



OPEN

DATA DESCRIPTOR

A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata

Gabriel Kasmi^{1,2,5}✉, Yves-Marie Saint-Drenan^{1,5}, David Trebosc^{3,5}, Raphaël Jolivet^{1,5}, Jonathan Leloux⁴, Babacar Sarr⁴ & Laurent Dubus^{1,2}

Photovoltaic (PV) energy generation plays a crucial role in the energy transition. Small-scale, rooftop PV installations are deployed at an unprecedented pace, and their safe integration into the grid requires up-to-date, high-quality information. Overhead imagery is increasingly being used to improve the knowledge of rooftop PV installations with machine learning models capable of automatically mapping these installations. However, these models cannot be reliably transferred from one region or imagery source to another without incurring a decrease in accuracy. To address this issue, known as distribution shift, and foster the development of PV array mapping pipelines, we propose a dataset containing aerial images, segmentation masks, and installation metadata (i.e., technical characteristics). We provide installation metadata for more than 28000 installations. We supply ground truth segmentation masks for 13000 installations, including 7000 with annotations for two different image providers. Finally, we provide installation metadata that matches the annotation for more than 8000 installations. Dataset applications include end-to-end PV registry construction, robust PV installations mapping, and analysis of crowdsourced datasets.

Background & Summary

In 2021, photovoltaic (PV) power generation amounted to 821 TWh worldwide and 14.3 TWh in France¹. With an installed capacity of about 633 GW_p worldwide² and 13.66 GW_p in France, PV energy represents a growing share of the energy supply. The integration of growing amounts of solar energy in energy systems requires an accurate estimation of the produced power to maintain a constant balance between electricity demand and supply. However, small-scale PV installations are generally invisible to transmission system operators (TSOs), meaning their generated power is not monitored³. For TSOs, the lack of reliable rooftop PV measurements increases the flexibility needs, i.e., the means for the grid to compensate for load or supply variability^{4–7}. Estimating the PV power generation from meteorological data is common practice to overcome the lack of power measurements. However, this requires precise information on PV installed capacity and installation metadata^{8,9}. Detailed information regarding small-scale PV is therefore of interest for integrating renewable energies into the grid¹⁰, or for understanding the factors behind its development¹¹.

Currently, such PV installation registries covering large areas are neither easily available nor available everywhere. Recent research to construct global PV inventories^{12,13} is limited to solar farms and does not include rooftop PV. A recent crowdsourcing effort allowed the mapping of 86% of the United Kingdom's rooftop PV installations¹⁴. Other available datasets are aggregated at the communal scale (census level)¹⁵.

Remote sensing-based methods^{13,15–17} recently emerged as a promising solution to quickly and cheaply acquire detailed information on PV installations¹⁸. These methods rely on overhead imagery and deep neural networks. The DeepSolar initiative led to the mapping of rooftop PV installations over the continental

¹Mines Paris - PSL University, Centre Observation, Impacts, Energy (O.I.E.), 06904, Sophia Antipolis, France.

²RTE France, 7C place du Dôme, 92073, Paris La Défense, France. ³BDPV, 1 Rue du Capitaine Fracasse, 31320, Castanet Tolosan, France. ⁴LuciSun, Rue Saint-Jean, 29, Sart-Dames-Avelines, Villers-la-Ville, Belgium. ⁵These authors contributed equally: Gabriel Kasmi, Yves-Marie Saint-Drenan, David Trebosc, Raphaël Jolivet. ✉e-mail:

gabriel.kasmi@minesparis.psl.eu

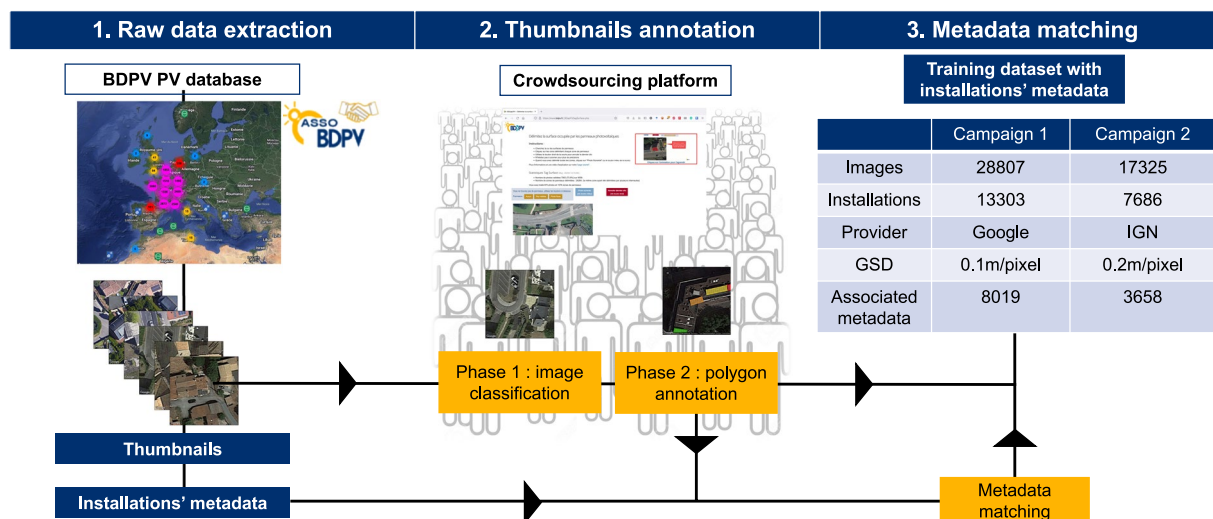


Fig. 1 Flowchart of the training dataset generation based on the BDPV PV data and crowdsourcing. “GSD” stands for the ground sampling distance, i.e., the distance between the centers of two adjacent pixels measured on the ground.

United-States¹⁵ or the state of North-Rhine Westphalia¹⁹. These remote sensing-based methods cannot scale to unseen regions without a sharp decrease in accuracy^{20,21}. It is caused by the sensitivity of deep learning models to distribution shifts²² (i.e., when “the training distribution differs from the test distribution”²³). These distribution shifts typically correspond to acquisition conditions and architectural differences across regions²⁴ (e.g., building shape, vegetation). The lack of robustness to distribution shifts limits the reliability of deep learning-based registries for constructing official PV statistics¹⁰. Therefore, developing PV mapping algorithms that are robust to distribution shifts is necessary.

To encourage the development of such algorithms, we introduce a training dataset containing data for (i) addressing distribution shifts in remote sensing applications and (ii) helping design algorithms capable of extracting small-scale PV metadata from overhead imagery.

To address distribution shifts, we gathered ground-truth annotations from two image providers for installations located in France. The double annotation allows researchers to evaluate the robustness of their approach to a shift in data provider (which affects acquisition conditions, acquisition device, and ground sampling distance) while keeping the same observed object. Our dataset provides ground truth installation masks for 13303 images from Google Earth²⁵ and 7686 images from the French national institute of geographical and forestry information (IGN). To address architectural differences, researchers can either use the coarse-grained location included in our dataset or use our dataset in conjunction with other training datasets that mapped different areas (e.g., Bradbury *et al.*²⁶ or Khomiakov *et al.*²⁷).

To extract PV systems’ metadata, we release the installation metadata for 28807 installations. This metadata includes installed capacity, surface, tilt, and azimuth angles, sufficient for regional PV power estimation⁸. We linked the installation metadata and the ground truth images for 8019 installations. To the best of our knowledge, it is the first time a training dataset contains PV panel images, ground truth labels, and installation metadata. We hope this data contributes to the ongoing effort to construct more detailed PV registries.

We obtained our labels through two crowdsourcing campaigns conducted in 2021 and 2022. Crowdsourcing is common practice in the machine learning community for annotating training datasets^{28,29}. We developed our crowdsourcing platform and collected up to 50 annotations per image to maximize the accuracy of our annotations. Besides, multiple annotations per image facilitate measurement of the annotator’s agreement or limit their individual annotations biases. Indeed, we found that some annotators were more cautious when annotating than others. We make the raw crowdsourcing data publicly available. It allows the replication of our annotations, but we also hope this will help research crowdsourcing, e.g., on the efficient combination of labels³⁰.

Our dataset targets practitioners and researchers in machine learning and crowdsourcing. Our data can serve as training data for remote PV mapping algorithms and test machine learning models’ robustness against acquisition conditions or geographical shift. Additionally, we release the raw annotation data from the crowdsourcing campaigns for the community to carry out further studies on the fusion of multiple annotations into ground truth labels. The training dataset and the data coming from the crowdsourcing campaigns are accessible on our Zenodo repository³¹.

Methods

We illustrate our training dataset generation workflow in Fig. 1. It comprises three main steps: raw data extraction, thumbnails annotation, and metadata matching.

Raw data extraction. Our annotation campaign leverages the database of PV systems operated by the non-profit association Asso BDPV (*Base de données Photovoltaïque* - Photovoltaic database). Asso BDPV (BDPV)

	Google	IGN
Total number of actions	349394	119528
Total number of annotators	1901	1021
Actions during phase 1	291597	90084
Actions during phase 2	68162	29444
Active annotators during phase 1	1043	51
Active annotators during phase 2	960	980
Active annotators during both phases	102	10

Table 1. Summary statistics of the contributions during the crowdsourcing campaigns.

gathers metadata (geolocation and metadata of the PV systems) and the energy production data of PV installations provided by individual system owners, mainly in France and Western Europe. The primary purpose of the BDPV database is to monitor system owners' energy production. BDPV also promotes PV energy by disseminating information and data to the general public and public authorities.

The BDPV data contains the localization of more than 28000 installations. We used this localization to extract the panels' thumbnails. During the first annotation campaign, we extracted 28807 thumbnails using Google Earth Engine (GEE)²⁵ application programming interface (API). For the second campaign, we extracted 17325 thumbnails from the IGN Geoservices portal (<https://geoservices.ign.fr/bdortho>).

Our thumbnails all have a resolution of 400×400 pixels. Thumbnails extracted from GEE API correspond to a ground sampling distance (GSD) of 0.1 m/pixel. The API directly generates this thumbnail by setting the zoom level to 20, the localization to the ground truth localization contained in BDPV, and the output size to be 400×400 pixels. For IGN images, the resolution of the thumbnails corresponds to a GSD of 0.2 m/pixel. The procedure for generating IGN thumbnails differs from Google. First, we downloaded geo-localized tiles from IGN's Geoservices portal. These tiles have a resolution of 25000×25000 pixels, covering an area of 25 square kilometers. Then, we extracted the thumbnail by generating a 400×400 pixels raster centered around the location of the PV panel. Finally, we export this raster as a .png file. We do not publish the exact location of the panels for confidentiality reasons.

The crowdsourcing campaigns took place on a dedicated platform called BDAPPV, which stands for “*Base de données apprentissage profond PV*” (PV database for deep learning). The BDAPPV platform is a web page where users can ergonomically annotate aerial images by clicking on the panel (phase 1) or delineating polygons around the PV panels (phase 2). Table 1 summarizes the contribution of the annotators during the crowdsourcing campaigns. The web page is accessible at this URL: https://www.bdpv.fr/_BDapPV/.

Thumbnails annotation. We extracted thumbnails based on the geolocation of the installations recorded in the BDPV dataset. However, this geolocation can be inaccurate, so before asking users to draw polygons of PV installations, we asked them to classify the images. It corresponds to the first phase of the annotation campaign. Once users classified images, we asked them to draw the PV polygons on the remaining images. It corresponds to the second phase of the crowdsourcing campaign.

We designed our campaign to get at least five annotations per image. It enabled us to derive so-called *consensus metrics*, targeted at measuring the quality of our labels. This way, we go further than the consensus between two annotators reported in previous work²⁶ to measure annotation quality. The analysis of the users' annotations during phases 1 and 2 are reproducible using the notebook annotations available on the public repository, accessible at this URL: <https://github.com/gabrielkasmi/bdappv/>.

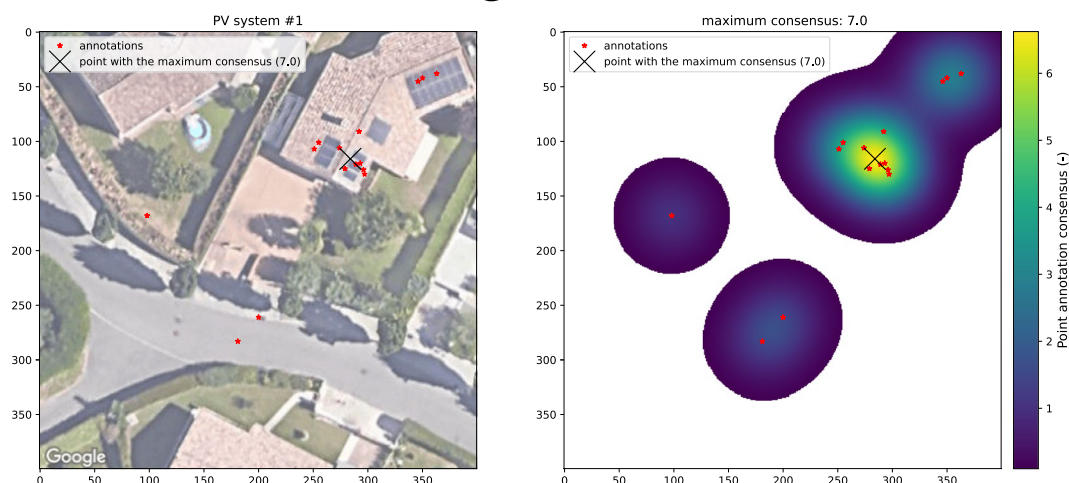
Phase 1: image classification. During the first phase, the user clicks on an image if it depicts a PV panel. We recorded the localization of the user's click and instructed them to click *on* the PV panel if there was one. We collected an average of 10 actions (click with localization or no click) per image. The left panel of Fig. 2 provides an example of annotations during phase 1. We apply the kernel density estimate (KDE) algorithm to the annotations to estimate a confidence level for the annotations and the approximate localization of the PV panel on the image. The likelihood $f_{\sigma}(x_i)$ of presence of a panel for each pixel x_i is given by:

$$f_{\sigma}(x_i) = \frac{1}{N} \sum_{k=1}^N K_{\sigma}(x_k - x_i) \quad (1)$$

where K_{σ} is a Gaussian kernel with a standard deviation σ , x_k is the coordinate of the k^{th} annotation, and N is the total number of annotations.

After an empirical investigation, we calibrated the standard deviation of the kernel to reflect the approximate spatial extent of an array on the image. We set its value to 25 pixels for Google images and 12 for IGN images. It corresponds to a distance of 2.5 m. As illustrated on Fig. 2, the KDE yields a heatmap whose hotspot locates on the solar array. The maximum value of the KDE quantifies the confidence level of the annotation. We refer to it as the *pixel annotation consensus* (PAC). This metric is proportional to the number of annotations. We use the PAC to determine whether an image contains an array.

Phase 1 : image classification



Phase 2 : polygon annotation

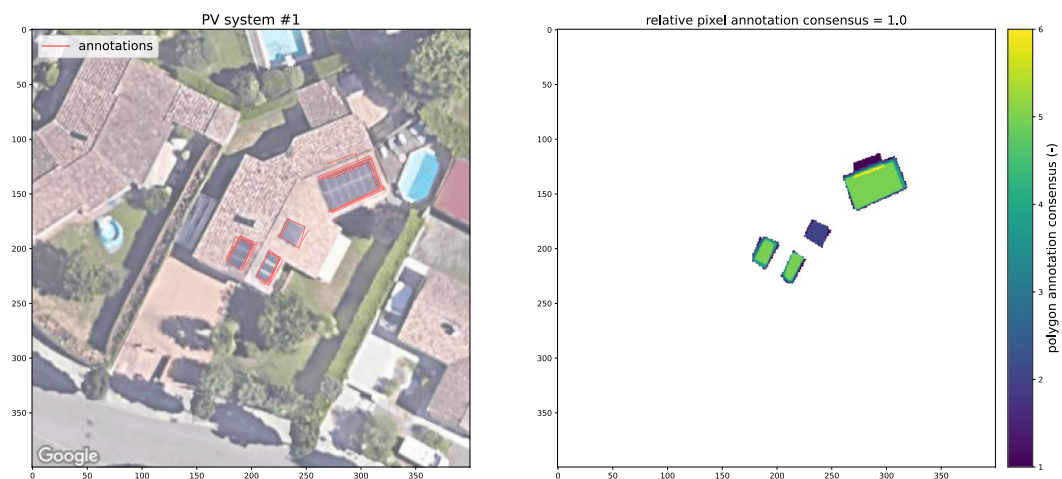


Fig. 2 Screenshot of the annotations notebook, showing analysis of click annotations (phase 1, above) and polygon annotations (phase 2, below). During phase 1 (above), each red dot corresponds to an annotation. The density of annotations is greater near one of the panels, but we also see that other panels also received clicks.

Phase 2: polygon annotation. During the second phase, annotators delineate the PV panels on the images validated during phase 1. Users can draw as many polygons as they want. On average, we collected five polygons per image. We collect the coordinates of the polygons drawn by the annotators. As illustrated in the lower left panel of Fig. 2, a set of polygons is available for each array in an image. We can note from the annotation illustrated in Fig. 2 that some polygons may be erroneous. However, these false positives have fewer annotations than true positives. To select only the true positives, we compute the PAC through the following steps:

1. We convert each user's polygon into a binary raster;
2. We compute the normalized PAC by summing all rasters and dividing by the number of annotators,
3. We apply a relative threshold and keep only the pixels whose PAC is greater than the threshold;
4. We compute the coordinates of the resulting mask using OpenCV's polygon detection algorithm (https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html).

In step 2., the unnormalized PAC takes values between 0 and the number N_i of annotators for the i^{th} image. 0 means no user included the pixel into his polygon, and N_i means that *all* annotators encapsulated the corresponding pixel in their polygons.

Metadata matching. Once we generate our PV panels polygons (i.e. segmentation masks), we match them with the installations' metadata reported in the BDPV dataset. Our matching procedure follows three steps:

Filter	Installation metadata from the BDPV database matched with an image	
	Google images	IGN images
Before filtering	28408	
After internal consistency filtering	27780	
After mask uniqueness*	10523	5883
After external consistency	8019	3658

Table 2. Number of installations filtered through the different filtering steps during the association between the masks and the installations' metadata. *During the mask uniqueness step, we account for the fact that (a) not all BDPV installations were identified on images (13303 were identified on Google images and 7686 on IGN images) and (b) among these identified installations, some of the masks contained more than one polygon.

Field	Attribute name	Description	Format	Unit
Image ID	id	ID of image	String	—
City	city	City of the image	String	—
Department	department	Département of the image	String	—
Region	region	Region of the image	String	—
Installation ID	install_id	ID of the corresponding installation in the BDPV database.	Integer	—
Clicks list	clicks[]	List of clicks	List	—
Click Pixel x coordinate	clicks[].x	x position of click in image	Integer	Pixel
Click Pixel y coordinate	clicks[].y	y position of click in image	Integer	Pixel
Click metadata	clicks[].action	Metadata of the click action	List	—
Click country	clicks[].action.country	Country of actor of the click	String	—
Click region	clicks[].action.region	Region of actor of the click	String	—
Click date	clicks[].action.date	Date/time of click	String	Date
Click author	clicks[].action.actorId	ID of the author of the click	Integer	—
Polygons list	polygons[]	List of polygons	List	—
Polygon points list	polygons[].points[]	List of point of the polygon	List	—
Polygon point Pixel x coordinate	polygons[].points[].x	x position of the point of polygon	Integer	Pixel
Polygon point Pixel y coordinate	polygons[].points[].y	y position of the point of polygon	Integer	Pixel
Polygon metadata	polygons[].action	Meta data of the polygon action	List	—
Polygon country	polygons[].action.country	Country of actor of the polygon	String	—
Polygon region	polygons[].action.region	Region of actor of the polygon	String	—
Polygon date	polygons[].action.date	Date/time of polygon	String	Date
Polygon author	polygons[].action.actorId	ID of the actor of the polygon	Integer	—

Table 3. Data attributes and description of the `input-google.json` and `input-ign.json` data files.

internal consistency, unique matching, and external consistency. Note that we only apply these filters when matching the metadata and the masks.

Internal consistency. ensures that the entries in the BDPV dataset are coherent before any matching. It is simply a cleaning of the raw dataset. To do this cleaning, we verify whether the information in one column of the BDPV dataset is coherent with the records from the other columns. For instance, if a PV system's record says it has ten modules and a surface of 3 squared meters, this would mean that each PV module has a surface of 0.3 squared meters, which is impossible (the smallest size being 1.7 squared meters).

Unique matching. Our segmentation masks may depict more than one array. It occurs if, for instance, more than one panel is on the image shown to the annotators. In this case, we adopt a conservative view: if the segmentation mask depicts more than one panel, we cannot know which one corresponds to the installation reported in the BDPV dataset. We do not match the segmentation mask with an installation in this case.

External consistency. After internal consistency filtering and unique matching, we are left with segmentation masks depicting single panels with coherent metadata. A final filtering step consists in making sure that the characteristics reported in the database match those that can be deduced from the segmentation mask. We assess the adequacy between the surface of the installation's mask and its true surface reported in the BDPV dataset by computing the ratio between them. We keep only installations whose ratio is equal to 1 (with a tolerance bandwidth of $\pm 25\%$). We apply this bandwidth to accommodate the possible approximations in the segmentation

Field	Attribute name	Description	Format	Unit
Image ID	id	ID of the image	String	—
Points list	clicks[]	List of points whose score is greater than the threshold	List	—
Point Pixel x coordinate	clicks[].x	x position of point in image	Integer	Pixel
Point Pixel y coordinate	clicks[].y	y position of point in image	Integer	Pixel
Point score	clicks[].score	Value of the Pixel annotation consensus for this point. This value is greater than the chosen threshold and lower than the number of clicks on this image	Float	—

Table 4. Data attributes and description of the `click-analysis.json` and `click-analysis-thres=1.0.json` data files.

Field	Attribute name	Description	format	Unit
Image ID	id	ID of the image	String	—
Polygons list	polygons[]	List of filtered polygons	List	—
Polygon's points list	polygons[].points[]	List of the points composing the polygon	List	—
Polygon's point pixel x coord.	polygons[].points[].x	x position of one point of the polygon	Integer	Pixel
Polygon's point pixel y coord.	polygons[].points[].y	y position of one point of the polygon	Integer	Pixel
Polygon area	polygons[].area	Area of the polygon	Float	Pixel
Polygon score	polygons[].score	Value of the Pixel annotation consensus for this polygon. This value is greater than the chosen the threshold and lower than the number of clicks on this image.	Float	—

Table 5. Data attributes and description of the `polygon-analysis.json` and `polygon-analysis-thres=1.0.json` data files.

mask. The reported surface excludes the inter-panel space and the distortions induced by the panel's projection on the image, as images are not perfectly orthorectified.

Data Records

The data records consist of two separate datasets, accessible on our Zenodo repository³¹, at this URL: <https://zenodo.org/record/7358126>.

1. The *training dataset*: input images, segmentation masks, and PV installations' metadata,
2. The *crowdsourcing and replication data*: annotations from the users, for each image, provided in.json format and the raw installations' metadata.

Besides, the source code and notebooks used to generate the masks from the users' annotations are accessible on our public Git repository at this URL: <https://github.com/gabrielkasmi/bdappv>. This repository contains the source code used to generate the segmentation masks. It contains the notebooks annotations and metadata, which can be used to visualize the threshold analysis or the metadata matching procedure.

Training dataset. The training dataset containing RGB images, ready-to-use segmentation masks of the two campaigns, and the file containing PV installations' metadata is accessible on our Zenodo repository. It is organized as follows:

- `bdappv/` Root data folder
 - `google/ign` One folder for each campaign
 - `img` Folder containing all the images presented to the annotators. This folder contains 28807 images for Google and 17325 for IGN. We provide all images as.png files.
 - `mask` Folder containing all segmentation masks generated from the polygon annotations of the annotators. This folder contains 13303 masks for Google and 7686 for IGN. We provide all masks as.png files.
 - `-metadata.csv` The.csv file with the metadata of the installations. Table 7 describes the attributes of this table.

Crowdsourcing and replication data. The Git repository contains the raw crowdsourcing data and all the material necessary to re-generate our training dataset and technical validation. It is structured as follows: the raw subfolder contains the raw annotation data from the two annotation campaigns and the raw PV installations' metadata. The replication subfolder contains the compiled data used to generate our segmentation masks. The validation subfolder contains the compiled data necessary to replicate the analyses presented in the technical validation section.

Field	Attribute name	Description	Format	Unit
Installation ID	idInstallation	The ID of the installation	Integer	—
Identifier	identifiant	The name of the image of the installation	String	—
Inverter ID	idInverter	The ID of the inverter of the installation	Integer	—
Inverter name	nameInverter	The name of the inverter of the installation	String	—
Number of inverters	countInverters	The number of inverters of the installation	Integer	—
Arrays ID	idArrays	The ID of the solar arrays used by the installation	Integer	—
Arrays' name	nameArrays	The name of the solar arrays used by the installation	Float	—
Number of arrays	countArrays	The number of PV arrays (modules) of the installation	Integer	—
Surface	surface	The surface (in square meters) of the installation	Float	m ²
Azimuth	azimuth	The azimuth angle in degrees relative to the north (south = 180) of the installation.	Float	Degrees
Installation type	typeInstallation	Indicates on which infrastructure the installation is mounted:	Integer	—
		- 0: rooftop		
		- 1: unknown		
		- 2: rooftop of a non-livable building		
		- 3: ground		
		- 4: other		
		- 5: shade house		
		- 6: sunshade		
		- 7: solar tracker with 1 axis		
		- 8: solar tracker with 2 axes		
Tilt	tilt	The tilt angle of the installation	Integer	Degrees
Installed capacity	kWp	The installed capacity of the installation in Wp	Float	Wp
Date of installation	dateInstalled	The date (month, year) the installation has been installed	String	Date
Is integrated	isIntegrated	Indicates if the installation is integrated (on the rooftop)	Boolean	—
Self-consumption	selfConsumption	Indicates if the installation is used for self-consumption (alternative is that PV power is reinjected into the grid)	Boolean	—
Département	departement	The <i>département</i> (county) in which the installation is located	Integer	—
City	city	The city where the installation is located	String (UTF-8)	—

Table 6. Data attributes and description of the `raw-metadata.csv` data file.

- `data/` Root data folder
 - `raw/` Folder containing the raw crowdsourcing data and raw metadata;
 - `input-google.json`: Input data containing all information on images and raw annotators' contributions for both phases (clicks and polygons) during the first annotation campaign;
 - `input-ign.json`: Input data containing all information on images and raw annotators' contributions for both phases (clicks and polygons) during the second annotation campaign;
 - `raw-metadata.csv`: The file containing the PV systems' metadata extracted from the BDPV database before filtering. It can be used to replicate the association between the installations and the segmentation masks, as done in the notebook `metadata`. Table 6 describes the attributes of the `raw-metadata.csv` table.
 - `-replication/` Folder containing the compiled data used to generate the segmentation masks;
 - `campaign-google/campaign-ign`. One folder for each campaign
 - `click-analysis.json`: Output on the click analysis, compiling raw input into a few best-guess locations for the PV arrays. This dataset enables the replication of our annotations;
 - `polygon-analysis.json`: Output of polygon analysis, compiling raw input into a best-guess polygon for the PV arrays.
 - `validation/` Folder containing the compiled data used for technical validation.
 - `campaign-google/campaign-ign`. One folder for each campaign
 - `click-analysis-thres=1.0.json`: Output of the click analysis with a lowered threshold to analyze the effect of the threshold on image classification, as done in the notebook annotations;
 - `polygon-analysis-thres=1.0.json`: Output of polygon analysis, with a lowered threshold to analyze the effect of the threshold on polygon annotation, as done in the notebook annotations.
 - `metadata.csv` the filtered installations' metadata.

Raw crowdsourcing data and raw installations' metadata. The files `input-google.json` and `input-ign.json` provide the raw crowdsourcing data for both annotation campaigns. The two files are identically formatted. They contain all metadata for all images and the associated annotators' contributions for both phases (clicks and polygons). It enumerates all the clicks and polygons attached to the image. Coordinates

Field	Attribute name	Description	Format	Unit
Installation ID	idInstallation	The ID of the installation	Integer	—
Identifier	identifiant	The name of the image of the installation	String	—
Inverter ID	idInverter	The ID of the inverter of the installation	Integer	—
Inverter name	nameInverter	The name of the inverter of the installation	String	—
Number of inverters	countInverters	The number of inverters of the installation	Integer	—
Arrays ID	idArrays	The ID of the solar arrays used by the installation	Integer	—
Arrays' name	nameArrays	The name of the solar arrays used by the installation	Float	—
Number of arrays	countArrays	The number of PV arrays (modules) of the installation	Integer	—
Surface	surface	The surface (in square meters) of the installation	Float	m ²
Azimuth	azimuth	The azimuth angle in degrees relative to the north (south = 180) of the installation.	Float	Degrees
Installation type	typeInstallation	Indicates on which infrastructure the installation is mounted:	Integer	—
		- 0: rooftop		
		- 1: unknown		
		- 2: rooftop of a non-livable building		
		- 3: ground		
		- 4: other		
		- 5: shade house		
		- 6: sunshade		
		- 7: solar tracker with 1 axis		
		- 8: solar tracker with 2 axes		
Tilt	tilt	The tilt angle of the installation	Integer	Degrees
Installed capacity	kWp	The installed capacity of the installation in Wp	Float	Wp
Date of installation	dateInstalled	The date (month, year) the installation has been installed	String	Date
Is integrated	isIntegrated	Indicates if the installation is integrated (on the rooftop)	Boolean	—
Self-consumption	selfConsumption	Indicates if the installation is used for self-consumption (alternative is that PV power is reinjected into the grid)	Boolean	—
Département	departement	The <i>département</i> (county) in which the installation is located	Integer	—
City	city	The city where the installation is located	String (UTF-8)	—
Controlled	Controlled	Indicates whether the installations' metadata are clean	Boolean	—
Matched with IGN image	IGNControlled	Indicates whether the installation corresponds to a unique segmentation mask corresponding to an IGN image	Boolean	—
Matched with Google image	GoogleControlled	Indicates whether the installation corresponds to a unique segmentation mask corresponding to a Google image	Boolean	—

Table 7. Data attributes and description of the `metadata.csv` data file.

are expressed in pixels relative to the upper-left corner of the image. We also provide information on the click or the polygon (annotator's ID, date, country). The field "Clicks list" is empty when an image contains no clicks. The field "Polygons list" is empty if it does not contain polygons. Table 3 describes the attributes of the `input` tables.

The raw metadata datasheet corresponds to the extraction of the BDPV database. This file contains all the installations' metadata. Table 6 provides a list of the complete attributes. We use this file as input to associate the segmentation masks to the installations' metadata.

Replication of the image classification and polygon annotation. We compiled the files `click-analysis.json` and `polygon-analysis.json` from the raw inputs to classify the images and generate the segmentation masks, respectively. We provide these files to enable users to replicate our classification process and the generation of our masks.

The output of click analysis contains a list of detected PV installations' positions for each image. Each image contains at least one point, corresponding to the number of panels found in it. The `score` variable summarizes the PAC associated with each point. By construction, the `click-analysis.json` files only contain points with a PAC greater than 2.0 (see the technical validation section for more details on the threshold tuning). Table 4 describes the attributes of the `click-analysis.json` data file.

The output of polygon analysis contains a list of polygons as a compilation of all polygons annotated by annotators. It contains one or more polygons for each image corresponding to the PV arrays. Analogous to the click-analysis files, this file summarizes the polygon annotations of the users. The variable `score` records the relative PAC associated with each polygon. By construction, the `polygon-analysis.json` files only contain polygons with a relative PAC greater than 0.45 (see the technical validation section for more details on the threshold tuning). Table 5 describes the attributes of the `polygon-analysis.json` data files.

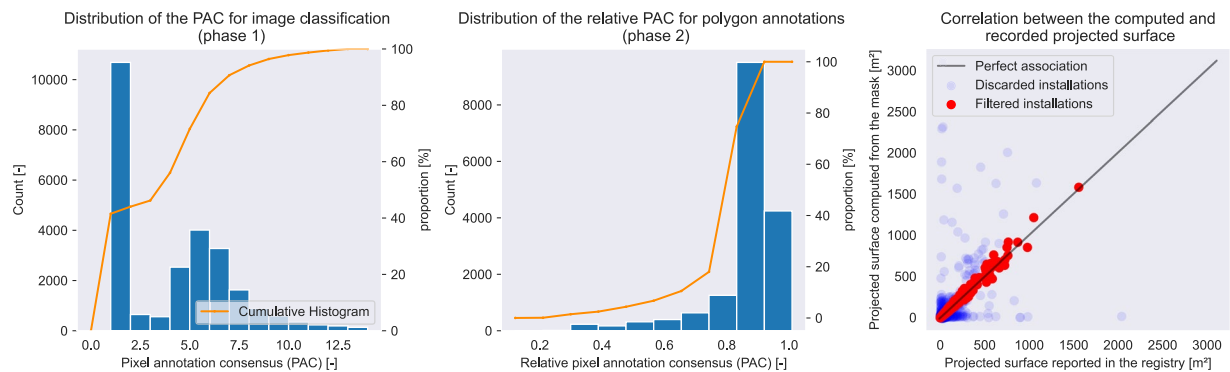


Fig. 3 Validation by comparison of the surface estimated from the masks and the surface reported in the PV installations' metadata.

Validation data. We compiled the files `click-analysis-thres=1.0.json` and `polygon-analysis-thres=1.0.json` to enable users to study how the thresholds chosen to generate our annotations affect the segmentation masks that we generate. The notebook `annotations` enables us to carry out this study.

The `metadata.csv` file corresponds to the output file of the notebook `metadata`. We provide this file to enable users to replicate our analysis of the fit between the filtered installations and their segmentation masks.

Technical Validation

Throughout the generation of the training dataset, we tested whether the threshold values chosen to classify the images, construct the polygon and associate the polygons to the installations' metadata yielded as few errors as possible. We base our approach on a consensus metric to classify images and construct the polygons, namely the pixel annotation consensus (PAC). Thus, we improve on Bradbury *et al.*²⁶, who proposed a confidence value based on the Jaccard Similarity Index³² between the two annotations. As for the association between the polygons and installations' metadata, we balance between accuracy and keeping as many installations as possible.

Analysis of the consensus value for image classification. As mentioned in the methods section, the choice criterion for image classification during phase 1 is the consensus among users. We empirically investigated a range of thresholds and determined that a value of 2.0 yielded the most accurate classification results. In other words, we require that at least three annotators click around the same point to validate the classification.

We use an *absolute* (unnormalized by the number of annotators for this image) threshold to decide whether the image contains a panel. The threshold is absolute because users could only click once on the image during the annotation campaign, even if the latter contained more than one array. As such, an absolute threshold does not dilute the consensus among users when there is more than one panel on the image.

The leftmost plot of Fig. 3 plots the histogram of the absolute PAC. Visual inspection revealed that the peak for values below 2.0 corresponded to false positives. We enable replication of the threshold analysis in the notebook `annotation`.

Analysis of the consensus value of the polygon annotation. Like the click annotation, we used a consensus metric to merge the users' annotations. After empirical investigations, we found that a *relative* threshold (expressed as a share of the total number of annotators) was the most effective for yielding the most accurate masks and that its value should be 0.45. In other words, we consider that a pixel depicts an installation if at least 45% of the annotators included it in their polygons.

The center plot of Fig. 3 depicts the histogram of the relative PAC. Visual inspection revealed that the few values below 0.45 corresponded to remaining false positives (e.g., roof windows). The use of a relative threshold is motivated by the fact that the users can annotate as many polygons as they want. We enable replication of the threshold analysis in the notebook `annotation`.

Consistency between polygon annotations and metadata of the PV installations. We link segmentation masks and installations metadata according to the steps described in the section "Metadata matching." To measure the quality of this linkage, we measure the Pearson correlation coefficient (PCC) between the surface reported in the installation' metadata dataset (referred to as the "target" surface) and the surface estimated from the segmentation masks (referred to as the "estimated" surface). The higher the PCC, the better our matching procedure.

Figure 3 plots estimated and target surfaces. After filtering, we obtain a PCC coefficient of 0.99 between the target and estimated surfaces. Without filtering, the PCC coefficient equals 0.68 for Google images and 0.61 for IGN images. It shows that our metadata-matching procedure enabled us to pick the installations with the best fit between the reported surface and the surface estimated from the segmentation masks.

Our matching procedure comprises three steps: internal consistency, uniqueness and external consistency. Each of these steps discards installations from the BDPV database. Table 2 summarizes the number of installations filtered at each process step. We can see that most of the filtering happens when we discard segmentation masks on which there is more than one installation.

Usage Notes

We designed the complete dataset records to be directly used as training data in machine learning projects. The ready-to-use data is accessible on our Zenodo repository accessible at this URL <https://zenodo.org/record/7358126>. This repository also stores the raw crowdsourcing data and the files necessary to reproduce our segmentation masks and analyses. We compiled the files `click-analysis.json` and `polygon-analysis.json` using the Python scripts `click-analysis.py` and `polygon-analysis.py`, provided in our repository from the raw input data `input-{provider}.json`. This repository also contains the notebooks annotations and metadata. The notebook annotations presents the analysis of crowdsourced data from the crowdsourcing campaigns. The notebook metadata filters the raw-metadata.csv datasheet.

Between phases 1 and 2, we generated new thumbnails re-centered on the PV installations. The new center corresponds to the coordinates of the estimated center of the (first in the list) detected PV installation. Therefore, to replicate the click analysis on the corresponding image, interested users need to download the corresponding image accessible on the BDAPPV website as illustrated in the notebook annotations. We re-center images by generating a new thumbnail centered around an updated location, according to the procedure described in the section “methods”.

The centering of the images will not induce a bias during learning because our thumbnails have a larger resolution (400×400 pixels) than the typical input size of typical neural networks (224×224 pixels). Adding a random crop transform during training will result in panels not being centered anymore. Besides, during the IGN campaign, we only re-centered about 13% of the images.

Rights and permissions. Code, raw crowdsourcing data, and compiled data are accessible on the project repository. All materials are provided under the CC-BY license. This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, as long as attribution is given to the creator. The license allows for commercial use.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or third-party material in this article are included in the article's Creative Commons license unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use. In that case, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

Code availability

Our public repository accessible at this URL <https://github.com/gabrielkasmi/bdappv> contains the code to generate the masks, filter the metadata and analyze our results. Interested users can clone this repository to replicate our results or conduct analyses.

Received: 8 September 2022; Accepted: 10 January 2023;

Published online: 28 January 2023

References

1. RTE France. *Bilan électrique 2021*. <https://bilan-electrique-2021.rte-france.com/> (2022).
2. IEA. *Solar PV*. <https://www.iea.org/reports/solar-pv> (2022).
3. Shaker, H., Zareipour, H. & Wood, D. A data-driven approach for estimating the power generation of invisible solar sites. *IEEE Transactions on Smart Grid* **7**, 2466–2476 (2015).
4. Kazmi, H. & Tao, Z. How good are TSO load and renewable generation forecasts: Learning curves, challenges, and the road ahead. *Applied Energy* **323**, 119565 (2022).
5. Saint-Drenan, Y.-M., Good, G. H., Braun, M. & Freisinger, T. Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method. *Solar Energy* **135**, 536–550 (2016).
6. Saint-Drenan, Y.-M. *et al.* Bayesian parameterisation of a regional photovoltaic model—Application to forecasting. *Solar Energy* **188**, 760–774 (2019).
7. Huber, M., Dimkova, D. & Hamacher, T. Integration of wind and solar power in Europe: Assessment of flexibility requirements. *Energy* **69**, 236–246 (2014).
8. Saint-Drenan, Y. M., Good, G. H. & Braun, M. A probabilistic approach to the estimation of regional photovoltaic power production. *Solar Energy* <https://doi.org/10.1016/j.solener.2017.03.007> (2017).
9. Killinger, S. *et al.* On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading. *Solar Energy* **173**, <https://doi.org/10.1016/j.solener.2018.08.051> (2018).
10. De Jong, T. *et al.* Monitoring Spatial Sustainable Development: semi-automated analysis of Satellite and Aerial Images for Energy Transition and Sustainability Indicators. *arXiv preprint arXiv:2009.05738* (2020).
11. Wang, Z., Arlt, M.-L., Zanolto, C., Majumdar, A. & Rajagopal, R. DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models. *Joule* **6**, 2611–2625 (2022).
12. Dunnett, S., Sorichetta, A., Taylor, G. & Eigenbrod, F. Harmonised global datasets of wind and solar farm locations and power. *Scientific data* **7**, 1–12 (2020).
13. Kruitwagen, L. *et al.* A global inventory of photovoltaic solar energy generating units. *Nature* **598**, 604–610 (2021).
14. Stowell, D. *et al.* A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK. *Scientific Data* **7**, 1–15 (2020).
15. Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. *Joule* **2**, 2605–2617 (2018).

16. Zech, M. & Ranalli, J. Predicting PV Areas in Aerial Images with Deep Learning. In *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 0767–0774 (IEEE, 2020).
17. Malof, J. M., Bradbury, K., Collins, L. M. & Newell, R. G. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied energy* **183**, 229–240 (2016).
18. Hu, W. *et al.* What you get is not always what you see—pitfalls in solar array assessment using overhead imagery. *Applied Energy* **327**, 120143 (2022).
19. Mayer, K. *et al.* 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy* **310**, 118469 (2022).
20. Wang, R., Camilo, J., Collins, L. M., Bradbury, K. & Malof, J. M. The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: an empirical study with solar array detection. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–8 (IEEE, 2017).
21. Kasmi, G., Dubus, L., Blanc, P. & Saint-Drenan, Y.-M. Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping. In *Workshop on Machine Learning for Earth Observation (MACLEAN), in Conjunction with the ECML/PKDD 2022* (2022).
22. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528 (IEEE, 2011).
23. Koh, P. W. *et al.* Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664 (PMLR, 2021).
24. Tuia, D., Persello, C. & Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine* **4**, 41–57 (2016).
25. Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* **202**, 18–27 (2017).
26. Bradbury, K. *et al.* Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data* **3**, 1–9 (2016).
27. Khomiakov, M. M. *et al.* SolarDK: A high-resolution urban solar panel image classification and localization dataset. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning* (2022).
28. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).
29. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
30. Lefort, T., Charlier, B., Joly, A. & Salmon, J. Improve learning combining crowdsourced labels by weighting Areas Under the Margin. *arXiv preprint arXiv:2209.15380* (2022).
31. Kasmi, G. *et al.* A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Zenodo* <https://doi.org/10.5281/zenodo.7358126> (2022).
32. Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**, 34–35 (1971).

Acknowledgements

We want to thank all annotators who participated in the crowdsourcing campaigns. We also would like to thank the association BDPV and the users of the online forum *Forum Photovoltaïque* (<https://forum-photovoltaïque.fr/>). They actively participated in the execution of the crowdsourcing elaboration of this project.

This project is carried out as part of the Ph.D. thesis of Gabriel Kasmi, sponsored by the French transmission system operator RTE France and partly funded by the national agency for research and technology (ANRT) under the CIFRE contract 2020/0685.

The European Commission partially funds the work of Jonathan Leloux and Babacar Sarr through the Horizon 2020 project SERENDI-PV (<https://serendipv.eu/>), which belongs to the Research and Innovation Programme, under Grant Agreement 953016.

Author contributions

G.K.: Contribution to the concept development, contributions to the manuscript review and editing of the final draft, second campaign management (generation of the thumbnails from IGN data for the two phases), matching between the metadata and the installations, Figs. 1,3 tables formatting, crowdsourcing campaign analysis, manuscript revision. R.J.: Contribution to manuscript and code for annotations analysis (scripts and notebooks), generation of the ground truth labels, Fig. 2 and tables formatting. Y.-M. S.-D.: development of the concept of BDAPPV, contribution and reviews to the original manuscript, analysis of the annotation, Figs. 2,3 formatting, manuscript revision. D.T.: development of the concept of BDAPPV, conception and development of the annotation platform, communication with the community of annotators, collection and management of the annotation data, and contribution to the manuscript. J.L.: contribution to the concept development and review of the manuscript. B.S.: contribution to the concept development and review of the manuscript. L.D.: contribution to the concept development and review of the manuscript.

Competing interests

Y.-M. Saint-Drenan and D. Trebosc are members of the non-profit association *Asso BDPV*, which collects and maintains a database of rooftop PV plants. Y.-M. Saint-Drenan, D. Trebosc, J. Leloux, and B. Sarr are members of the initiative coBDPV, aiming to improve the analysis conducted within the BDPV platform. G. Kasmi is carrying out a Ph.D. funded by RTE (the French transmission system operator) and the national agency for research and technology (ANRT). L. Dubus is a senior research scientist at RTE. R. Jolivet declares no conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to G.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023