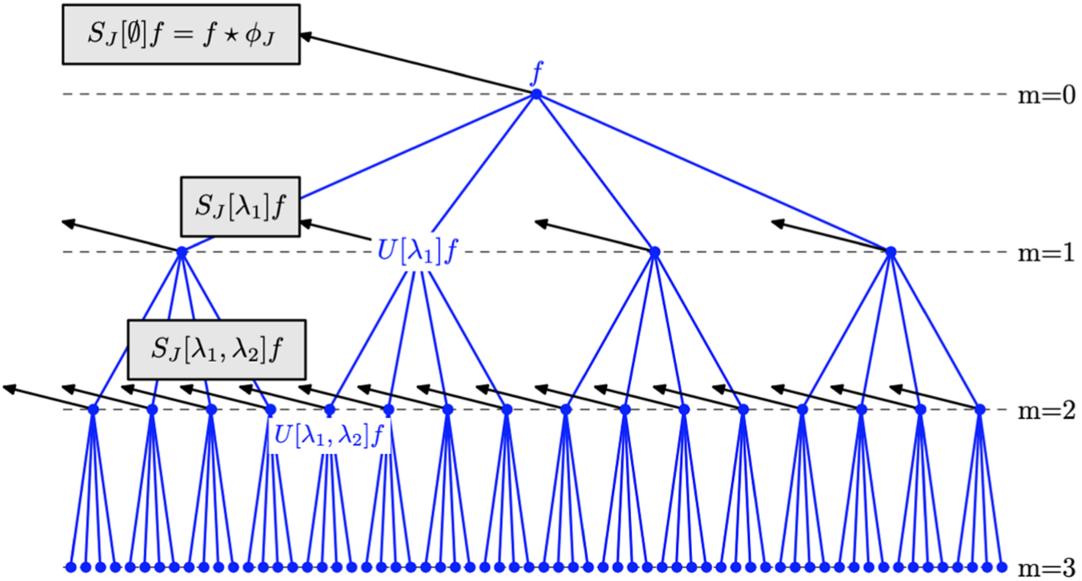
$S_J[\emptyset]f = f \star \phi_J$
f
m=0
$S_J[\lambda_1]f$
$U[\lambda_1]f$
m=1
$S_J[\lambda_1, \lambda_2]f$
m=2
$U[\lambda_1, \lambda_2]f$
m=3

Figure 5. *A scattering propagator U_J applied to x computes each $U[\lambda_1]x = |x \star \psi_{\lambda_1}|$ and outputs $S_J[0/]x = x \star \phi_{2^J}$ (black arrow). Applying U_J to each $U[\lambda_1]x$ computes all $U[\lambda_1, \lambda_2]x$ and outputs $S_J[\lambda_1] = U[\lambda_1] \star \phi_{2^J}$ (black arrows). Applying U_J iteratively to each $U[p]x$ outputs $S_J[p]x = U[p]x \star \phi_{2^J}$ (black arrows) and computes the next path layer. Figure borrowed from Bruna and Mallat, 2013. Note: In the image, the input x corresponds to f and $\lambda = 2^j r$ is a frequency variable corresponding to the j^{th} scale with r rotations.*

to the penultimate layer of a CNN. We train a linear classifier on this feature vector. Our implementation of the Scattering transform is based on the Python library Kymatio (Andreux et al., 2020).

4.2.3. Post-hoc XAI method: the wavelet scale attribution method (WCAM)

Traditional feature attribution methods (Petsiuk et al., 2018; Selvaraju et al., 2020; Simonyan and Zisserman, 2015) highlight the important areas for the prediction of a classifier in the pixel (spatial) domain. The WCAM (Kasmi et al., 2023b) generalizes attribution to the wavelet (space-scale domain). The WCAM provides us with two pieces of information: where the model sees and what scale it sees at this location. The decomposition of the prediction in terms of scales points towards actual elements on the input image since orthoimagery scales are indexed in meters. For example, on Google images, details at the 1–2 pixel scale correspond to physical objects with a size between 0.1 and 0.2 m on the ground. Thus, we know what the model sees as a panel; we can interpret it and assess whether it is sensitive to varying acquisition conditions. We refer the reader to appendix B or to Kasmi et al., 2023b for more details on the computation of the WCAM.

Reading a WCAM. Figure 6 presents an example of an explanation computed using the WCAM. On the right panel, we can see the important areas in the model prediction highlighted in the wavelet domain. On the left panel, we can see the spatial localization of the important components. We can see two main spatial locations: the center of the image, which depicts the PV panel, and the bottom left, which depicts a pool. Disentangling the scales, we can see that the PV panel's importance spreads across three scales (orange arrows), while the pool is only important at the 4–8 pixel scale. This underlines that the model focuses on the PV panel because it sees details ranging from small details in the PV modules to the cluster of modules.

4.3. Improving the robustness through implicit regularization

Improving the robustness to noise and scale perturbations. Since we know that varying acquisition conditions induce perturbations which primarily affect high-frequency components (i.e., the finest scales, Lone and Siddiqui, 2018, we primarily focus on implicit regularization and more precisely data augmentations. Indeed, data augmentation is sufficient to enforce invariance to alterations in the frequency domain. Besides, they are easier to implement for deep learning practitioners and do not require having access to samples from the target domain. For the sake of completeness, we compare our results with explicit regularization techniques. We evaluate popular data augmentation methods to

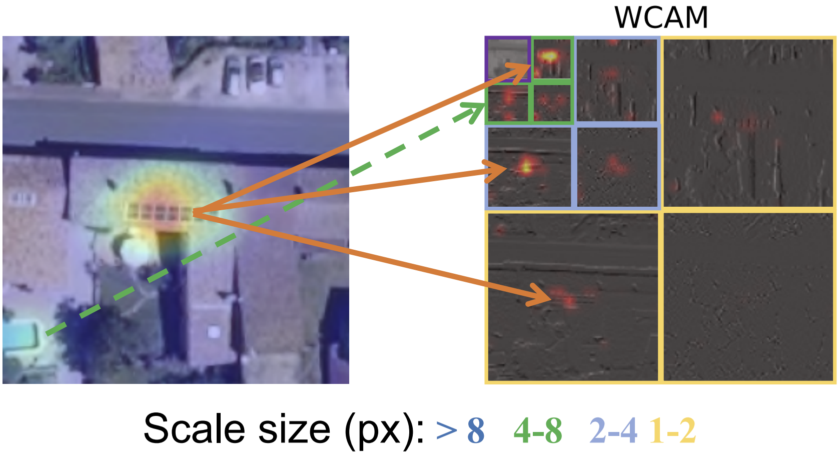

Figure 6. Decomposition in the wavelet domain of the important regions for a model's prediction with the WCAM.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

improve the robustness of classification models to image corruptions (Cubuk et al., 2019; Cubuk et al., 2020; Geirhos et al., 2019; Hendrycks et al., 2020, 2022). We consider the AugMix method (Hendrycks et al., 2020) and the recently-proposed RandAugment (Cubuk et al., 2020) and AutoAugment (Cubuk et al., 2019) methods. We refer the reader to appendix D.1 for a detailed presentation of these methods.

Proposed data augmentation methods. As a baseline, we propose blurring the input image and refer to this method as **Blurring**. We apply a nonrandom Gaussian blur to the image. The value is set by comparing visually Google and IGN images and trying to remove details from Google images that are not visible on IGN images. After a manual inspection, we set the blur level to discard the details at 0.1–0.2 m scale from the image. It corresponds to a blurring value $\sigma = 2.$ in the `ImageFilter.GaussianBlur` method of the Python Imaging Library (PIL). Our proposed method consists of combining blurring, which removes the small-scale details of the image with a random perturbation of the wavelet transform of the image. We randomly set some wavelet coefficients to 0 sand reconstructed the image from its perturbed coefficients. The perturbation is done across all scales, and the set of coefficients set to 0 is determined using uniform sampling. This results in a random perturbation that removes information for some precise scales and locations. We then reconstruct the image from its perturbed wavelet coefficients. For each call, 20% of the coefficients are canceled. This value balances between the loss of information and the input perturbation. We perturb each color channel independently. The wavelet perturbation aims to disrupt information at specific scales, as it can happen with varying acquisition conditions. The resulting data augmentation method is referred to as **Blurring + Wavelet perturbation (WP)**. Figure 7 presents examples of perturbed images using our method.

Domain adaptation. We complement our analyses by comparing our approach with popular domain adaptation techniques. These techniques are more demanding as unlabeled data from the target domain is

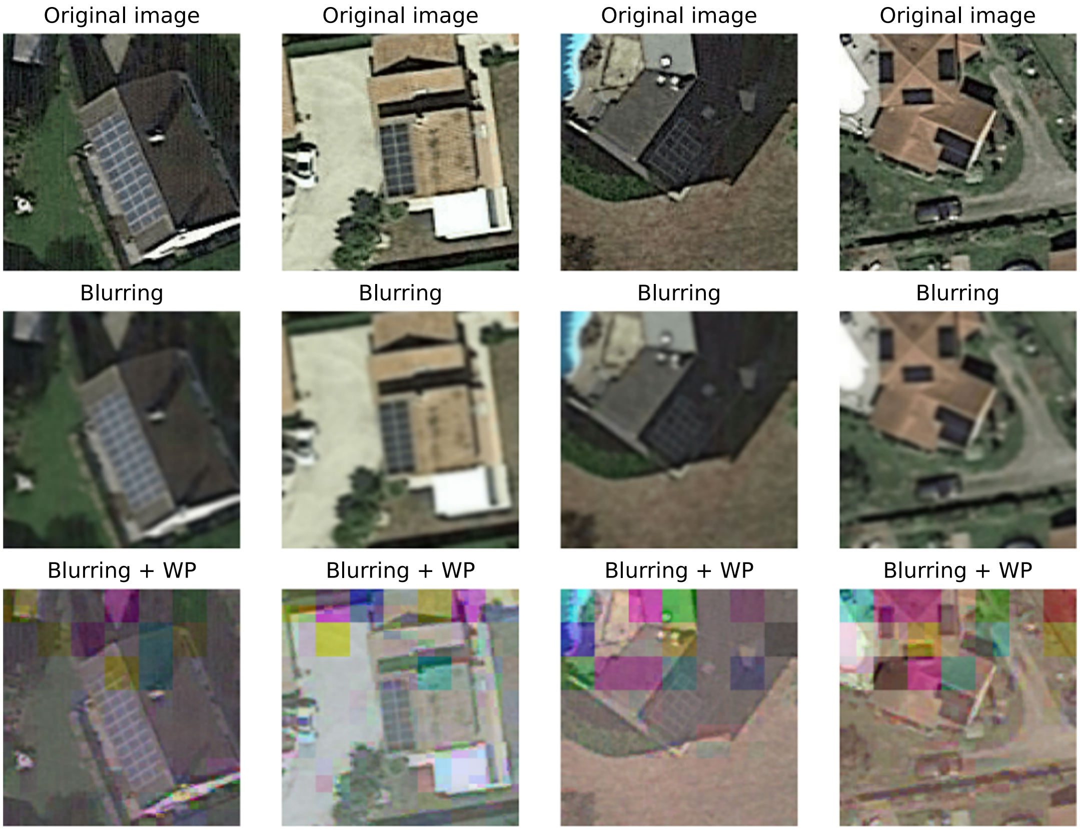

Figure 7. *Illustration of the effect of our data augmentation method on a sample of images.*

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

required. We refer the reader to appendix E for a discussion of the results obtained with the domain adaptation techniques.

5. Results

5.1. Deep models are mostly sensitive to varying acquisition conditions, leading to an increase in the number of false negatives

Table 1 shows the results of the decomposition of the effect of distribution shifts into three components: resolution, acquisition conditions, and spatial shift. We can see that the F1 score drops the most when the model faces new acquisition conditions. The second most significant impact comes from the change in the resolution. However, the performance drop remains relatively small compared to the effect of the acquisition conditions (which can also be assimilated to variations in the image quality). In our framework, there is no evidence of an effect of geographical variability once we isolate the effects of acquisition conditions and resolution. This effect is probably underestimated, as images of our dataset that are not in France are near France. However, the effect of the acquisition conditions is sizeable enough to seek methods for addressing it.

5.2. The Scattering transform shows that clean, fine-scale features are transferable but poorly discriminative

Discriminative and transferable features. In the following, we distinguish between two kinds of features: the *discriminative* and the *transferable* features. Discriminative features enable the model to discriminate well between PV and non-PV images. Relying on discriminative features ensures a low number of false positives. On the other hand, transferable features correspond to features that generalize well across domains. If a model relies on transferable features, its performance should remain even across domains. Ideally, we would like a model to rely on discriminative and transferable features.

Accuracy of the Scattering transform. Table 2 presents the accuracy results of the Scattering transform and compares it with a random classifier and the ERM (which is the same model as the one evaluated in Table 1). We can see that the performance on the source domain lags behind the performance of the CNN, but the Scattering transform generalizes better to IGN than the CNN. However, this comes at the cost of a

Table 1. ***F1 Score*** *and decomposition in true positives, true negatives, false positives, and false negatives rates of the classification accuracy of a CNN model trained on Google images (Google baseline) and tested on the three instances of distributions shifts: GSD (Google 0.1 m/pixel), the geographical variability (Google Spatial Shift) and the acquisition conditions (IGN).*

Shift instance	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
None (Google baseline)	0.98	0.99	0.98	0.02	0.01
GSD (Google 0.1 m/pixel)	0.89	0.81	1.00	0.00	0.19
Geography (Google spatial shift)	0.98	0.99	0.98	0.02	0.01
Acquisition conditions (IGN)	0.46	0.32	0.95	0.03	0.68

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Table 2. ***F1 Score*** *and decomposition in true positives, true negatives, false positives, and false negative rate of the classification accuracy of the Scattering Transform model trained on Google images and deployed on IGN images. The best results are* ***bolded****.*

Dataset	Model	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
Google baseline	Scattering transform	0.57	0.89	0.10	0.56	0.48
	CNN (ERM)	**0.98**	**0.99**	**0.98**	**0.02**	**0.01**
	Random classifier	0.47	0.50	0.50	0.55	0.45
IGN	Scattering transform	**0.59**	**0.54**	0.31	0.62	0.54
	CNN (ERM)	0.46	0.32	**0.95**	**0.03**	0.68
	Random classifier	0.47	0.50	0.50	0.56	**0.44**

high false positive rate. Table F2 in appendix F.2 presents similar accuracy results for variants of the Scattering transform model in the depth and number of features.

Implications for the CNN. We know which features the Scattering transform relies on. It leverages information at the two-pixel scale after downsampling the input image. In other words, the Scattering transform makes predictions based on *clean* features at the two-pixel scale. Therefore, we can deduce that these features are *transferable*, as the performance remains even across datasets, but not very *discriminative* as the false positives rate is high (across both datasets). Therefore, the analysis of the errors of the Scattering transform and the CNN highlights a potential trade-off between transferable and discriminative features

On the other hand, the CNN should rely on *discriminative* features located at coarser scales than 8 pixels, and on noisy features. In Section 5.3, we investigate how the distortion of the input image's coarse scales impacts the CNN's decision process and the shift in its predicted probability. In Section 5.4.1, we discuss how noise in input images affects the generalization ability of the CNN.

5.3. CNNs are sensitive to the distortion of coarse-scale discriminative features

Predicted probability shifts. The CNN outputs a predicted probability of a PV panel on the input image. When evaluating the CNN on the same scene from two providers, we compute *predicted probability shift* $\Delta p = |p_{ign} - p_{google}|$ when the model trained on Google is evaluated on IGN images. p_{google} denotes the predicted probability on Google images and p_{ign} on IGN images. By construction, $\Delta p \in [0,1]$. If $\Delta p = 0$, the predicted probability did not change when changing the provider. On the other hand, if $\Delta p \rightarrow 1$, then it means that the model made a different prediction solely because of the new acquisition condition.

Correlations between the probability shift and low-scale similarity of the images. For all images in our test set ($n = 4321$), we compute the similarity between the low-scale components of the input image across the two domains. This enables us to assess how similar images depicting the same scene on Google and IGN are, with respect to their low-scale components, i.e., components larger than 8 pixels, which correspond to the approximation coefficients of a 3-level dyadic decomposition of the image.

On the other hand, we compute the predicted probability shift for each image across two domains. The predicted probability shift indicates how much the model's prediction changed when facing the IGN image.

Suppose the CNN is indeed sensitive to low-scale perturbations of the input image. In that case, we expect a correlation between the dissimilarity between the approximation coefficients (which only contain

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

the low-scale components of the image) and the predicted probability shift (which indicates whether the model changed its prediction once faced with a new image).

We evaluate the similarity between the approximation coefficients using two metrics: the Structural similarity index measure (SSIM, Wang et al., 2004) and the Euclidean distance between the approximation coefficients. The SSIM takes values between -1 and 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates perfect anti-correlation. On the other hand, the Euclidean distance takes positive values; the greater the distance, the greater the dissimilarity between the images.

We evaluate the correlation between the similarity of the approximation coefficients and the magnitude of the probability shift using the Pearson correlation coefficient (PCC, Pearson and Galton, 1895). The PCC is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. The PCC value ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 is a perfect positive linear relationship, and 0 is no linear relationship (the variables are uncorrelated). Given two random variables X and Y, the PCC is given by

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\mathrm{Cov}(X,Y)$ denotes the covariance between X and Y and σ is the standard deviation. In addition to computing the PCC, we report its p-value to assess whether the reported value significantly differs from 0, thus rejecting the hypothesis that the variables are uncorrelated.

As expected, we obtain a negative Pearson correlation coefficient equal to -0.41 (with a p-value $< 10^{-5}$) between the input images' SSIMs and the predicted probability shift. Using the Euclidean distance, we obtain a correlation coefficient of 0.250 ($p < 10^{-5}$). These results back the idea that the CNN is sensitive to low-scale perturbations of the input image, which results in a shift in the predicted probability.

Visualization of the model's response with the WCAM. The WCAM disentangles the important scales in a model's prediction. It enables us to see which scales were disrupted. On Figure 8, we present an example of an image that was initially identified as a PV panel but turned out to be no longer recognized on IGN images.

We can see that in both cases, the approximation details are important in the model's prediction. The model responds to distortions at this scale by no longer focusing on a single area. Indeed, the model weights more components located at the 2–4 and 4–8 pixel scale (orange circles), which were not as important initially. At the level of the perturbed scales, we can also witness that the model is disrupted by factors lying next to the PV panel (green circle). We supply more examples of such cases in appendix G and discuss the quantitative analysis of this result in appendix C.

5.4. Pathways towards improving the robustness to acquisition conditions

5.4.1. Blurring and wavelet perturbation improve accuracy

Table 3 reports the results of our data augmentation techniques and compares them with existing methods. We can see that augmentations that explicitly discard small scales (high frequencies, i.e. Blurring and Blurring + WP) information perform the best. However, the blurring method sacrifices the recall (which drops to 0.6) to improve the F1 score. In Table 3, this can be seen by the increase in false positives rate. Therefore, this method is unreliable for improving the robustness to acquisition conditions. We recall that the true positive rate and the false negative rate divide the number of true positives (resp. false negatives) by the number of positive samples. Similarly, the true negative and false positive rates divide the number of true negatives (resp. false positive) by the number of negative samples in the dataset. The true positive rate corresponds to the recall, and the true negative rate to the specificity.

On the other hand, adding wavelet perturbation (WP) contributes to restoring the accuracy of the classification model without sacrificing the precision or the recall. While the drop in accuracy is still sizeable compared to the Oracle, the gain is consistent compared to other data augmentation techniques. Compared to RandAugment, the best-benchmarked method, our Blurring + WP is closer to the targets regarding true positives and true negatives and makes lower false negatives. This experiment shows that it is possible to consistently and reliably improve the robustness of acquisition conditions using a data augmentation technique, which does not leverage any information on the IGN dataset.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

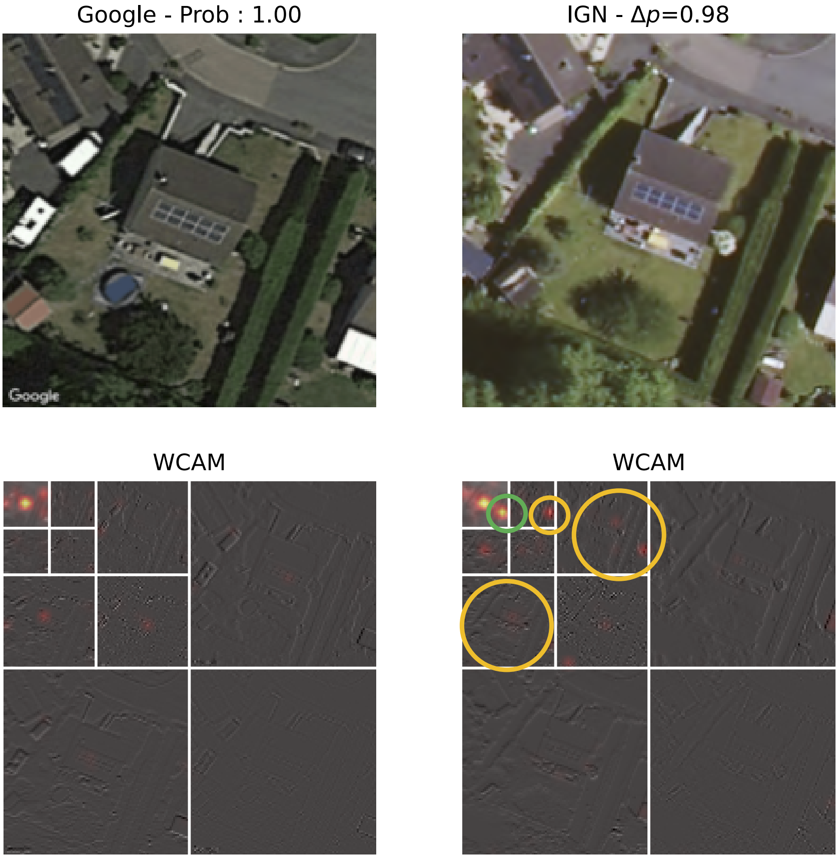

Figure 8. *Analysis with the WCAM of the CNNs prediction on an image no longer recognized as a PV panel.*

Table 3. ***F1 Score*** *and decomposition in true positives, true negatives, false positives, and false negatives rate for models trained on Google with different mitigation strategies. Evaluation on IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. Best results are* ***bolded****, second-best results are underlined, values highlighted in red indicate the worst performance, and values in orange indicate the second-to-last worst performance*

	Model	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
Augmentations	Oracle	0.88	0.96	0.82	0.18	0.04
	None (ERM)	0.44	0.30	**0.96**	0.04	0.70
	AutoAugment	0.46	0.31	0.96	**0.04**	0.69
	AugMix	0.48	0.33	0.96	0.04	0.67
	RandAugment	0.51	0.37	0.94	0.06	0.63
	Blurring (Ours)	**0.74**	**0.98**	0.49	0.51	**0.02**
	Blurring + WP (Ours)	0.58	0.47	0.87	0.13	0.53

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Table 4. ***F1 Score*** *and true positives, true negatives, false positives, and false negatives rates. Evaluation computed on the Google dataset. ERM was trained on Google and Oracle on IGN images*

Model	F1 Score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
ERM (Vapnik, 1999)	0.98	0,98	0.98	0,02	0,02
Oracle (ERM trained on IGN)	0.91	0,94	0,89	0,11	0,06

5.4.2. On the role of the input data: towards some practical recommendations regarding the training data.

Generalizability of the feature representation. Our results show that lowering the reliance on high-frequency content in the image improves generalization. This content is located on the 0.1-0.2 m scale and only appears on Google images. In Table 4, we flip our experiment to study how a model trained on IGN images generalizes to Google images. Results show that the model trained on IGN generalizes better to the downscaled Google images than the opposite. This result further supports the idea that higher resolution is not necessarily better for good robustness to acquisition conditions.

Reliability trade-offs. The training data is often considered as given in many practical settings, especially given the high cost of annotating samples. This motivates the use of domain adaptation techniques, such as those described in this work. From Tables 3 and E1 in appendix E, it should be noted that the F1 score can be misleading regarding how the different methods attenuate the effects of the distribution shifts. In particular, it should be noted that Blurring achieves a very high F1 score at the expense of the number of false positives. On the other hand, the domain adaptation method Wasserstein Distance Guided Representation Learning (WDGRL) exhibits a relatively false negative rate. Depending on the task at hand, different methods can be preferred. In the case of the remote sensing of rooftop PV systems, false negatives can be an issue, thus leading to favor solutions such as Blurring or Adversarial Discriminative Domain Adaptation (ADDA), even though we know that these methods generate a lot of false detections.

6. Discussion

6.1. Conclusion

This work aims to explain why convolutional neural networks (CNNs) applied to detect PV panels on orthoimages are sensitive to distribution shifts. We first set up an experiment to disentangle the effects of the three main distribution shifts occurring in remote sensing (Murray et al., 2019; Tuia et al., 2016), namely geographical variability, varying acquisition conditions, and varying resolution. We showed that the varying acquisition conditions contribute significantly to the observed performance drop. To explain why this drop occurs, we leverage space-scale analysis to disentangle the different scales from the input images. We combine two types of explainable AI methods grounded in the wavelet decomposition of the input images to show that the CNN relies on noisy features (at the finest scales) and features that are not very well transferable across domains (at the coarsest scales).

We then introduced a data augmentation technique to improve the model's robustness to distortions of the coarse-scale features and remove noise from the fine-scale features. We compare this method against popular data augmentation techniques and show that our approach outperforms these baselines. We also compared our approach with more demanding domain adaptation techniques and showed that our approach remains competitive. We then discussed several practical takeaways of this study for the training or the choice of the training data for the initial training of the deep learning model.

Broader impact. Mapping rooftop PV systems is a recurring issue in many countries, and a lot of actors interested in such rooftop PV registries require reliable data (De Jong et al., 2020; Kasmi, 2024). While offering the possibility to quickly and cheaply map PV systems over vast areas, current methods for mapping rooftop PV installations lack reliability owing to their poor generalization abilities beyond their

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

training dataset (De Jong et al., 2020). This work addresses this gap and thus demonstrates that remote sensing of PV installations is a reliable way to construct registries of rooftop PV systems.

The methodology introduced in this work, which consists of first isolating the main source of performance drop among possible types of distribution shifts, then leveraging XAI methods to grasp better the impact of these shifts on the model's predictions to finally highlight *how* invariance to these shifts can be mitigated can be replicated to other case studies.

6.2. Limitations and future works

Further discussion of the geographical variability. Our training data was limited to a narrow area around France. Therefore, we suspect the effect of the geographical variability to be underestimated. For instance, Freitas et al., 2023 showed that fine-tuning a model with data that is *not far* from the target area (e.g., France when the goal is to map PV systems in Portugal) enables accuracy gains compared to directly transferring a model trained over the United States. It could be interesting to study how the performance varies with the distance between the training data and the target mapping area once all other factors (acquisition conditions, resolution) are accounted for.

Extensions to other models. Over the last couple of years, foundation models (Bommasani et al., 2022) have been redefining the standards in deep learning. These very large models, trained on large data corpora, have shown remarkable performance for many challenging tasks, especially for text (Brown et al., 2020) and image (Rombach et al., 2022) generation. These models are used for more conventional and specialized tasks such as image segmentation (Kirillov et al., 2023) and achieve superior performance to conventional approaches while only requiring a few samples to learn their new task. Extending this benchmark and evaluating the performance of foundation models fine-tuned for segmenting PV panels, such as Yang et al., 2024, under distribution shifts could be interesting.

Application to other case studies. The key ingredients to replicate our methodology are the disentanglement of the effect of the different types of shifts occurring in the case study at hand and the combination of various XAI methods to build a relevant intuition regarding the characterization of the feature representation of the model, in terms of semantic relevance and in terms of robustness to the types of shifts that are recurring in the case study. Introducing new domain adaptation methods should not be prioritized over thorough decomposition of the distribution shifts and analysis of their effects.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.13.

Acknowledgments. The authors would like to thank Reviewer 1 for the detailed comments and feedback on our manuscript, which significantly contributed to improving the quality and depth of this work.

Author contribution. Conceptualization, Gabriel Kasmi; Formal analysis, Gabriel Kasmi; Funding acquisition, Laurent Dubus; Investigation, Gabriel Kasmi; Methodology, Gabriel Kasmi; Project administration, Laurent Dubus; Software, Gabriel Kasmi; Supervision, Philippe Blanc, Yves-Marie Saint-Drenan and Laurent Dubus; Validation, Gabriel Kasmi; Writing—original draft, Gabriel Kasmi; Writing – review & editing, Gabriel Kasmi, Philippe Blanc, Yves-Marie Saint-Drenan Laurent Dubus. All authors approved the final submitted draft.

Competing interests. The authors declare no conflicts of interest.

Data availability statement. Code for replicating the results of this paper can be found at: https://github.com/gabrielkasmi/robust_pv_mapping. Model weights can be found at: https://zenodo.org/records/12179554.

Funding statement. This research was supported by a grant from the ANRT (CIFRE funding 2020/0685) and was funded by the French transmission system operator RTE.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Achtibat R, **Dreyer M**, **Eisenbraun I**, **Bosse S**, **Wiegand T**, **Samek W and Lapuschkin S** (2022) From "Where" to "What": Towards human-understandable explanations through concept relevance propagation. [arXiv:2206.03208 [cs]]. https://doi.org/10.48550/arXiv.2206.03208

Alipour M, Salim H, Stewart RA and Sahin O (2020) Predictors, taxonomy of predictors, and correlations of predictors with the decision behaviour of residential solar photovoltaics adoption: A review. *Renewable and Sustainable Energy Reviews 123*, 109749. https://doi.org/10.1016/j.rser.2020.109749

Andeol L, Fel T, de Grancey F and Mossina L (2023) Confident object detection via conformal prediction and conformal risk control: An application to railway signaling. In Papadopoulos H, Nguyen KA, Boström H and Carlsson L (eds.), *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, Vol. *204*. PMLR, pp. 36–55

Andreux M, Angles T, Exarchakis G, Leonarduzzi R, Rochette G, Thiry L, Zarka J, Mallat S, Andén J, Belilovsky E, Bruna J, Lostanlen V, Chaudhary M, Hirn MJ, Oyallon E, Zhang S, Cella C and Eickenberg M (2020) Kymatio: Scattering transforms in python. *Journal of Machine Learning Research 21*(60), 1–6.

Arjovsky M, Bottou L, Gulrajani I and Lopez-Paz D (2019) Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arjovsky M, Chintala S and Bottou L (2017). Wasserstein Generative Adversarial Networks. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, Vol *70*. PMLR.

Arnaudo E, Blanco G, Monti A, Bianco G, Monaco C, Pasquali P and Dominici F (2023) A comparative evaluation of deep learning techniques for photovoltaic panel detection from aerial images. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2023.3275435

Ben-David S, Blitzer J, Crammer K and Pereira F (2006) Analysis of representations for domain adaptation. In Schölkopf B, Platt J and Hoffman T (eds.), *Advances in Neural Information Processing Systems*, Vol. *19*. MIT Press.

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, … Liang P (2022) On the opportunities and risks of foundation models. [arXiv:2108.07258 [cs]].

Bouaziz C, El Koundi M and Ennine G (2024) High-resolution solar panel detection in Sfax, Tunisia: A UNet-based approach. *Renewable Energy* 121171. https://doi.org/10.1016/j.renene.2024.121171

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, … Amodei D (2020) Language models are few-shot learners. [arXiv:2005.14165 [cs]]. https://doi.org/10.48550/arXiv.2005.14165

Bruna J and Mallat S (2013) Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 1872–1886. https://doi.org/10.1109/TPAMI.2012.230

Casanova A, Careil M, Verbeek J, Drozdzal M and Romero Soriano A (2021) Instance-Conditioned GAN. In Ranzato M, Beygelzimer A, Dauphin Y, Liang PS and Vaughan JW (eds.), *Advances in Neural Information Processing Systems*, Vol. *34*. Curran Associates, Inc, pp. 27517–27529

Chen Y, Ren Q and Yan J (2022) Rethinking and improving robustness of convolutional neural networks: A shapley value-based approach in frequency domain. In Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds.), *Advances in Neural Information Processing Systems*, Vol. *35*. Curran Associates, Inc, pp. 324–337

Cohen T and Welling M (2016) Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, 2990–2999.

Colas M and Saulnier E (2024). Means-Tested Solar Subsidies (CESifo Working Paper No. 11378). CESifo. https://doi.org/10.2139/ssrn.4991926

Csurka G (2017) A comprehensive survey on domain adaptation for visual applications. In Csurka G (ed.), *Domain Adaptation in Computer Vision Applications*. Springer International Publishing, pp. 1–35. https://doi.org/10.1007/978-3-319-58347-1_1

Csurka G, Volpi R and Chidlovskii B (2021) Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey. [arXiv:2112.03241 [cs]]. https://doi.org/10.48550/arXiv.2112.03241

Cubuk ED, Zoph B, Mane D, Vasudevan V and Le QV (2019) AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cubuk ED, Zoph B, Shlens J and Le QV (2020) RandAugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3008–3017. https://doi.org/10.1109/CVPRW50498.2020.00359

Dardouillet P, Benoit A, Amri E, Bolon P, Dubucq D and Credoz A (2023) Explainability of image semantic segmentation through SHAP values. In Rousseau J-J and Kapralos B (eds.), *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*. Springer Nature Switzerland, pp. 188–202. https://doi.org/10.1007/978-3-031-37731-0_19

De Jong T, Bromuri S, Chang X, Debusschere M, Rosenski N, Schartner C, Strauch K, Boehmer M and Curier L (2020). *Monitoring spatial sustainable development: Semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators*. arXiv preprint arXiv:2009.05738.

Fel T, Cadene R, Chalvidal M, Cord M, Vigouroux D and Serre T (2021) Look at the Variance! efficient black-box explanations with sobol-based sensitivity analysis. In Ranzato M, Beygelzimer A, Dauphin Y, Liang PS and Vaughan JW (eds.), *Advances in Neural Information Processing Systems*, Vol. *34*. Curran Associates, Inc, pp. 26005–26014.

Flora M, Potvin C, McGovern A and Handler S (2022) Comparing explanation methods for traditional machine learning models part 1: An overview of current methods and quantifying their disagreement. [arXiv:2211.08943 [physics, stat]].

Freitas S, Silva M, Silva E, Marceddu A, Miccoli M, Gnatyuk P, Marangoni L and Amicone A (2023) An artificial intelligence-based framework to accelerate data-driven policies to promote solar photovoltaics in Lisbon. *Solar RRL*, n/a(n/a), 2300597. https://doi.org/10.1002/solr.202300597

Frimane Â, **Johansson R**, **Munkhammar J**, **Lingfors D and Lindahl J** (2023) Identifying small decentralized solar systems in aerial images using deep learning. *Solar Energy 262*, 111822. https://doi.org/10.1016/j.solener.2023.111822

Gal Y (2016) Uncertainty in deep learning. Doctoral dissertation, University of Cambridge.

Ganin Y, **Ustinova E**, **Ajakan H**, **Germain P**, **Larochelle H**, **Laviolette F**, **Marchand M and Lempitsky V** (2016) Domain-adversarial training of neural networks. *The Journal of Machine Learning Research 17*(1), 2096–2030.

Geirhos R, **Rubisch P**, **Michaelis C**, **Bethge M**, **Wichmann FA. and Brendel W** (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Golovko V, **Kroshchanka A**, **Bezobrazov S**, **Sachenko A**, **Komar M and Novosad O** (2018) Development of solar panels detector. In *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, pp. 761–764. https://doi.org/10.1109/INFOCOMMST.2018.8632132

Goodfellow I, **Pouget-Abadie J**, **Mirza M**, **Xu B**, **Warde-Farley D**, **Ozair S**, **Courville A and Bengio Y** (2014) Generative adversarial nets. In Ghahramani Z, Welling M, Cortes C, Lawrence N and Weinberger KQ (eds.), *Advances in Neural Information Processing Systems*, Vol. *27*. Curran Associates, Inc.

Gorelick N, **Hancher M**, **Dixon M**, **Ilyushchenko S**, **Thau D and Moore R** (2017) Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment 202*, 18–27.

Graziano M and Gillingham K (2015) Spatial patterns of solar photovoltaic system adoption: The influence of neighbors and the built environment. *Journal of Economic Geography 15*(4), 815–839. https://doi.org/10.1093/jeg/lbu036

Guan H and Liu M (2022) Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering 69*(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3117407

Gulrajani I and Lopez-Paz D (2021) In search of lost domain generalization. In *International Conference on Learning Representations*.

Haegel NM, **Margolis R**, **Buonassisi T**, **Feldman D**, **Froitzheim A**, **Garabedian R**, **Green M**, **Glunz S**, **Henning H-M**, **Holder B**, et al (2017) Terawatt-scale photovoltaics: Trajectories and challenges. *Science 356*(6334), 141–143.

He K, **Zhang X**, **Ren S and Sun J** (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hendrycks D, **Mu N**, **Cubuk ED**, **Zoph B**, **Gilmer J and Lakshminarayanan B** (2020) AugMix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations*, ICLR 2020, *Addis Ababa, Ethiopia, April 26–30, 2020.*

Hendrycks D, **Zou A**, **Mazeika M**, **Tang L**, **Li B**, **Song D and Steinhardt J** (2022) PixMix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.48550/arXiv.2112.05135

Ho J, **Saharia C**, **Chan W**, **Fleet DJ**, **Norouzi M and Salimans T** (2022). Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research 23*(1), Article 47, 33 pages.

Hu W, **Bradbury K**, **Malof JM**, **Li B**, **Huang B**, **Streltsov A**, **Sydny Fujita K and Hoen B** (2022) What you get is not always what you see—pitfalls in solar array assessment using overhead imagery. *Applied Energy 327*, 120143. https://doi.org/10.1016/j.apenergy.2022.120143

IGN (2024) BD ORTHO® | Géoservices. https://geoservices.ign.fr/bdortho

Jansen MJW (1999) Analysis of variance designs for model output. *Computer Physics Communications 117*(1), 35–43. https://doi.org/10.1016/S0010-4655(98)00154-4

Kasmi G (2024) Enhancing the reliability of deep learning algorithms to improve the observability of french rooftop photovoltaic installations. Doctoral dissertation, Université Paris sciences et lettres.

Kasmi G, **Dubus L**, **Saint-Drenan Y-M and Blanc P** (2024). *Leveraging Artificial Intelligence to Improve the Integration of Photovoltaic Energy into the Grid*. The Transition Institute 1.5. https://doi.org/10.23646/9NEV-PY65

Kasmi G, **Dubus L**, **Saint-Drenan Y-M and Blanc P** (2022) Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping. In Corpetti T, Ienco D, Interdonato R, Pham M-T and Lefèvre S (eds.), *Proceedings of MACLEAN: MAChine Learning for EArth ObservatioN Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2022)*, *September 18–22, 2022*, Grenoble, France, Vol. *3343*. CEUR-WS.org.

Kasmi G, **Dubus L**, **Saint-Drenan Y-M and Blanc, P** (2023a) Can we reliably improve the robustness to image acquisition of remote sensing of PV systems? In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*.

Kasmi G, **Dubus L**, **Saint-Drenan Y-M and Blanc P** (2023b) Assessment of the reliablity of a model's decision by generalizing attribution to the wavelet domain. In *XAI in Action: Past, Present, and Future Applications Workshop at NeurIPS 2023*. https://doi.org/10.48550/arXiv.2305.14979

Kasmi G, **Saint-Drenan Y-M**, **Trebosc D**, **Jolivet R**, **Leloux J**, **Sarr B and Dubus L** (2023c) A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Scientific Data 10*(1), 59. https://doi.org/10.1038/s41597-023-01951-4

Kausika BB (2022) GIS4PV: A technological impact assessment of the application of GIS for Photovoltaic Solar Energy. Doctoral dissertation, Utrecht University. https://doi.org/10.33540/1371

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Kausika BB, Nijmeijer D, Reimerink I, Brouwer P and Liem V (2021) GeoAI for detection of solar photovoltaic installations in the Netherlands. *Energy and AI 6* 100111. https://doi.org/10.1016/j.egyai.2021.100111

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P and Girshick R (2023) Segment anything. [arXiv:2304.02643 [cs]]. https://doi.org/10.48550/arXiv.2304.02643

Koh PW, Sagawa S, Marklund H, Xie S M, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E, Stavness I, Guo W, Earnshaw B, Haque I, Beery SM, Leskovec J, Kundaje A, … Liang P (2021) WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664.

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller K-R (2019) Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications 10*(1), 1–8.

Li P, Zhang H, Guo Z, Lyu S, Chen J, Li W, Song X, Shibasaki R and Yan J (2021) Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning. *Advances in Applied Energy 4*, 100057. https://doi.org/10.1016/j.adapen.2021.100057

Lindahl J, Johansson R and Lingfors D (2023) Mapping of decentralised photovoltaic and solar thermal systems by remote sensing aerial imagery and deep machine learning for statistic generation. *Energy and AI* 100300. https://doi.org/10.1016/j.egyai.2023.100300

Lone AH and Siddiqui AN (2018) Noise models in digital image processing. *Global Science & Technology 10*(2), 63–66.

Lundberg SM and Lee S.-I (2017) A unified approach to interpreting model predictions. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. *30*. Curran Associates, Inc.

Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11*(7), 674–693. https://doi.org/10.1109/34.192463

Mallat S (1999) *A Wavelet Tour of Signal Processing*. Elsevier.

Malof JM, Bradbury K, Collins LM and Newell RG (2016) Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied Energy 183*, 229–240. https://doi.org/10.1016/j.apenergy.2016.08.191

Malof JM, Li B, Huang B, Bradbury K and Stretslov A (2019). Mapping solar array location, size, and capacity using deep learning and overhead imagery. CoRR, abs/1902.10895.

Malof JM, Rui Hou, Collins LM, Bradbury K and Newell R (2015) Automatic solar photovoltaic panel detection in satellite imagery. In *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*, pp. 1428–1431. https://doi.org/10.1109/ICRERA.2015.7418643

Mayer K, Rausch B, Arlt M-L, Gust G, Wang Z, Neumann D and Rajagopal R (2022) 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy 310*, 118469. https://doi.org/10.1016/j.apenergy.2021.118469

Mayer K, Wang Z, Arlt M-L, Neumann D and Rajagopal R (2020) DeepSolar for Germany: A deep learning framework for PV system mapping from aerial imagery. In *2020 International Conference on Smart Energy Systems and Technologies (SEST)*, pp. 1–6. https://doi.org/10.1109/SEST48500.2020.9203258

Murray J, Marcos D and Tuia D (2019) Zoom in, zoom out: Injecting scale invariance into landuse classification CNNs. In *IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5240–5243.

Parekh J (2023) Un cadre flexible pour l'apprentissage automatique interprétable: Application à la classification d'images et d'audio. Theses, Institut Polytechnique de Paris [Issue: 2023IPPAT032].

Parhar P, Sawasaki R, Todeschini A, Reed C, Vahabi H, Nusaputra N and Vergara F (2021) HyperionSolarNet: Solar panel detection from aerial images. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*.

Pearson K and Galton F (1895) VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Pena Pereira S, Rafiee A and Lhermitte S (2024) Automated rooftop solar panel detection through convolutional neural networks. *Canadian Journal of Remote Sensing 50*(1), 2363236. https://doi.org/10.1080/07038992.2024.2363236

Petsiuk V, Das A and Saenko K (2018) RISE: Randomized input sampling for explanation of black-box models. [arXiv: 1806.07421 [cs]]. https://doi.org/10.48550/arXiv.1806.07421

Pfungst O (1911) *Clever Hans: (The Horse of Mr. Von Osten.) A Contribution to Experimental Animal and Human Psychology.* Holt, Rinehart; Winston.

Pooch EHP, Ballester P and Barros RC (2020) CanWe trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification. In Petersen J, San José Estépar R, Schmidt-Richberg A, Gerard S, Lassen-Schmidt B, Jacobs C, Beichel R and Mori K (eds.), *Thoracic Image Analysis*. Springer International Publishing, pp. 74–83. https://doi.org/10.1007/978-3-030-62469-9_7

Rausch B, Mayer K, Arlt M-L, Gust G, Staudt P, Weinhardt C, Neumann D and Rajagopal R (2020) An enriched automated PV registry: Combining image recognition and 3D building data. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Recht B, Roelofs R, Schmidt L and Shankar V (2019) Do imagenet classifiers generalize to imagenet?. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400.

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer, B (2022) High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Ross AS, Hughes MC and Doshi-Velez F (2017) Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2662–2670.

RTE France (2022) *Energy Pathways to 2050* (Tech. Rep.). RTE France.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. and Fei-Fei L (2015) ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2020) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shen J, Qu Y, Zhang W and Yu Y (2018) Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence 32*(1). https://doi.org/10.1609/aaai.v32i1.11784

Simonyan K and Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In Bengio Y and LeCun Y (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.*

Sudjianto A and Zhang A (2021, November) Designing inherently interpretable machine learning models. [arXiv:2111.01743 [cs, stat]].

Sun B and Saenko K (2016) Deep CORAL: Correlation alignment for deep domain adaptation. In Hua G and Jégou H (eds.), *Computer Vision—ECCV 2016 Workshops*. Springer International Publishing, pp. 443–450. https://doi.org/10.1007/978-3-319-49409-8_35

Sun J, Mehra A, Kailkhura B, Chen P-Y, Hendrycks D, Hamm J and Mao ZM (2022a) A spectral view of randomized smoothing under common corruptions: Benchmarking and improving certified robustness. In Avidan S, Brostow G, Cissé M, Farinella GM and Hassner T (eds.), *Computer Vision—ECCV 2022*, Vol. *13664*. Springer Nature Switzerland, pp. 654–671. https://doi.org/10.1007/978-3-031-19772-7_38

Sun T, Segu M, Postels J, Wang Y, Van Gool L, Schiele B, TombariF and Yu F (2022b) SHIFT: A synthetic driving dataset for continuous multi-task Domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21371–21382.

Thimonier H, Popineau F, Rimmel A, Doan B-L and Daniel F (2024) Comparative evaluation of anomaly detection methods for fraud detection in online credit card payments. In *International Congress on Information and Communication Technology*, pp. 37–50.

Torralba A and Efros AA (2011) Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

Tuia D, Persello C and Bruzzone L (2016) Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine 4*(2), 41–57. https://doi.org/10.1109/MGRS.2016.2548504

Tuia D, Schindler K, Demir B, Zhu XX, Kochupillai M, Džeroski S, van Rijn JN, Hoos HH, Del Frate F, Datcu M, Markl V., Le Saux B., Schneider R and Camps-Valls G (2024) Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geoscience and Remote Sensing Magazine*, 2–25. https://doi.org/10.1109/MGRS.2024.3425961

Tzeng E, Hoffman J, Saenko K and Darrell T (2017) Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971. https://doi.org/10.1109/CVPR.2017.316

Vapnik V (1999) *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Wang R, Camilo J, Collins LM, Bradbury K and Malof JM (2017) The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: An empirical study with solar array detection. *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* 1–8. https://doi.org/10.1109/AIPR.2017.8457965

Wang Z, Arlt M-L, Zanocco C, Majumdar A and Rajagopal R (2022) DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models. *Joule 6*(11), 2611–2625. https://doi.org/10.1016/j.joule.2022.09.011

Wang Z, Bovik AC, Sheikh HR and Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*(4), 600–612.

Yang R, He G, Yin R, Wang G, Zhang Z, Long T and Peng Y (2024) Weakly-semi supervised extraction of rooftop photovoltaics from high-resolution images based on segment anything model and class activation map. *Applied Energy 361*, 122964. https://doi.org/10.1016/j.apenergy.2024.122964

Yu J, Wang Z, Majumdar A and Rajagopal R (2018) DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. *Joule 2*(12), 2605–2617. https://doi.org/10.1016/j.joule.2018.11.021

Yuan J, Yang H-HL, Omitaomu OA and Bhaduri BL (2016). Large-scale solar panel mapping from aerial images using deep convolutional networks. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2703–2708. https://doi.org/10.1109/BigData.2016.7840915

Zech M and Ranalli J (2020) Predicting PV areas in aerial images with deep learning. In *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, pp. 0767–0774. https://doi.org/10.1109/PVSC45281.2020.9300636

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

Zhang C, **Bengio S**, **Hardt M**, **Recht B and Vinyals O** (2017) Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*.

Zhang Y, **Hong D**, **McClement D**, **Oladosu O**, **Pridham G and Slaney G** (2021) Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods 353*, 109098. https://doi.org/10.1016/j.jneumeth.2021.109098

Zhou K, **Liu Z**, **Qiao Y**, **Xiang T and Loy CC** (2023) Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3195549

Zhou X, **Lin Y**, **Zhang W and Zhang T** (2022) Sparse invariant risk minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 27222–27244. https://proceedings.mlr.press/v162/zhou22e.html

A. Limitations of the GradCAM and related feature attribution methods for our use case

Figure A1 presents the explanations obtained using the GradCAM (Selvaraju et al., 2020). We can see two different prediction patterns depending on whether the model predicts a positive (true or false) or a negative (true or false). In the case of a true positive prediction, the model will focus on a specific, narrow region of the image, which corresponds to a PV panel. However, for false positives, the model also focuses on a narrow image region. Inspecting the samples of Figure A1 reveals that this region of the image depicts items that *resemble* PV panels. In the image on the first row (second column) of Figure A1, we can see that the model confuses a shade house that shares the same color and overall shape as a PV panel with an actual panel. In the image on the second row, the verandas with groves fool the model.

On the other hand, when the model does not see a PV panel, it does not focus on a specific image region. This remains true for the false negatives, where we can see that the model does not see the panels on any of the images.

However, we can also see that as the GradCAM only assesses where the model is looking, it is challenging to understand why it focused on a given area that resembles a PV panel on false positives and why it did not identify the PV panel on the false negatives. Achtibat et al. (2022) underline the necessity for reliable model evaluation to assess where models are looking at and *what* they are looking at on input images. The choice of the WCAM as an attribution method and, more broadly, the space-scale decomposition attempts to address this question by assessing the scales the models consider when making their predictions.

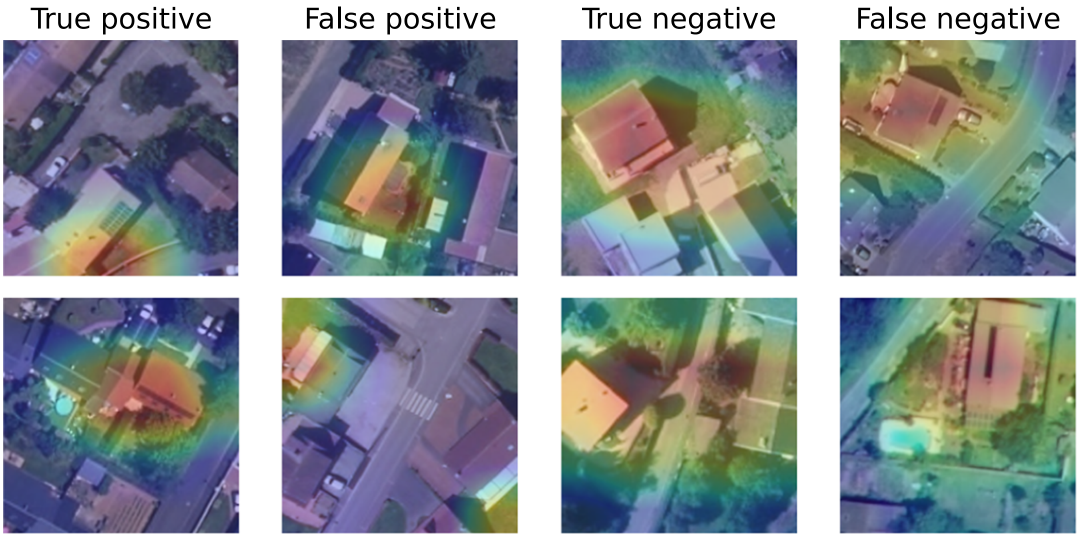

Figure A1. *Model explanations using the GradCAM (Selvaraju et al., 2020) for some true positives, false positives, true negatives, and false negatives. The redder, the higher the contribution of an image region to the predicted class (1 for true and false positives, and 0 for true and false negatives).*

B. Computation of the WCAM (from Kasmi et al., 2023b)

Figure B1 depicts the principle of the WCAM. The importance of the regions of the wavelet transform of the input image is estimated by **(1)** generating masks from a Quasi-Monte Carlo sequence, **(2)** evaluating the model on perturbed images. We obtain these images

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

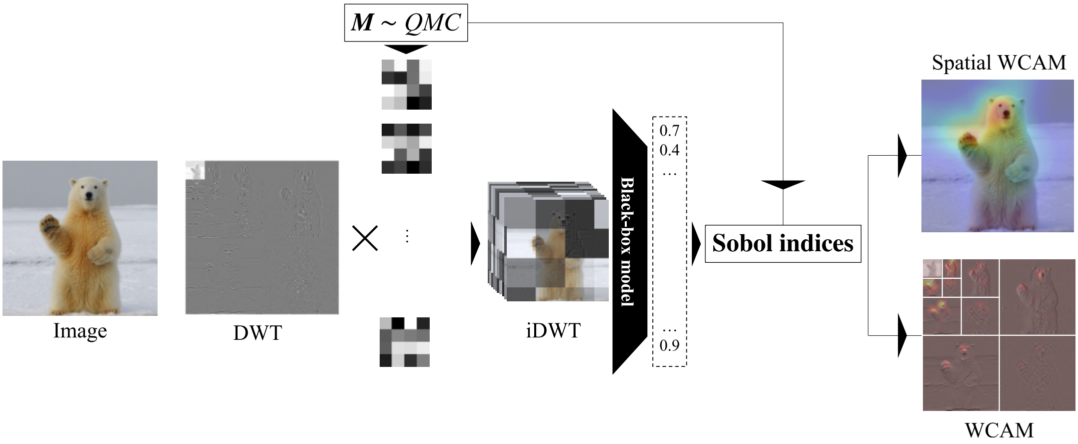

Figure B1. *Flowchart of the wavelet scale attribution method (WCAM). Source: Kasmi et al., 2023b.*

by computing the discrete wavelet transform (DWT) of the original image, applying the masks on the DWT to obtain perturbed DWT, and inverting the perturbed DWT to generate perturbed images. On an RGB image, we apply the DWT channel-wise and apply the same perturbation to each channel. We generate $N(K+2)$ perturbed images for a single image, and **(3)** We estimate the total Sobol indices of the perturbed regions of the wavelet transform using the masks and the model's outputs using Jansen's estimator (Jansen, 1999). Fel et al. (2021) introduced this approach to estimate the importance of image regions in the pixel space. We generalize it to the wavelet domain.

C. Quantitative relationship between the WCAM's scale embeddings and the model's response to distribution shifts

Definition. A *scale embedding* is a vector $z=(z_1,\ldots,z_L)\in\mathbb{R}^L$ where each component z_s encodes the importance of the l^{th} scale component in the prediction.

Scale embeddings compute the importance of each scale and each direction and summarize it into a vector $z\in\mathbb{R}^L$ where L indicates the number of levels. In our case, we have ten levels (1 corresponding to the approximation coefficients and $L=3\times 3$ corresponding to the three scales of details coefficients and their three respective orientations. Scale embeddings summarize the importance of each scale, irrespective of the spatial localization of importance.

Results. We computed the distance (measured by the Euclidean distance) between the two images' scale embeddings and computed the correlation between this distance and the predicted probability shift. As a baseline, we also computed the distance between the two WCAMs.

We obtained correlation coefficients of 0.18 $(p=0.19)$ for the scale embedding and 0.17 $(p=0.19)$ for the raw WCAM. Although weaker than the correlation between the distortion and the predicted probability shift, this result highlights that the WCAM consistently captures the change in behavior of the model resulting from the shift in acquisition conditions.

D. Overview of the data augmentation strategies

D.1. Description of the data augmentations

AugMix (Hendrycks et al., 2020). The data augmentation strategy "Augment-and-Mix" (AugMix) consists of producing a high diversity of augmented images from an input sample. A set of operations (perturbations) to be applied to the images are sampled, along with sampling weights. The image resulting x_{aug} is obtained through the composition $x_{aug}=\omega_1 op_1\circ\ldots\omega_n op_n(x)$ where x is the original image. Then, the augmented image is interpolated with the original image with a weight m that is also randomly sampled. We have $x_{augmix}=mx+(1-m)x_{aug}$.

AutoAugment (Cubuk et al., 2019). This strategy aims at finding the best data augmentation for a given dataset. The authors determined the best augmentation strategy S as the outcome of a reinforcement learning problem: a controller predicts an augmentation policy from a search space. Then, the authors train a model, and the controller updates its sampling strategy S based on the train loss. The goal is for the controller to generate better policies over time. The authors derive optimal augmentation

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

strategies for various datasets, including ImageNet (Russakovsky et al., 2015), and show that the optimal policy for ImageNet generalizes well to other datasets.

RandAugment (Cubuk et al., 2020). This strategy's primary goal is to remove the need for a computationally expansive policy search before model training. Instead of searching for transformations, random probabilities are assigned to the transformations. Then, each resulting policy (a weighted sequence of K transformations) is graded depending on its strength. The number of transformations and the strength are passed as input when calling the transformation.

D.2. Plots

Figure D1 plots examples of the different data augmentations implemented in this work. Along with these augmentations, we apply random rotations, symmetries, and normalization to the input during training. At test time, we only normalize the input images.

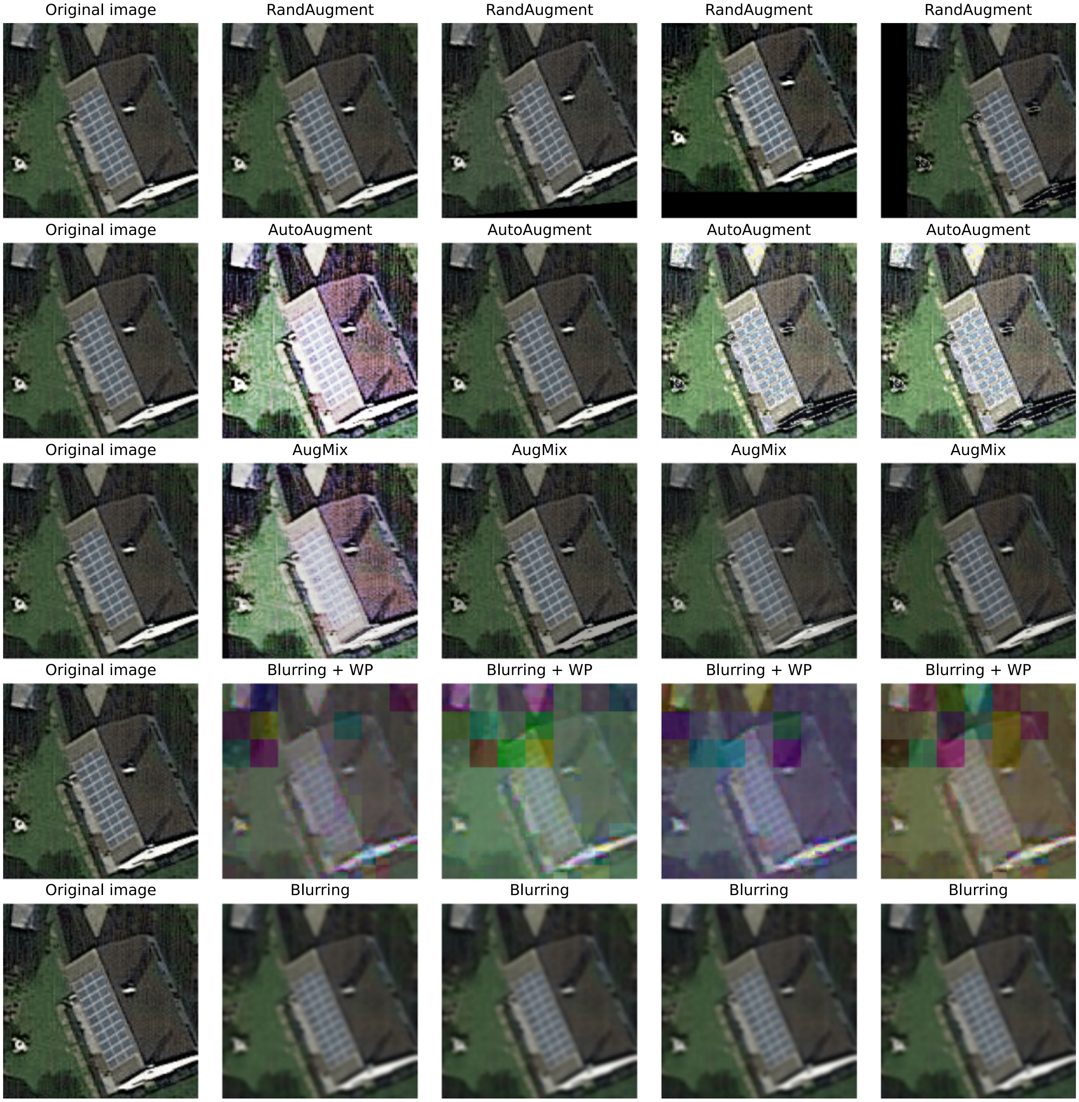

Figure D1. *Visualization of the different data augmentation techniques implemented in this work.*

E. Evaluation of domain adaptation techniques

E.1. Overview of the selected methods

We selected various popular unsupervised domain adaptation (UDA) methods. The common point between these methods is that they aim to learn a domain invariant representation using labeled samples from the source domain S (in our case, Google images) and unlabeled samples from the target domain T (in our case, IGN images). The central difference with our approach is that these UDA approaches require unlabeled samples from the target domain, which is not the case with data augmentation strategies.

DeepCORAL (Sun and Saenko, 2016). DeepCORAL (CORrelation ALignment) expands CORAL to deep neural networks. The original CORAL framework consists of aligning the source and target domain's distributions by aligning their second-order statistics. Denoting S and T the source and target domains respectively and $C_{\cdot} \in \mathbb{R}^{d\times d}$ denotes the covariance matrix of the features of dimension d. The CORAL Loss is then defined as

$$\mathcal{L}_{\text{CORAL}} = \frac{1}{4d^2}\|C_S - C_T\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The CORAL loss is added as a penalty term in the target loss of the model. Denoting $\mathcal{L}_{CLF}$ the loss of a classification model (e.g., the classification loss) derived from the source dataset, the loss of the model is modified as

$$\mathcal{L} = \mathcal{L}_{CLF} + \lambda\mathcal{L}_{CORAL}.$$

DeepCORAL (Sun and Saenko, 2016) adapts this framework by aligning the covariance matrices of the feature representation matrix $Z_{\cdot}$ retrieved from a deep learning encoder: $Z = \Phi(X)$ where X corresponds to the input data and Φ denotes the feature extractor of the deep learning model. The dimensionality d then denotes the dimensionality of the model's latent space rather than the input space's dimensionality.

Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017). ADDA is based on the generative adversarial networks Goodfellow et al. (2014). It aims to learn a representation that is aligned between the source and target domains. To do so, given a feature extractor or encoder Φ_S trained on the source domain (source feature extractor), an adversarial game between a discriminator D and an encoder trained on the target domain (target feature extractor) Φ_T is set up to train the target feature extractor to generate features from the target domain that are undistinguishable with the features generated by the source feature extractor. The *domain* $\mathcal{L}_d$ is the combination of two components, the loss of the discriminator $\mathcal{L}_d^D$ and the loss of the feature extractor $\mathcal{L}_d^{\Phi_T}$, where

$$\mathcal{L}_d^D = -\mathbb{E}_S[\log(D(\Phi_S(X_S)))] - \mathbb{E}_T[\log(1 - D(\Phi_T(X_T)))]$$

where S and T, with a slight abuse of notation, denote the distributions of the source and target domains, respectively. The discriminator D indicates whether the feature representation comes from the source or the target extractor. The loss of the target feature extractor is

$$\mathcal{L}_d^{\Phi_T} = -\mathbb{E}_T[\log(D(\Phi_T(X_T)))]$$

Combining the losses we get $\mathcal{L}_d = \mathcal{L}_d^D + \mathcal{L}_d^{\Phi_T}$ and finally,

$$\mathcal{L}_{ADDA} = \mathcal{L}_s + \lambda\mathcal{L}_d$$

where $\mathcal{L}_s$ denotes the source supervised loss, i.e., the ERM on the source domain to train the source feature extractor. The adversarial game is formulated as $\min_{\Phi_T}\max_D \mathcal{L}_d$.

Unsupervised domain adaptation by backpropagation (RevGrad) (Ganin et al., 2016). RevGrad, like ADDA, aims at learning a representation that is aligned across domains through adversarial training. Unlike ADDA, the feature extractor Φ is shared across domains. In addition, RevGrad uses the Gradient Reversal Layer (GRL) during training. The GRL reverses the gradient during backpropagation when computing the domain loss. This layer enables the feature extractor to learn domain-invariant features by making adversarial training more effective.

Wasserstein Distance Guided Representation Learning (WDGRL) (Shen et al., 2018). This method is based on the Wasserstein GANs (Arjovsky et al., 2017), which use the Wasserstein distance to measure the difference between the generated and real data distributions. Unlike the standard GAN, which uses the Jensen-Shannon divergence, Wasserstein GANs (WGAN) provide a more stable training process and avoid issues like mode collapse. It uses a critic instead of a discriminator and enforces a 1-Lipschitz constraint on the critic using a gradient penalty to ensure proper convergence. This approach leads to more meaningful and smooth loss gradients to improve the generator. WDGRL builds on WGAN to learn a domain invariant feature representation. A critic C replaces the discriminator D to discriminate between the source and target domain-based representations.

E.2. Benchmark details

Overview. We implemented the four methods described above in an unsupervised domain adaptation (UDA) setting. During training, we assume we have access to the samples and the labels of the source dataset (i.e., Google images) and only to the samples of the target dataset (i.e., IGN). We only had to assume we had access to the source domain data (samples and labels) for the data augmentation techniques. We opted for the UDA setting as it is the closest setting to the one used for evaluating the data augmentation techniques. It is also the easiest to implement in practice as by definition, when deploying a model on new data, we have by definition access to this data, although without labels.

On the other hand, domain generalization often requires multiple source domains. As in our setting, we only have one source domain to train our model on, so we discarded these methods. Our implementation of DeepCORAL is based on the repository accessible at this URL https://github.com/DenisDsh/PyTorch-Deep-CORAL and our implementation of ADDA, RevGrad, and WDGRL is based on the repository accessible at this URL https://github.com/jvanvugt/pytorch-domain-adaptation. The trained model weights and the source code to replicate our results are accessible on our Git repository.

Implementational details. Our approach is the following: We trained four UDA methods on labeled Google images and unlabeled IGN images. We evaluate the performance of the models on IGN images with our usual metrics, namely the F1 score, and report the associated true positives, false positives, true negatives, and false negatives rates. Table F1 presents the accuracy results on Google images

During training, we looked for optimal parameters for DeepCORAL and WDGRL. This was done through a grid search. For DeepCORAL, we searched for the optimal learning rate, momentum, and weight for the CORAL term in the loss λ_{CORAL}. For WDGRL, we looked for the optimal parameters γ, which controls the weight of the gradient penalty term in the critic loss, K_{CLF}, which controls for the number of iterations for training the classifier in each training step and WD_{CLF}, which controls the weight of the Wasserstein distance in the classifier loss.

E.3. Results

E.3.1. Quantitative results

Table E1 presents the evaluation results of the domain adaptation methods on our benchmark. We reproduced the results of the data augmentation methods for completeness and to ease the comparisons.

Judging solely according to the F1 score, we can see that our data augmentation techniques match or surpass the performance of the domain adaptation techniques while requiring less information as *no* information on the target domain is required. In detail, however, we can see that the UDA methods, especially ADDA, outperform our method, especially as it achieves a higher true positive rate and a lower false negative rate. On the other hand, our Blurring + WP method's performance is in line with DeepCORAL.

Table E1. F1 Score *and decomposition in true positives, true negatives, false positives, and false negatives rate for models trained on Google with different mitigation strategies. Evaluation of IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. Best results are* ***bolded****, second-best results are underlined, values highlighted in red indicate the worst performance, and values in orange indicate the second-to-last worst performance*

	Model	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
	Oracle	0.88	0.96	0.82	0.18	0.04
Augmentations	None (ERM)	0.44	0.30	0.96	0.04	0.70
	AutoAugment	0.46	0.31	0.96	0.04	0.69
	AugMix	0.48	0.33	0.96	0.04	0.67
	RandAugment	0.51	0.37	0.94	0.06	0.63
Adaptation	DeepCoral	0.54	0.54	0.64	0.36	0.46
	ADDA	0.61	0.95	0.09	0.91	0.05
	WDGRL	0.66	0.58	0.86	0.14	0.42
	RevGrad	0.30	0.18	**0.98**	**0.02**	0.82
	Blurring (Ours)	**0.74**	**0.98**	0.49	0.51	**0.02**
	Blurring + WP (Ours)	0.58	0.47	0.87	0.13	0.53

E.3.2. Qualitative analysis with the WCAM

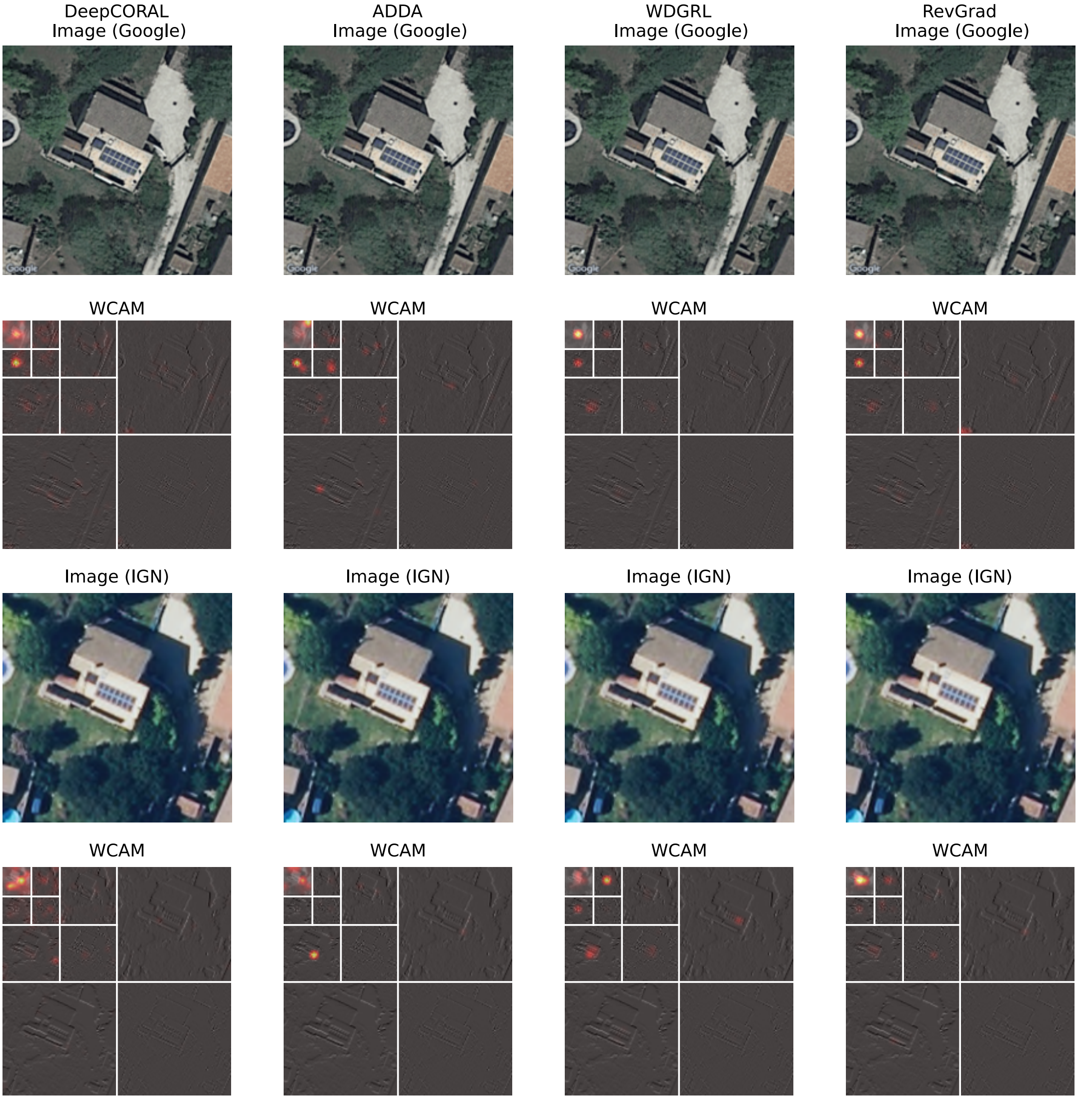

Figure E1. *Evaluation of the different domain adaptation methods with the WCAM. Each column represents a column. The first and third rows depict the images from Google and IGN, respectively, and the second and fourth rows are the associated WCAMs.*

E.4. Discussion and limitations

Our results show that the data augmentation methods can achieve performance that matches some popular domain adaptation techniques while being easier to implement in practice *and* requiring less information as no information on the target domain is required. However, UDA methods, and especially WDGRL, remain more reliable as their false negative rate is lower than the false negative rate of our approach.

This benchmark, however, is limited by the fact that the methods evaluated here are relatively old. More recent methods, such as Invariant Risk Minimization (Arjovsky et al., 2019) or methods featured in DomainBed (Gulrajani and Lopez-Paz, 2021) do not scale very well to architectures as large as ResNets, so we discarded them.

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

F. Complementary results

F.1. Accuracy results of the mitigation methods on Google images

Table F1 displays the accuracy results of the models trained with various data augmentation and domain adaptation strategies on the source domain (i.e., Google images).

Table F1. ***F1 Score*** *and decomposition in true positives, true negatives, false positives, and false negatives rate for models trained on Google with different mitigation strategies. Evaluation of Google images*

	Model	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
Augmentations	None (ERM)	0.98	0.98	0.98	0.02	0.02
	AutoAugment	0.98	0.99	0.98	0.02	0.01
	AugMix	0.98	0.98	0.98	0.02	0.02
	RandAugment	0.98	0.99	0.98	0.02	0.01
Adaptation	DeepCoral	0.67	1.00	0.19	0.81	0.00
	ADDA	0.62	0.99	0.03	0.97	0.01
	WDGRL	0.97	0.99	0.95	0.05	0.01
	RevGrad	0.97	0.96	0.98	0.02	0.04
	Blurring (Ours)	0.82	0.85	0.82	0.18	0.15
	Blurring + WP (Ours)	0.90	0.93	0.89	0.11	0.07

F.2. Accuracy results for variants of the Scattering transform

Table F2 presents the accuracy of the Scattering transform for two depth variants (labeled $m = 1$ and $m = 2$). We can see that the performance of the Scattering transform remains relatively poor regardless of the depth of the scattering coefficients. Contrary to the claims of Bruna and Mallat, 2013, including second-order coefficients does not seem enough to discriminate between images, as the number of false positives remains high. This could be caused by the fact that our task, namely the detection of small objects on orthoimagery, is more challenging than digit classifications.

Table F2. ***F1 Score*** *and decomposition in true positives, true negatives, false positives, and false negative rate of the classification accuracy of the Scattering Transform model trained on Google images and deployed on IGN images*

Depth	Dataset	F1 score (↑)	True positive rate (↑)	True negative rate (↑)	False positive rate (↓)	False negative rate (↓)
$m = 1$	*Google baseline*	0.57	0.84	0.09	0.57	0.57
	IGN	0.57	0.71	0.40	0.52	0.36
$m = 2$	*Google baseline*	0.57	0.89	0.10	0.56	0.48
	IGN	0.59	0.54	0.31	0.62	0.54
ERM	*Google baseline*	0.98	0.99	0.98	0.02	0.01
	IGN	0.46	0.32	0.95	0.03	0.68
Random classifier	*Google baseline*	0.47	0.5	0.50	0.55	0.45
	IGN	0.47	0.50	0.50	0.56	0.44

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

G. Additional figures

G.1. Assessment of the effects of distribution shifts on the model's predictions

Figures G1 to G3 present additional examples of qualitative assessment of the effects of distribution shifts on the model's prediction. In Figure G1, we can see that the model initially primarily relied on the gridded pattern, which is discernible at the 4–8 pixel scale. The acquisition conditions discarded this factor, thus explaining why the model could no longer recognize the PV panel. A similar phenomenon occurs in Figure G2. Figure G3 presents an example of a prediction not affected by the acquisition conditions. We can see that the important scales (especially at the 4–8 pixel scale) remain the same.

G.1.1. Comparison of the behavior of the data augmentation methods on IGN images

Figure G4 compares some data augmentation techniques' behavior on an image from the IGN dataset.

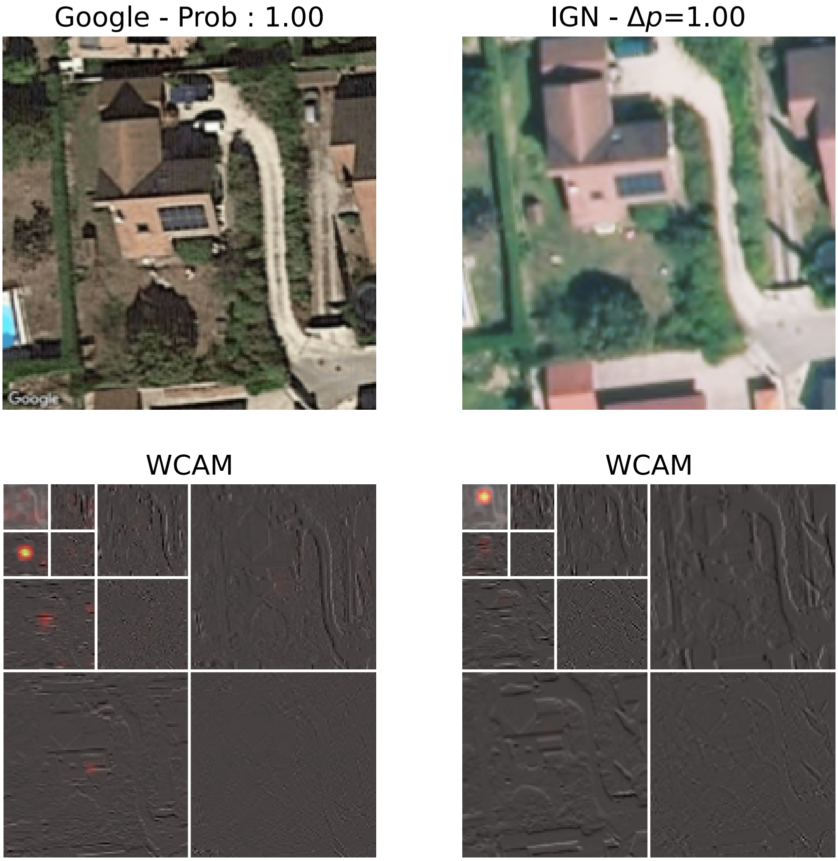

Figure G1. *Analysis with the WCAM of the CNNs prediction on an image no longer recognized as a PV panel.*

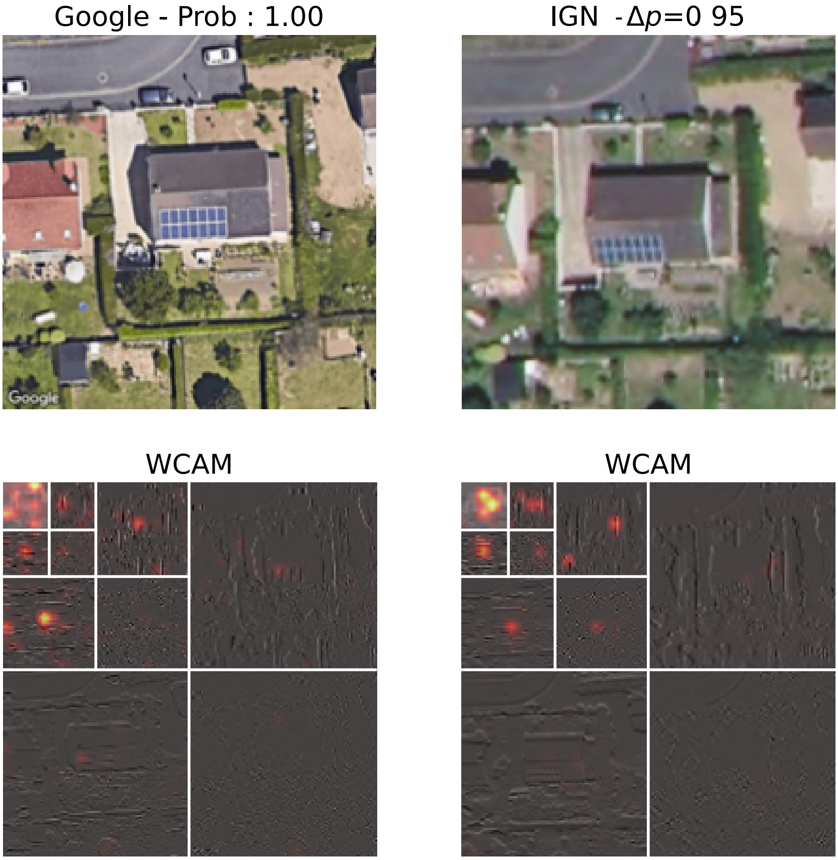

Figure G2. *Analysis with the WCAM of the CNNs prediction on an image no longer recognized as a PV panel.*

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

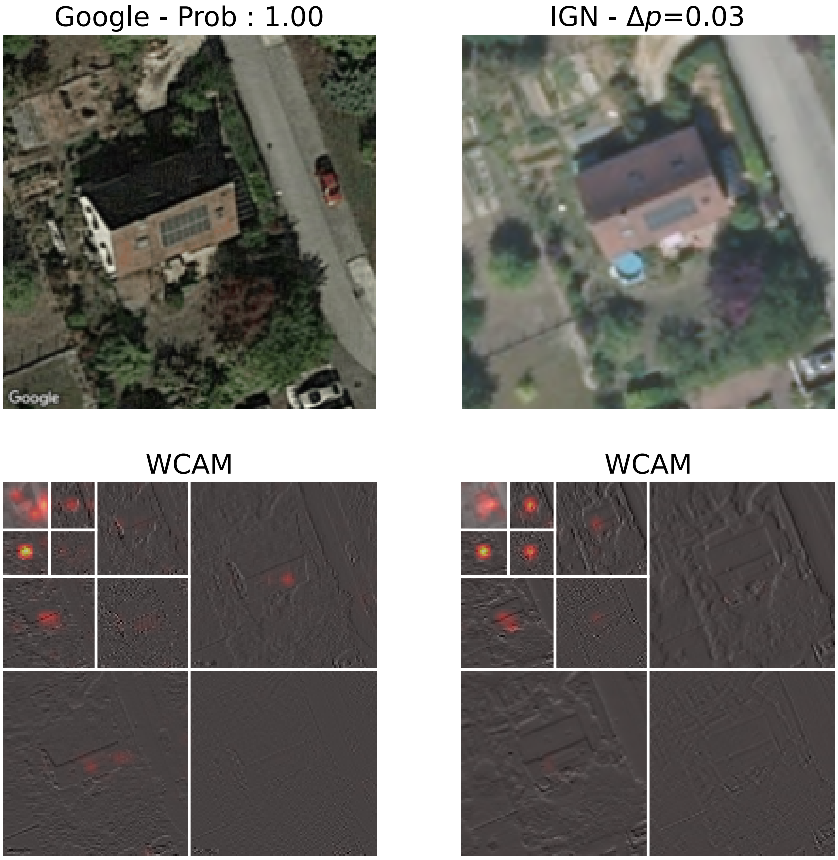

Figure G3. *Analysis with the WCAM of the CNNs prediction on an image that remains insensitive to varying acquisition conditions.*

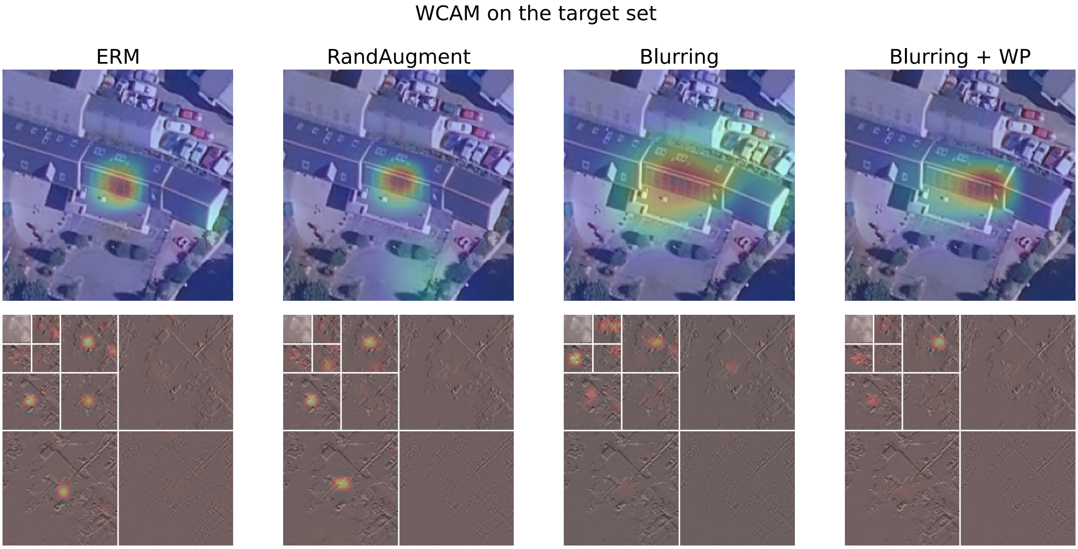

Figure G4. *WCAMs on IGN of models trained on Google with different augmentation techniques.*

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

G.2. Effect of the distribution shifts on the domain adaptation methods

Figure G5 and G6 plot additional examples of the effect of the varying acquisition conditions on the domain adaptation methods evaluated in this work.

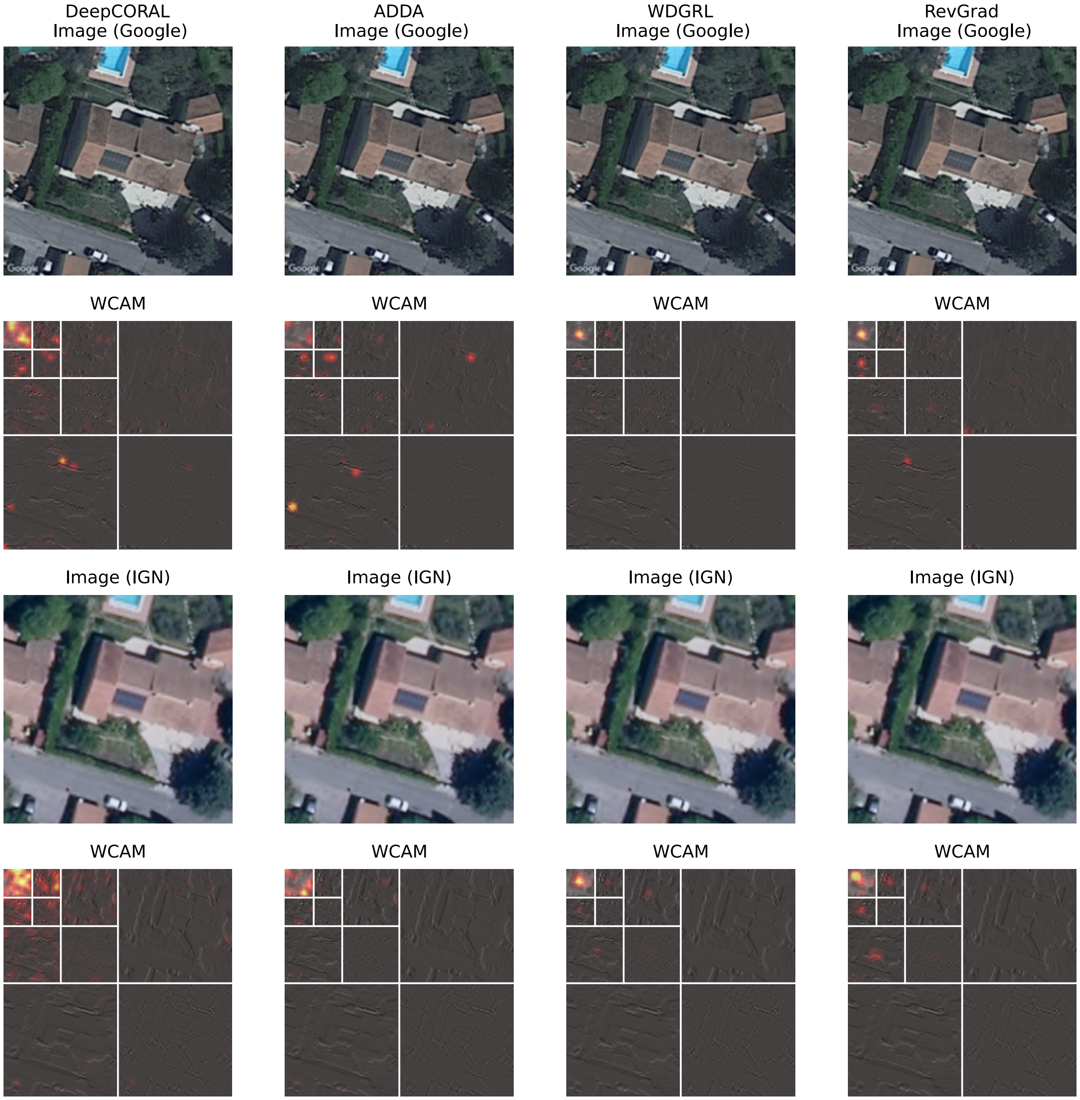

Figure G5. *Evaluation of the different domain adaptation methods with the WCAM. Each column represents a column. The first and third rows depict the images from Google and IGN, respectively, and the second and fourth rows are the associated WCAMs.*

https://doi.org/10.1017/eds.2025.13 Published online by Cambridge University Press

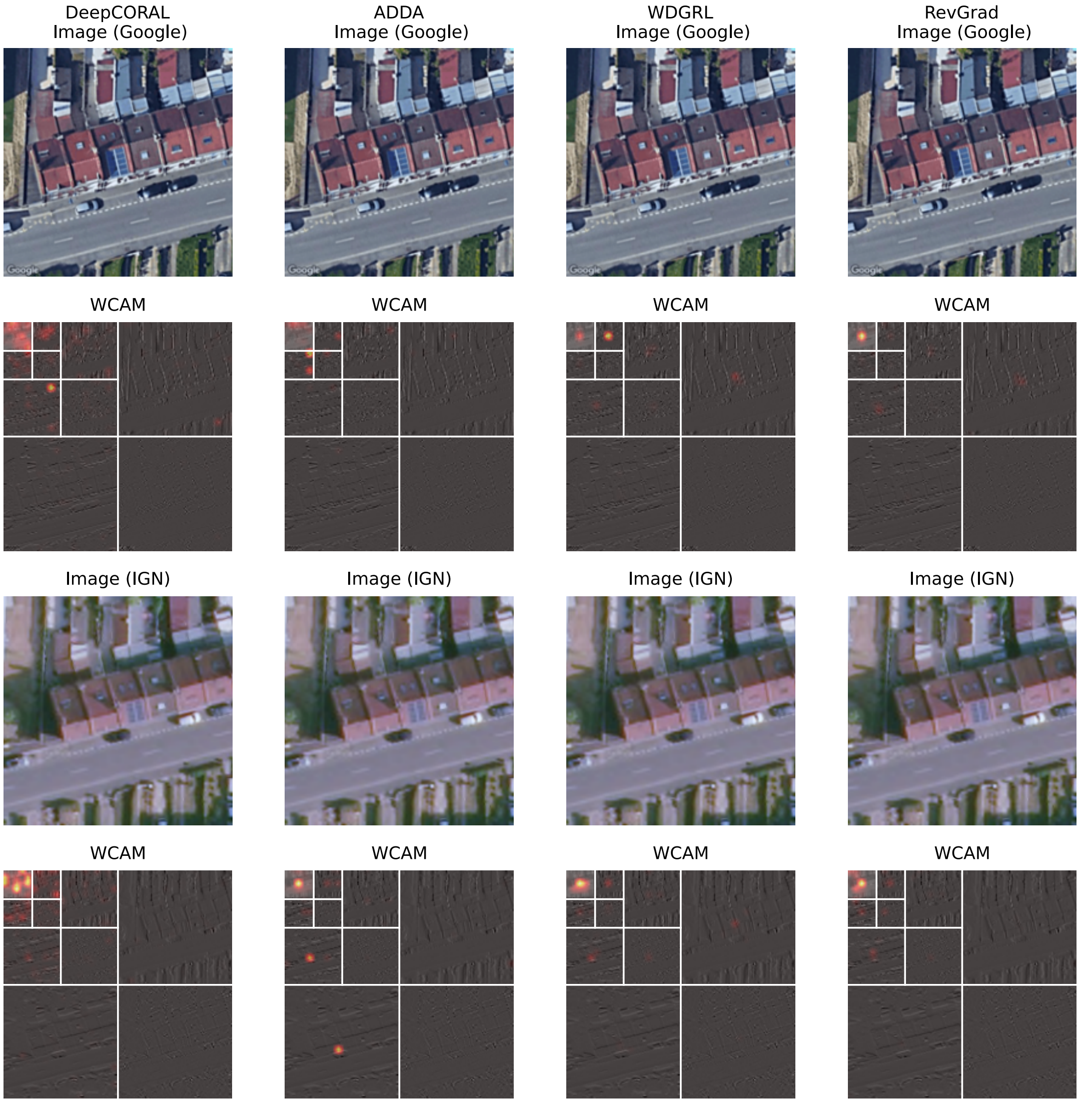

Figure G6. *Evaluation of the different domain adaptation methods with the WCAM. Each column represents a column. The first and third rows depict the images from Google and IGN, respectively, and the second and fourth rows are the associated WCAMs.*

Cite this article: Kasmi G, Dubus L, Saint-Drenan Y.-M. and Blanc P (2025). Space-scale exploration of the poor reliability of deep learning models: the case of the remote sensing of rooftop photovoltaic systems. *Environmental Data Science*, 4: e22. doi:10.1017/eds.2025.13