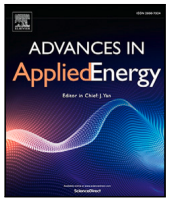
Advances in Applied Energy 16 (2024) 100191

Contents lists available at ScienceDirect

Advances in Applied Energy

journal homepage: www.elsevier.com/locate/adapen



# Toward global rooftop PV detection with Deep Active Learning

Matthias Zech [a,*], Hendrik-Pieter Tetens [a], Joseph Ranalli [b]

[a] German Aerospace Center (DLR), Institute of Networked Energy Systems, Carl-von-Ossietzky-Straße 15, Oldenburg, 26129, Germany
[b] Penn State Hazleton, 76 University Drive, Hazleton, 18202, PA, USA

ARTICLE INFO

Dataset link: Code to run the experiments and create plots, will be published on github once accepted (Original data)

*Keywords:*
Deep Active Learning
PV panel detection
Machine Learning
Semantic segmentation
Remote sensing

ABSTRACT

It is crucial to know the location of rooftop PV systems to monitor the regional progress toward sustainable societies and to ensure the integration of decentralized energy resources into the electricity grid. However, locations of PV are often unknown, which is why a large number of studies have proposed variants of Deep Learning to detect PV panels in remote sensing data using supervised Deep Learning. However, these methods are based on annotating datasets and therefore often require relabeling when fine-tuned or extended to a different region. Recent advances in Deep Active Learning offer the opportunity to significantly reduce the number of required annotated images by intelligently selecting the images to label next based on their informative value for the model. In this study, we compare different Deep Active Learning algorithms using a variety of datasets from different regions and compare different model training variants. In the simulations, the entropy-based acquisition function shows the highest performance with only 3% of the data needed in case-imbalanced data, while remaining simple to implement. We believe that Deep Active Learning provides an elegant solution to maintain high model accuracy while reducing annotation effort substantially. This facilitates the development of generalizable models for worldwide rooftop PV detection.

## 1. Introduction

### 1.1. Background

Solar photovoltaics (PV) have shown unprecedented global annual growth rates of 50% during the last decade [1] and are expected to become the main energy supply technology in 2050, with electricity production shares of 30 to 50% in competitive markets [2]. PV modules are granular, meaning that identical PV panels can be combined in various configurations, from a few PV panels used in residential applications, up to millions of PV panels for utility-scale applications. This modularity has contributed to rapid scaling and cost reductions [1,3]. The rapid development of PV technology has largely been driven by rooftop PV systems [4], which account for approximately 40% of the total installed PV capacity [5]. Despite their importance, relatively little information is available on the location and capacity of these rooftop PV systems [6], although such information is crucial for the operation of renewable energy-based systems. For example, it is essential for electricity grid operators to calculate actual and predicted regional PV inputs [7,8]. Moreover, accurate PV location data enables policymakers to track regional progress toward sustainable energy system goals and to design and evaluate equitable policies. Energy research also relies on accurate data, in particular for the modeling of urban energy systems on a community scale [9,10]. While Open Street Map data provide reasonable estimates [11], more advanced methods are required to obtain accurate and updated global PV registries.

### 1.2. Related works

To tackle the issue of limited information about global PV systems, a large number of studies have proposed Machine Learning (ML) techniques in combination with Remote Sensing data to detect PV panels from above. To categorize existing approaches, we conducted a literature review, which is summarized in Table 1. Note that the literature search is limited to the task of detecting PV panels and thus excludes the large research body about PV fault classification [31,49], in which the module locations are already known.

Since 2014, a large number of different approaches have been used to detect PV panels in satellite and aerial imagery. Early works [12,13] were based on Support Vector Machines, while later works are mainly driven by the latest advances in deep neural networks. This shift toward Deep Learning is typical for the use of ML in remote sensing [50]. Different Deep Learning (DL) model architectures have been proposed such as U-Net, DeepLabV3 and recently, transformer-based models, such as SegFormer. Furthermore, authors proposed customized architectures for PV detection, such as PV-UNet [43], GenPV [34], or

* Corresponding author.
*E-mail address:* matthias.zech@dlr.de (M. Zech).

https://doi.org/10.1016/j.adapen.2024.100191
Received 5 August 2024; Received in revised form 6 September 2024; Accepted 25 September 2024
Available online 30 September 2024

**Table 1**
Models that use Machine Learning and remote sensing data for PV panel detection.

| | Year | Region | Method | Data | Resolution [m] | Channels | Strengths |
|---|---|---|---|---|---|---|---|
| [12] | 2014 | Abu Dhabi | SVM | Google | – | RGB | Pioneering work |
| [13] | 2015 | Lemoore (California) | SVM | Aerial imagery | 0.3 | RGB | Pioneering work |
| [14] | 2016 | Washington D.C.;San Fransisco;Boston | CNN | Aerial imagery | 0.3 | RGB | Pioneering work using CNNs |
| [15] | 2017 | – | CNN | Google | – | RGB | Low quality image data |
| [16] | 2018 | Fresno (California) | SegNet | Aerial imagery | 0.3 | RGB | Publication of dataset; high-resolution |
| [17] | 2018 | United States (50 cities) | Inception-v3 | Google | 0.15 | RGB | Large-scale coverage (US) |
| [18] | 2019 | Switzerland | U-Net | Aerial imagery | 0.25 | RGB | Feasibility study |
| [19] | 2019 | China | SolarNet (FCN,EMANet) | Satellite | – | RGB | Better performance |
| [20] | 2020 | Oldenburg (Germany) | U-Net | Google | 0.27 | RGB | Uncertainty quantification |
| [21] | 2020 | Fresno (California) | Crossnet (U-Net) | Aerial imagery | 0.3 | RGB | Cross learning |
| [22] | 2021 | Worldwide | U-Net | Satellite | 10 (Sentinel-2), 1.5 (SPOT6/7) | 12 bands (Sentinel-2), RGBIR (SPOT6/7) | Global inventory |
| [23] | 2021 | Netherlands | TernausNet (U-Net) | Aerial imagery | 0.05–0.107 | RGBIR | Feasibility study; infrared data fusion |
| [24] | 2021 | Jiangsu (China) | DeepLabv3+ | Aerial;Satellite;UAV | 0.1,0.3,0.8 | RGB | Model comparison (RefineNet, U-Net, DeepLabv3+) Cross-Application |
| [25] | 2021 | Brazil | U-Net (Efficient-net-B7) | Sentinel-2 | 10 | RGBIR | Model comparison (U-net, DeepLabv3+, Pyramid Scene Parsing Network, Feature Pyramid Network), Backbone comparison (EfficientNet, ResNet) |
| [26] | 2021 | Germany (Berlin;NRW;Thuringia) | RetinaNet | Aerial imagery | 0.2 | RGBIR | Partially automated (through address data) |
| [27] | 2022 | NRW (Germany) | DeepLabV3 | Aerial imagery | 0.1 | RGB | Tilt and azimuth derivation |
| [28] | 2022 | India | U-Net | Sentinel-2 | 10,20,60 | 12 bands | Open dataset of solar farms in India |
| [29] | 2022 | China | Random Forest | Sentinel-1;Sentinel-2;VIIRS | 10;20;60 | All bands, monthly VIIRS | VIIRS information to detect human settlements |
| [30] | 2022 | Golmud (China) | XGBoost | Landsat-8 | 30 | 11 bands | Local terrain features |
| [31] | 2022 | Italy;Spain;Japan | Mask R-CNN | UAV (infrared) | | | Thermal images, Anomaly of the PV plant |
| [32] | 2022 | Netherlands | Random Forest | Sentinel-1;Sentinel-2 | 10;20 | 12 bands | Identify most important channels and indices |
| [33] | 2023 | Piedmont (Italy) | U-Net | Aerial imagery | 0.3 | RGBIR | Mono vs. Polycrystalline |
| [34] | 2023 | Heilbronn (Germany) | GenPV | Google | 0.15;0.3;0.6 | RGB | Novel method(loss function) |
| [35] | 2023 | United States | U-Net | Aerial imagery | 0.6 | RGB | Validation on capacity and energy generation data |
| [36] | 2023 | Heilbronn (Germany) | SegFormer | Aerial imagery | 0.15 | RGB | Constraint refinement (Color & shape loss) Model comparison (SegFormer;DeepLabV3+;FCN;UPerNet) |
| [37] | 2023 | Regions in Germany;China;France | DeepLabv3 | Bing Maps | 0.1;0.2;0.3;0.8 | RGB | Multi-resolution training |
| [38] | 2023 | Regions in China | PVNet | Google | 0.5;0.54;0.6 | RGB | Enhanced model |
| [39] | 2023 | California;Heilbronn (Germany) | Rooftop PV Segmenter | Aerial imagery | 0.3;0.15 | RGB | Semantic Refinement Module, Feature Aggregation Module, Deep Supervision Module |
| [40] | 2023 | Switzerland | Mask2Former | Aerial imagery | 0.1 | RGB | Show better performance for Mask2Former architecture |
| [41] | 2024 | China | Segment Anything Model | Google | 2 | RGB | Weakly-supervison by Segment Anything Model |
| [42] | 2024 | Selected solar farms in Europe | Solis-Seg+DeepLab | Sentinel-2 | 10 | 12 bands | Neural Architecture Search |
| [43] | 2024 | Ordos (China) | PV-UNet (attention-based) | Gaofen-2;Sentinel-2 | 1 (Gaofen-2), 10;20;60 (Sentinel-2) | RGBIR; 12 bands | Robustness to different measurements, integrates low-resolution and high-resolution |
| [44] | 2024 | Heilbronn (Germany) | SegFormer | Aerial imagery | 0.15 | RGB | Generate artificial images for data augmentation |
| [45] | 2024 | Germany;New-York;France;California | U-Net | Aerial imagery | 0.15;0.27;0.3;0.45 | RGB | Low generalizability of DL models for PV |
| [46] | 2024 | China;France | Mask2Former | Satellite, Aerial imagery | 0.1,0.2,0.3,0.8 | RGB | Architecture (compared against U-Net DeepLabv3+) |
| [47] | 2024 | Islamabad (Pakistan) | U-Net, DeepLabV3 | Google | 0.25 | RGB | Used for PV simulation for current and future PV system |
| [48] | 2024 | Heilbronn (Germany) | TransPV (U-Net, Transformer) | Aerial imagery | 0.15 (Heilbronn); 0.2 (France) | RGB | Refining loss function; better generalizability |

**Legend:** *Google* as a data source refers to using images from Google Earth, as retrieved from Google Static Maps. *RGB* refers to red, green and blue channels, *RGBIR* refers to RGB with an additional infrared channel.

TransPV [48], which modify elements of the model architecture or tailor the loss function to detect PV systems. Given the high activity in this research domain, and more broadly in architecture development for semantic segmentation, it can be expected that numerous novel DL architectures with better performance are proposed in academic literature. However, the general principle of using semantic segmentation in a supervised learning setting for PV detection, as applied in all the listed publications, is expected to persist.

The literature can be divided into models with a regional focus, using aerial imagery in selected regions with image resolutions below 1 m/px, and studies with a larger spatial domain that use globally available satellite imagery [22]. Regional studies are conducted primarily in developed and leading economies with significant PV capacity, such as Germany, France, Switzerland, the United States, and China. The required spatial resolution of around 0.15 to 0.3 meters [51] can only be achieved through aerial imagery from regional airborne campaigns or commercial satellite imagery. Publicly accessible satellites, such as Sentinel-2 [52] and Landsat-8 [53], have resolutions that are orders of magnitude lower than needed and therefore are only capable of capturing utility-scale solar farms. Consequently, global PV registries that include individual PV panels and rooftop PV systems will likely need to be created by combining coarse satellite data with a mosaic of regional aerial imagery.

The application of a DL model on different remote sensing datasets from different airborne campaigns or satellites is challenging, as DL models trained in one region are known to show poor performance in other regions [45,54], necessitating additional fine-tuning and re-labeling. Although efforts have been made to improve generalizability through the use of generative AI [44] or by Feature Pyramid Networks [51], the ability to generalize across datasets with differing

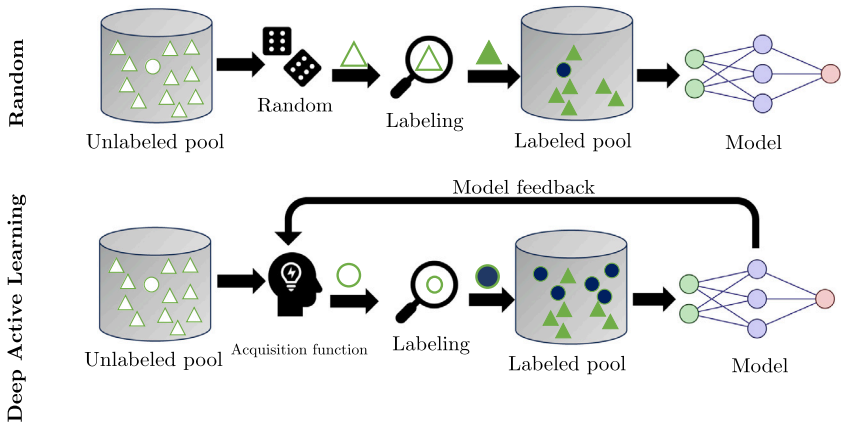



**Fig. 1.** Schematic overview of Deep Active Learning in comparison to the common practice of learning with randomly selected images.

measurement and local characteristics remains uncertain, requiring further model finetuning on additional labeled datasets. Furthermore, PV datasets are highly imbalanced [51], making the selection of relevant images particularly challenging.

Deep Active Learning (DeepAL) provides a strategy to identify model-critical images for the training of DL models. It aims to reduce the number of labels required for training deep learning models by selectively labeling images based on their potential to improve the model's performance, rather than randomly labeling uninformative images. Active Learning, the counterpart to traditional, non-batch trained ML models, has been widely implemented in the remote sensing domain. For example, Active Learning for Support Vector Machines has been evaluated for object classification in remote sensing, showing promise to reduce the number of needed labels [55–57]. It has also been used for object segmentation [58]. More recently, [59] showed that DeepAL algorithms can be used effectively to greatly reduce the number of labels needed for building segmentation. Although they illustrate the advantage of DeepAL in the remote sensing domain, it remains unknown how well DeepAL works on a realistic object detection dataset that contains different regions, has small signal-to-noise ratios, and is highly imbalanced, with many images only showing background information, as is the case with detecting PV panels. Furthermore, it needs to be demonstrated that DeepAL can also be used for fine-tuning an existing model on a different dataset or extending an existing model by an additional dataset.

### 1.3. Research contributions

This article contributes to the large body of academic literature on detecting PV panels using DL models. We believe that labeling effort is the core challenge hindering wide-scale adaption of scalable DL models for global PV registries. As this issue can be ameliorated by using DeepAL, we demonstrate its application as a missing step toward the development of global PV registries. More concretely, we contribute to the literature by investigating the following research questions:

1. Can DeepAL facilitate model training to detect PV systems?
2. Which acquisition function is preferable?
3. Can DeepAL handle imbalances in PV data?
4. How well does DeepAL work for different PV detection model tasks (fine-tuning and joint-learning)?

## 2. Deep active learning

Training DL models requires large datasets that are annotated through tedious labeling efforts, with time primarily spent on labeling randomly sampled images. By randomly selecting images, the image labeling and model training are completely decoupled, likely resulting in an inefficient method of labeling, e.g. providing redundant information and potentially missing out on important information. DeepAL proposes a more effective approach, as described in [60], by establishing a feedback loop between the model and the labeling process as depicted in Fig. 1. The feedback from the model aims to iteratively provide information on which images are most valuable to annotate next by making use of an acquisition function. To formalize, the scenario considers an *unlabeled pool* $\mathcal{U}$ of images, out of which DeepAL aims to iteratively sample a batch of images ($\mathcal{B} = x_1, x_2, \dots, x_b \subseteq \mathcal{U}$) and label them ($y_1, y_2, \dots, y_b$). The labeled batch is then added to the training set ($\mathcal{D}_{train}$) and the model is retrained with this extended training dataset. This procedure is repeated for a number of rounds until the model has sufficient accuracy or the annotation budget is spent.

Unsurprisingly, the choice of the acquisition function within DeepAL is crucial, leading to a large number of different proposals from academia. For this study, we investigate three promising acquisition functions.

### 2.1. Uncertainty-based acquisition functions

Uncertainty-based querying strategies assume that showing the model images that experience high uncertainty will provide the most valuable information for effective model training. In case of PV systems, this would ideally propose objects that are known to be difficult to discriminate, such as winter gardens, greenhouses or objects orientated in parallel lines similar to those of PV systems. Multiple different acquisition functions have been proposed following this rationale, as well-documented in literature [61,62]. We select two representatives of uncertainty-based acquisition functions, described in more detail below.

#### 2.1.1. Entropy

Entropy-based sampling (***Entropy***) described in [63] refers to an acquisition function that approximates the uncertainty using the Shannon entropy [64]. This metric stems from information theory and estimates the model confidence in its prediction [65]. It is defined as

$$\mathbb{H}[y|x, \mathcal{D}_{train}] = -\sum_{c} p(y = c|x, \mathcal{D}_{train}) \log p(y = c|x, \mathcal{D}_{train}) \quad (1)$$

with $p$ representing the posterior label probability of the classifier for class $c$. The conditional probability that a pixel belongs to that class can be derived from the model's sigmoid function in a binary classification setting, or the model's softmax function in a multi-class setting. Although these outputs are uncalibrated and often overconfident [66], they provide estimates of low-confidence (high entropy) and high-confidence (low entropy) model predictions. The runtime to calculate the Shannon entropy is nearly identical to running one forward pass through the DL model, and as the activation layers are already available

in DL model architectures, the entropy method is easy to implement and computationally efficient. The entropy score is calculated pixel-wise, which is why it is averaged to image-wise values as in [67].

#### 2.1.2. Bayesian active learning by disagreement (BALD)

The second uncertainty-based acquisition function in this study is ***Bayesian Active Learning by Disagreement (BALD)*** from [68]. This technique aims to maximize the mutual information between model predictions and model parameters and is formulated as

$$\mathbb{I}[y, \omega | x, \mathcal{D}_{train}] = \mathbb{H}[y|x, \mathcal{D}_{train}] - \mathbb{E}_{p(\omega|\mathcal{D}_{train})}\left[\mathbb{H}[y|x, \omega]\right] \quad (2)$$

The first term $\mathbb{H}[y|x, \mathcal{D}_{train}]$ expresses the entropy while the second term expresses the expected entropy of the prediction $y$ given $x$ and the model parameters $\omega$ averaged over the model posterior $p(\omega|\mathcal{D}_{train})$. This score increases when the model shows a large disagreement about the respective category. In large neural networks Bayesian inference is infeasible, which is why BALD is typically implemented using Monte Carlo Dropout [69]. This approach uses Dropout layers with fixed dropout rate during training and during inference time, leading to multiple different predictions that can be seen as a proxy for Bayesian inference [70]. In our simulation, we select eight stochastic forward passes and a dropout rate of 0.2. The Dropout layer is implemented between the encoder and decoder of the neural network architecture as in [20], where it is used to obtain uncertainty estimates of PV segmentation masks. The uncertainty is averaged using a simple sum over all pixels, though in case of static input image sizes, averaging and summation are identical (see Appendix A).

### 2.2. Core-set acquisition function

In contrast to uncertainty-based acquisition functions, diversity-based approaches aim to find the most representative sub-dataset of the full dataset. This assumes that to obtain best model performance, a model has to see all the diversity inherent in the full dataset. In the case of PV systems, this acquisition function ideally would propose objects that are representative of the full range of plausible PV systems in different environments. For instance, this might mean various suburban, urban or rural regions.

The selection of the most promising images following this approach can be formulated as an NP-hard optimization problem, referred to as the **Core-set** approach [71], making it intractable for large-scale datasets and for usage in operational contexts. [71] propose a simpler heuristic, the k-Center-Greedy algorithm, to obtain a solution of similar accuracy as the original optimization problem that works as follows: Choose a number of center points that minimize the largest distance between a data point and its nearest center and iteratively repeat until the batch of queried images is complete. Given that the embeddings are used to describe the distance of images, reasonable embeddings are crucial for Core-set efficiency. We tested multiple different embeddings in the model architecture and chose the output from the activations of the final convolutional layer of the encoder block in the U-Net architecture.

## 3. Experimental design

### 3.1. Datasets

The images used in this study are the same as those used in [45], which is a collection of six different aerial imagery datasets from different regions in the world. More specifically, one dataset covers the region of Northern Germany [20], two datasets are from France [72], and three datasets come from the United States [45,73]. These datasets are recognized for their limited generalizability, indicated by the low model performance when the DL models are evaluated in a different region than they were trained on [45].

More information about the six datasets is provided in Table 2. In total, around 100,000 images are available. The images without any PV are referred to as negatives while those with PV panels are referred to as positives. A notable observation across all datasets is the predominance of negatives over positives, highlighting a strong imbalance toward negative images. This imbalance is further exacerbated when considering the proportion of pixels depicting PV panels. In cases where exclusively positive images are considered, less than 2% of all pixels are covered by PV panels. When negatives are included, between 0.06% to 0.86% of the pixels show PV panels. This means that, in the most extreme scenario, fewer than one out of every 1000 pixels shows a PV panel.

Another interesting difference between the datasets is the different image resolutions. The spatial resolutions are between 0.15 and 0.45 m/px showing the large spread of horizontal resolutions. In the context of PV panels, this means that a PV panel of the same physical size covers nine times more pixels in the FR-G dataset than in the CA-F dataset. Using different resolutions provides the opportunity to study the realistic case of having different measurement devices for different regions, as typical in different airborne campaigns or satellite products. All images were resized to the same dimension (320 by 320 pixels) to combine the different datasets and to reduce model training time.[1]

Besides resolution, the datasets also differ in terms of their color balance and regional characteristics. For this purpose, Table 2 lists the standard deviation of the red, green, and blue channels as an indicator of the spread of color intensities. The datasets from Germany (DE-G) and New-York (NY-Q) show the highest variability, while the Californian datasets show the lowest variability.

To gain a better understanding of the different datasets, Fig. 2 depicts the dataset in the feature space of a U-Net model after model training in a 2D-reduced space by applying principal component analysis (PCA) on the 512 embeddings. The first principal component separates the dataset into urban regions with dense population (negative values), industrial regions (around zero) and rural regions (positive values). The NY-Q and DE-G datasets exhibit a wide range of values, ranging from highly dense to rural regions. The French datasets (FR-I, FR-G) display regions with lower population densities, tending toward more rural areas, while the Californian datasets primarily represent residential and industrial areas. Only NY-Q and DE-G have values on the left half of the plot, where densely populated residential houses are visible, showing the uniqueness of these two datasets.

### 3.2. Model training

For the model simulations in this study, we apply the same DL model architecture, namely the U-Net model originally proposed by [74]. This architecture has been selected due to its wide usage for PV detection as identified in the literature review. The U-Net architecture was originally proposed for segmentation in biomedical imaging, but found wide adaption in other domains. The model architecture is depicted in Fig. 3 showing the typical encoder–decoder structure of the U-Net model. The encoder path compresses and transforms the input from a high-resolution space to a high-dimensional feature space. In the decoder, these features are then upsampled, concatenated with the resolution-wise matching feature maps and convolved to halve the number of features multiple times until the final segmentation map is obtained. This final map has the same resolution as the inputs and thereby allows a semantic segmentation of each image pixel. The roughly symmetric encoder–decoder network forms a U-shape, giving it its name.

The U-Net is trained with a batch size of 32 images using the Adam optimizer with a learning rate of $10^{-4}$. The chosen loss function is the dice score loss function which measures the overlap between

[1] Note that these dimensions are only half as detailed as in [45] making the final model scores in [45] not comparable to the final scores in this study.

**Table 2**
Datasets used in this study.

| Dataset | CA-F | CA-S | DE-G | FR-G | FR-I | NY-Q |
|---|---|---|---|---|---|---|
| Location | Fresno (California) | Stockton (California) | Oldenburg (Germany) | France | France | New York |
| References | [73] | [73] | [20] | [72] | [72] | [45] |
| Number of training images (positives) | 4193 | 1045 | 1324 | 13 304 | 7684 | 1008 |
| Number of all training images | 26 368 | 6016 | 10 379 | 28 807 | 17 325 | 6336 |
| Number of test images (full dataset) | 200 (1000) | 200 (1000) | 200 (1000) | 200 (1000) | 200 (1000) | 200 (1000) |
| Share of positive pixels (with negatives) | 0.38% (0.06%) | 0.36% (0.06%) | 0.95% (0.12%) | 1.85% (0.86%) | 0.61% (0.27%) | 1.48% (0.23%) |
| Image resolution [m/px] | 0.45 | 0.45 | 0.27 | 0.15 | 0.3 | 0.23 |
| $\sigma$ (RGB), only positives | (44,35,30) | (45,36,37) | (61,54,51) | (54,50,48) | (48,42,42) | (61,56,57) |

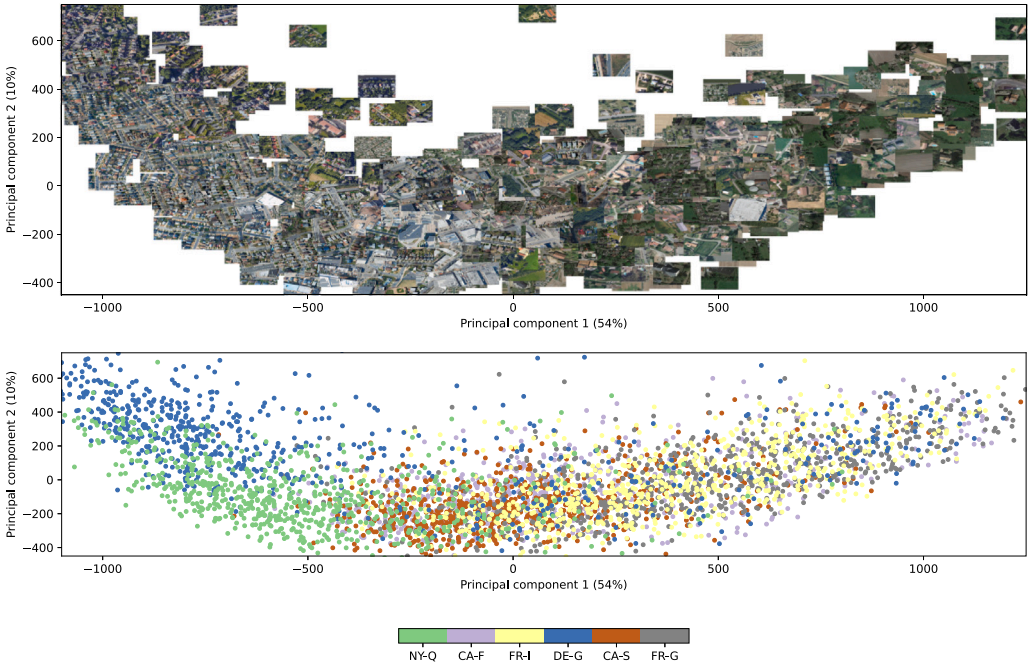



**Fig. 2.** Dataset description in the feature space. The feature space is created by running inference on the positive samples with a trained U-Net model and a subsequent PCA (explained variance in brackets) on the U-Net model embeddings (512). The upper plots show around 1000 images of the positive images in the feature space, while below each image represents one dot in feature space colorized by the dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

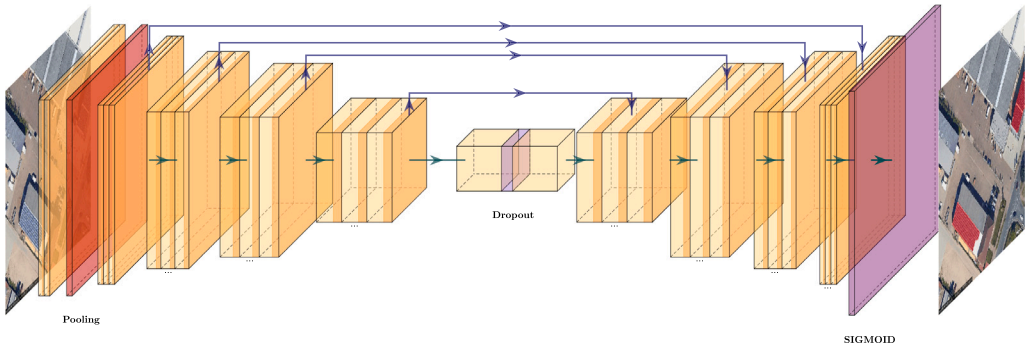



**Fig. 3.** Schematic image of the U-Net architecture used for the object segmentation in this study.

predicted and actual segmentation masks. The encoder is based on the resnet-18 architecture with pretrained weights from the ImageNet competition [75], known to increase the performance for remote sensing tasks [76,77]. The model is implemented using the `segmentation models` library [78].

For model training, we run a sufficiently large number of rounds until the model has converged (maximum of 215 rounds) and always query a batch of 16 additional images. Note that the number of queried images is different than the batch size. With these 16 additional images, the model is retrained at each round until model convergence, which is determined by the early stopping criterion [79]. This means that the validation IoU is inspected and the model is stopped when it did not improve on for three consecutive iterations. For validation data, a subset of the training data is used (20%), but the same validation data are used in the different model runs.

### 3.3. Model verification

In semantic segmentation, the main goal is the correctness of the predicted and the segmented masks. This is measured by the overlap between predicted $A$ and actual ground truth masks $B$ which can be expressed by the Intersection over Union (IoU):

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

While the IoU measures how accurately the predicted and actual masks overlap, it cannot provide information about the completeness and reliability of the model. Therefore, we use two additional metrics, namely precision and recall. The precision provides information on the reliability of the model by calculating the ratio between correctly predicted PV panels (True Positives) and the total number of predicted positives:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

In other words, the precision measures how often the model is correct when it makes a positive prediction. The completeness of the model, meaning how many of the labeled pixels showing PV panels are detected by the model, is measured by the recall. It can be calculated by the ratio between correctly predicted PV panel covered pixels and all pixels covered by PV panels:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

These three scores provide a detailed view of the model's performance and are commonly used in semantic segmentation.

### 3.4. Simulating and evaluating deep active learning

In order to demonstrate DeepAL following a human-in-the-loop modality, as depicted in Fig. 1, we simulated the behavior of human labeling by iteratively selecting annotated images from the labeled datasets. From a starting point for the model, images from the complete training set were ranked by the DeepAL acquisition function, treating each as if it were an unlabeled image. The acquisition function predicted which of these images would provide the most information to the model. While in an operational setting, these images would be shown to an annotator, the availability of labeled datasets allowed us to directly access the labeled masks for these images. The model was then retrained with the inclusion of these images and the process was repeated.

The effectiveness of DeepAL is evaluated based on first training a baseline model on randomly sampled images at each acquisition round. Comparing the baseline model performance at each training step against the performance of a model trained with a more sophisticated acquisition function helps us to determine whether images intelligently selected by DeepAL are more valuable than those chosen at random. This approach is commonly applied in the Active Learning literature [59,69]. As a second benchmark, we train the models on the entire annotated data to obtain an estimate of the DeepAL's performance compared to a model with access to the full dataset. The comparison of the final DeepAL's model performance with the score using the full dataset indicates whether DeepAL can reach the same model performance with less data.

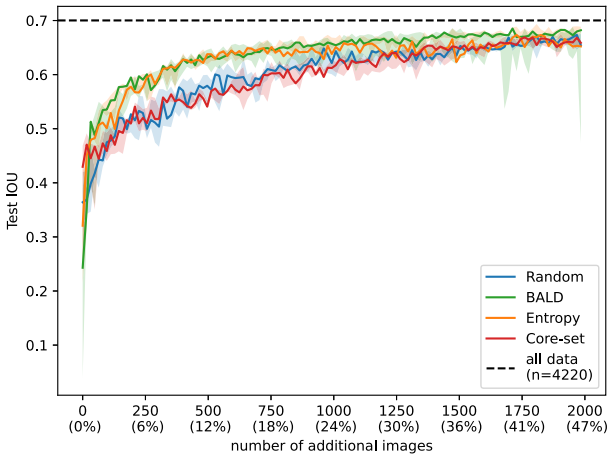



**Fig. 4.** Performance for the controlled study only using positive samples. The min–max bands and the median are constructed using three different seeds.

### 3.5. Computational implementation

The computational study is implemented using the `snakemake` library [80] as a reproducible workflow. The simulations are executed on a GPU cluster with five GPUs (2× Tesla P100, 2× Tesla V100 and 1× Tesla A40). Each simulation is repeated for three different seeds to estimate the sensitivity to randomness. The different simulations are explained in the results section for the sake of clarity.

## 4. Results

### 4.1. Choosing the acquisition functions

#### 4.1.1. Performance on positive images

As a first step, we consider only positive images from all different locations to evaluate the performance of the three acquisition functions (BALD, Entropy and Core-set). The results of the simulations are depicted in Fig. 4. Between the acquisition functions, Core-set shows the worst performance, with no significant improvements over random sampling. This makes it ineffective for the choice of an acquisition function in the context of detecting PV panels.

The uncertainty-based acquisition functions (BALD and Entropy) outperform Core-set and random sampling. This is evident by the higher test IoUs in the final iterations. As the final test IoUs are close to the baseline model, the uncertainty-based methods show that they can reach the model score of the model trained on the full dataset. This is important for practitioners, as it shows that not only DeepAL is more efficient in selecting appropriate images, but these images contain enough valuable information that the model can achieve the full model performance. In addition to better final scores, uncertainty-based acquisition functions can reach higher accuracy with many fewer images than random sampling. In numbers, both BALD and Entropy can reach test IoUs up to 0.6 with only 6% of the images while randomly sampled images reach only 0.5 with the same amount of images.

#### 4.1.2. Insights from queried images

To better understand how the acquisition functions differ, we investigate which images they queried during the first two rounds, as depicted in Fig. 5. The randomly selected images reflect the image diversity in the datasets with respect to spatial resolution, color contrast, topographic differences, and PV panel types. Furthermore, they show multiple images of urban regions with typical residential houses. This can be explained by considering that the dataset composition has a strong focus on suburban regions (DE-G, NY-G, CA-F and CA-S) or in the case of FR-I and FR-G, is based on known PV locations [72]. Similar to random sampling, Core-set shows a diverse set of images such as suburban, industrial buildings and agricultural fields. In contrast to Core-set, the uncertainty-based acquisition functions (Entropy, BALD) query more homogeneous images with a special emphasis on agricultural structures. This is remarkable, as there are only a few images of agriculture in the datasets, especially in the positive samples that need to show at least one PV panel by definition. A plausible explanation for this focus is that the geometrical structure of agricultural fields, which are organized in parallel lines, leads the model to exhibit high uncertainty in discriminating between these rows and rows present in large-scale PV systems. In addition to fields, Entropy and BALD query regions with large-scale PV systems which can be explained by the aggregation within the acquisition function, which favors images with large areas of high uncertainty.

As the full dataset is a collection of different locations, we can investigate which sources the acquisition functions query from during the different rounds, as depicted in the first row of Fig. 6. Note that by comparing the final share with the shares of random sampling, over- and undersampling of a dataset can be recognized. Based on the cumulative number, the Core-set approach significantly overrepresents the NY-Q and DE-G datasets, especially during the early rounds. Interestingly, these two are also the two datasets with the largest spread of RGB values (Table 2). In the second row of Fig. 6 the distribution of queried samples is depicted in the feature space derived from the U-Net model embeddings. The majority of images queried from the uncertainty-based methods stem from the region where the first principal component is close to zero. In this region, there is the overlap of the different datasets with a large variety of plausible image realizations such as industrial buildings, regions with rural regions and images with image-contrast such as the FR-G images.

Although the distributions of the uncertainty-based methods look similar and are similar to random sampling, the Core-Set approach shows a large oversampling of the high-contrast images on negative values of the first principal component where only the NY-Q and DE-G datasets are present. By contrast, low contrast images, such as rural regions with larger values for the first principal component, are not considered during the querying. A plausible explanation for this overrepresentation is that in the early iterations, the embeddings from the neural network are not expressive enough to be able to create differences between low-contrast images. As Core-set builds clusters based on the diversity of the embeddings, a high-diversity dataset may also lead to more clusters compared to other datasets, and thus finally to more querying. In contrast, more homogeneous datasets (CA-F, CA-S, FR-G and FR-I) with smaller color palettes lead to a smaller spread of embeddings and thus to fewer clusters. When thinking about impacts on model performance, the opposite is true, as objects from low-contrast images are harder to discriminate against the background than high-contrast images. This raises some questions about the usefulness of the Core-Set approach for the sake of PV detection.

The uncertainty-based acquisition functions in Fig. 6 show fundamentally different behavior. In the early rounds (< 500 images), the model prioritizes querying from datasets with highest spatial resolutions (FR-G, FR-I and NY-Q). As the training progresses, Entropy increasingly queries from low-resolution datasets (CA-F and CA-S). This behavior can be attributed to the model's training state. In the initial iterations, the model struggles to discriminate PV panels from the background and therefore focuses on large-scale PV systems, which also generate a large uncertainty due to their large covered area. Once the model is able to easily identify these, it shifts its focus to more challenging datasets with lower resolution. Furthermore, BALD and Entropy sample very similarly at the different rounds, corresponding to the similarity in their uncertainty-based approaches. More specifically, of the first 1500 images queried by Entropy and BALD, there is an overlap of 37% of sampled images, while Core-set sampling would result in only 25% and random sampling in only 23% of overlap.

#### 4.1.3. Choosing a suitable acquisition function

Table 3 summarizes the findings of the above analysis and the characteristics of the different acquisition functions. In addition to performance, we include storage requirements and computational expense to evaluate the acquisition function's ability to scale to large-scale datasets, as needed for a global PV registry on submeter resolution. Storage requirements determine whether very large datasets can be evaluated, while computational expenses determine the time a human-in-the-loop annotator, such as described in [59], would need to wait until new images are proposed. As remote sensing can sample from a nearly unlimited set of geographical regions and sensors, the time spent to evaluate the acquisition function could easily become the bottleneck in an operational system.

With respect to the simplicity of the implementation, Entropy does not require a specific model architecture, as it uses the already available probability outputs from the sigmoid or softmax layer. By contrast, BALD is based on Monte Carlo Dropout and requires Dropout layers to obtain uncertainty estimates. While Deep Ensembles could also be applied in BALD for networks without Dropout layers [81], this requires the training of multiple DL models and increases the computation time during inference. In the case of Core-set, the DL model is required to have extractable, meaningful embeddings. These embeddings are critical, because they determine the feature representation, but fortunately, embeddings are already available in the model architecture. Entropy has the lowest computational expenses, as only one forward pass over the unlabeled pool is needed, and the calculation of the Shannon entropy is not a new operation, as it is also performed during model inference. The image scores do not need to be kept in memory, allowing extensive parallelization with low storage requirements. In the case of BALD, the computational costs are multiples of the entropy method, scaling as a function of the number of desired uncertainty estimates. Core-set needs to calculate the distance matrix over all embeddings, resulting in large computational and memory expenses.

In terms of performance as an acquisition function, uncertainty-based methods outperform the diversity-based method by a large margin for the case of detecting PV panels. A plausible explanation for this can be derived from the characteristics of the dataset used in this study. In [67], different DeepAL algorithms are benchmarked on three different datasets that are classified into diverse and redundant datasets. In the case of redundant datasets, uncertainty-based methods often propose highly correlated batches that hinder the learning of reasonable models, which is also known as mode collapse [67]. Diverse datasets do not suffer from local clusters, which leads to uncertainty-based acquisition functions outperforming their diversity-based counterparts [67]. Given that the PV datasets used here cover wide geographical areas characterized by different climatic zones, vegetation, and different sensors, we argue that the PV segmentation task implemented on these can be considered to be a diverse dataset. This is also consistent with the results in [59] for the case of building segmentation, where, although only uncertainty-based methods have been compared, these show large performance improvements over random sampling.

Given the high performance of Entropy, its easy implementation within the existing neural network architecture and the low computational and storage costs (Table 3), we suggest Entropy as the preferable method of those tested, and focus on Entropy for the remainder of this study.

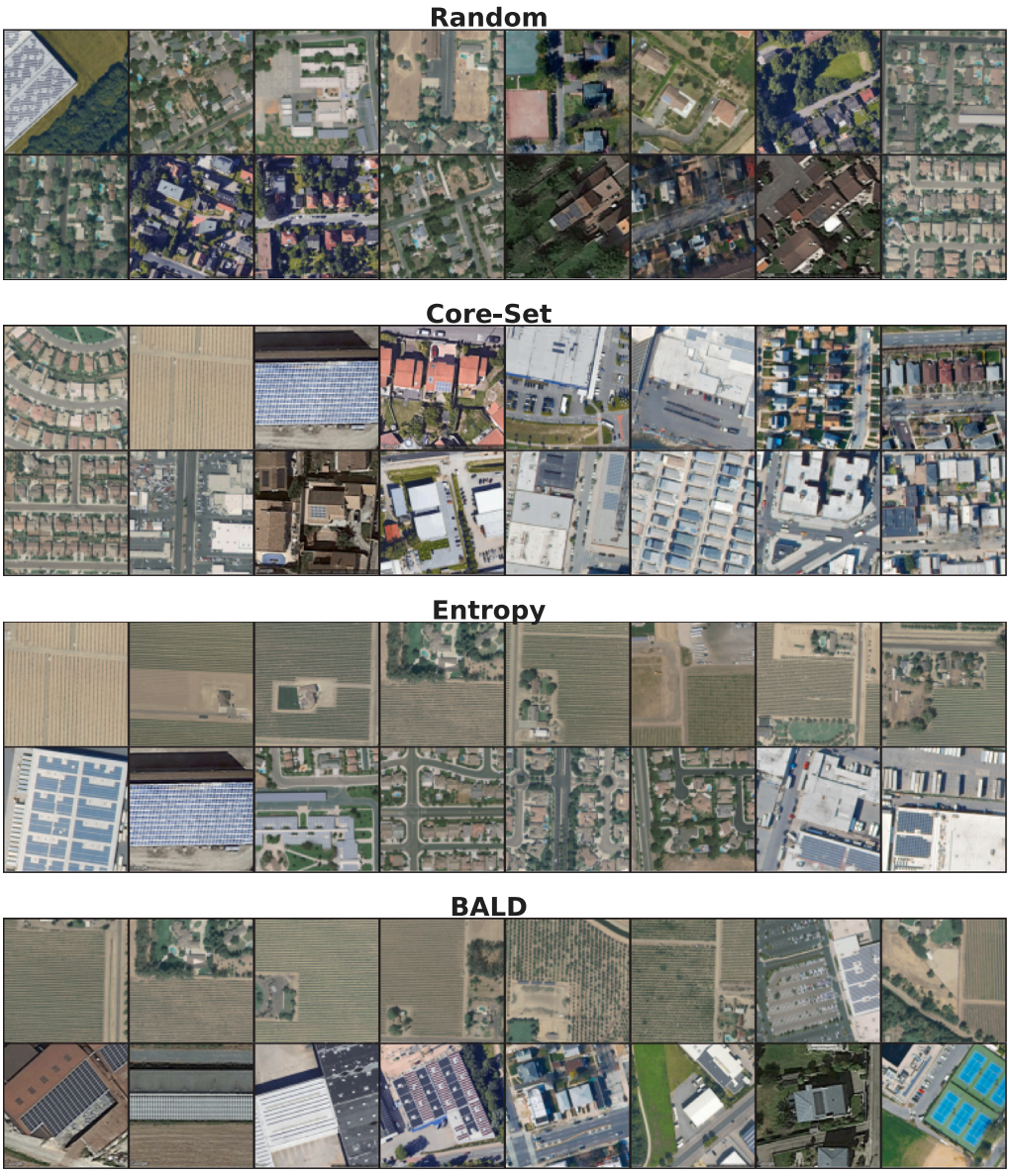



**Fig. 5.** Selected images of the different acquisition functions compared to random sampling as a reference. The two rows indicate the first two queried batches of images, only every second entry is selected for visibility.

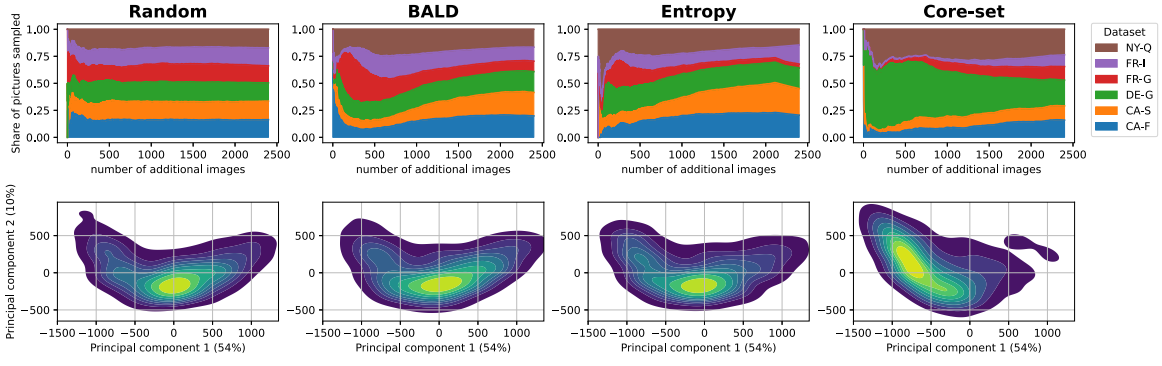



**Fig. 6.** Datasets sampled from when using different acquisition functions. Only positive samples are used, a representative seed out of the three is selected as all are similar. The second row depicts the 2D Kernel density estimator of the first 800 queried images represented in the U-Net feature space (reduced by PCA). Lighter color indicates more sampled images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
Comparison of acquisition functions. Favorable properties are printed in bold.

| | Entropy | BALD | Core-set |
|---|---|---|---|
| Acquisition function type | Uncertainty | Uncertainty | Diversity |
| Improvement | **High** | **High** | Medium |
| Simplicity of implementation | **High** | Medium | Low |
| Need to change NN architecture | **no** | eventually | **no** |
| Computational requirements | **Low** | Medium | High |
| Storage requirements | **Low** | Medium | High |

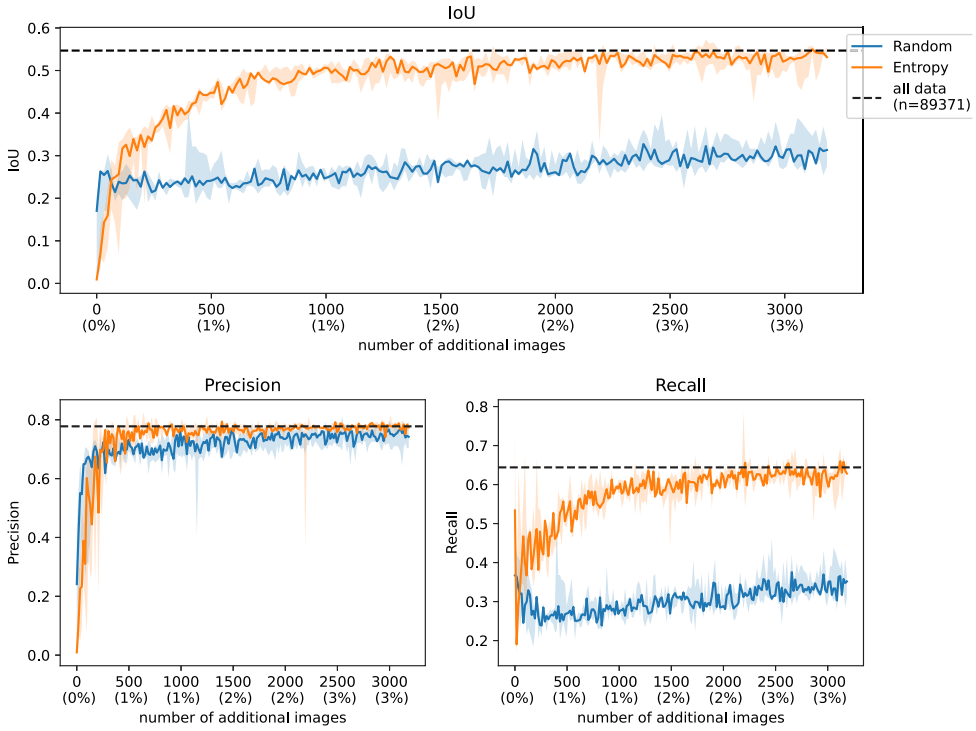



**Fig. 7.** Performance of DeepAL for imbalanced datasets. The min–max bands and the median are constructed using three different seeds.

### *4.2. Imbalanced data*

#### *4.2.1. Performance*

The simulations thus far assume that the unlabeled pool only contains positive images, meaning that each image contains at least one PV system. However, this approach does not fully encapsulate the challenges inherent in a realistic PV segmentation dataset that spans over wide regions and would not be expected to show a PV system in most images. To address this limitation and to provide a more realistic evaluation, we investigate the effect of imbalanced datasets by extending the datasets from Section 4.1.1 by their negative samples (66674 additional images, leading to a share of 70% negative images). This also applies to the test dataset which is extended by 800 negative images for each region (4800 in total).

Fig. 7 shows the performance of Entropy compared to random sampling. Interestingly, the performance disparity between random sampling and Entropy is stronger than when only positive samples are considered (Section 4.1.1). This was evident from the onset, as Entropy largely outperforms random sampling after around 100 queried images, and then consistently outperforms random sampling as additional images are added. In addition to scoring much better during the initial iterations, Entropy converges to much higher performance levels than random sampling. In numbers, Entropy can reach a test IoU of 0.55 while random sampling only reaches values between 0.25 and 0.3, around 50% lower. Remarkably, the final IoU scores of Entropy are equal to the model trained on the full dataset, which means that by using Entropy, only 3% of the total data are needed. This number directly corresponds to the number of images that a labeler would need to inspect, meaning that an annotator would need to observe 97% fewer images by using Entropy.

The IoU only provides information about the overlap between the estimated and actual masks. To obtain more information about the model performance, Fig. 7 shows the pixel-wise precision and recall values as a measure of how complete (recall) and reliable (precision) the model predictions are. The precision curve illustrates large precision values for Entropy and random sampling, even in the early iterations. Entropy plateaus quickly after less than 500 iterations (0.8) at the value of the baseline while random sampling only slowly converges to the final model score over the 3000 sampled images. The recall curves are much flatter than the precision curves. The final model scores are only reached at the simulation end of around 3000 images for Entropy, while random sampling never reaches the final score. On the contrary, the final recall score of randomly sampled images is 50% lower than achieved by Entropy. Furthermore, the recall curve is highly similar to the IoU curves for both Entropy and random sampling. This precision–recall relationship indicates that the model is highly conservative, meaning that it only makes predictions when it is highly certain about them (high precision). This leads to a smaller number of

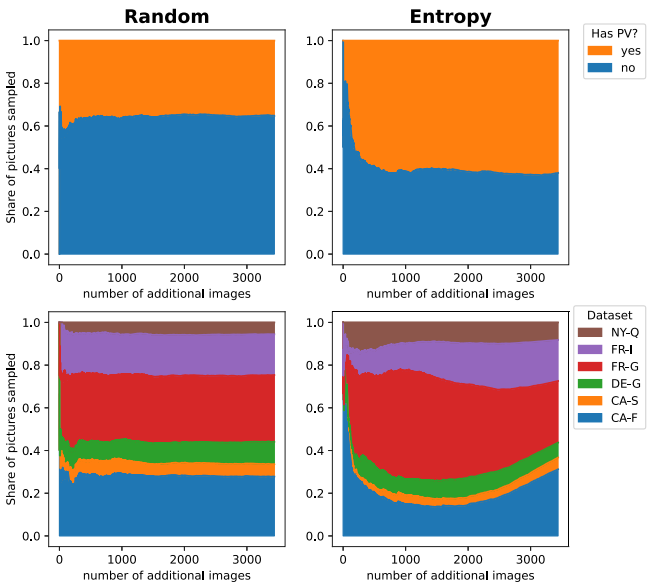



**Fig. 8.** Illustration of the cumulative shares of the dataset origin during sampling in terms of positive/negative samples (upper row) and the dataset location (bottom row).

predictions, making the results incomplete (low recall) during the early iterations. The model needs to incorporate a diverse set of PV systems in its training data to improve performance. Entropy's approach actively provides a diverse set of different PV systems that allows the model to improve quickly, while for random sampling, many more images are needed to obtain a sufficient set of diverse PV systems.

Another interesting observation in Fig. 7 is the low spread between the different model runs in the case of Entropy. Therefore, we can conclude that Entropy reliably leads to the same model performance during different model runs. This is crucial, as it means that the modeler can trust the result of individual training runs when using Entropy in a DeepAL approach.

#### 4.2.2. Insights from queried images

Analogous to the analysis in Section 4.1, we investigated the types of images sampled during the DeepAL rounds but when negative images are included, as shown in Fig. 8. During the early rounds, Entropy oversamples the FR-G dataset, similar to observations from the case of positive images. During later rounds, the ratio between the different datasets approaches the random population. This implies that Entropy is able to adapt by shifting its focus from the difficult images of the French datasets to other datasets in later stages. Around 90% of the performance improvements are achieved within the first 1000 additional images (Fig. 7), which aligns with the rounds when most sampling is from the FR-G dataset. This shows the importance of including the very difficult cases in the early rounds. In later rounds, a more diverse querying leads to a well-balanced performance over the different regions. In addition to the different datasets sampled from, Fig. 8 illustrates the balance between positive and full dataset. An oversampling of positive images is noticeable. Random sampling would be expected to result in a ratio of about 40% positive images, while Entropy instead shows 60% positive split. Interestingly, primarily negative samples are queried during the early iterations, while after around 100 iterations, an increasing number of positive images are sampled. The oversampling of positive samples shows that after only a few rounds of queried images, the model is able to propose images that are more likely to show PV panels than in the case of randomly sampled images.

The images with the highest uncertainty scores using Entropy are depicted in Fig. 9. These allow us to investigate how the model uncertainty changes over the course of the model training for both positive images and the full dataset. The first images contain particularly large-scale structures meaning large-scale PV systems (P1), agricultural fields (P2, P3, F1 and F2) and measurement error (F3). As discussed with the positive-only datasets in Section 4.1.1, these large-scale structures lead to a high value of aggregated uncertainty. Furthermore, these images show that parallel geometries similar to PV systems create large model uncertainty during the early rounds. This also applies to parking lots (P4, F4 and F8) with parallel-oriented cars, and train railways (F7).

Interestingly, the images with the highest uncertainty often do not show any PV panels, or there is only a PV system close to the object by coincidence (e.g. the primarily agricultural areas in P2 and P3), that leads to an inclusion of these images in the training dataset. This is crucial as it shows the importance of including these negative images to train a robust DL model. Furthermore, in case of the imbalanced data, the DeepAL method provides images, even at early iterations, that are also non-trivial to label for untrained human labelers. For example, this can be seen in selected images with low contrast differences between PV panel and background (P5). This shows that the uncertainty estimates from a model with a small training dataset and only a few training iterations can already query highly uncertain images. It is interesting that large-scale objects can still be challenging for the model even when it already shows reasonable detection skills. For instance, this applies for images 1400 to 1600, when the model is close to its final model IoU (Fig. 7), but negative images such as a large-scale agricultural field (F10) are still queried. This highlights the crucial importance of including negative samples to train robust segmentation models, and that Entropy is able to detect these negative samples. This is consistent with the fact that around 40% of images sampled are negative samples, even when the model has reached relatively sophisticated IoU scores at 1000 images.

### 4.3. Application across different PV detection models

#### 4.3.1. Different PV detection models

Academic studies mainly assume that DL models are trained from scratch, as investigated so far in this study. However, a PV detection model applied in an operational context needs to be frequently updated to extend the spatial domain of an existing model, or to account for updated measurement sensors. This is a crucial difference, as it is known that neural networks often perform poorly under spatial domain shifts for the detection of PV panels [45]. As a last section in this study, our objective is to evaluate the performance of DeepAL in the context of different PV model operational variants.

Fig. 10 depicts two other training variants, inspired by [82], besides training a model from scratch. **Fine-tuning** refers to the task of adapting an existing model, which was trained from scratch, to a different task without utilizing the data from the initial task. A plausible scenario for fine-tuning is the development of a specialized model for a region or the update of a model to a new generation of a dataset. Fine-tuning a model requires fewer data than training a model from scratch, which makes fine-tuned models highly applicable in remote sensing [83] and in PV detection [54]. **Joint-Learning** is similar to fine-tuning, as it also transfers an existing model to a novel task with updated labeled data, but it keeps the training data from the initial task. The intent is that the model should perform well for both the original and novel tasks, leading to a more sophisticated general model [82]. This is the scenario we consider for the development of a global PV inventory, in which a model might be iteratively extended by dataset or region.

In the following, we investigate the performance of Entropy for fine-tuning and joint-learning approaches. For the sake of clarity, we only make the comparisons using the full datasets that include negative samples. Further, we merge the datasets by location and dataset similarity from six to three, leading to a French dataset (FR: FR-G,

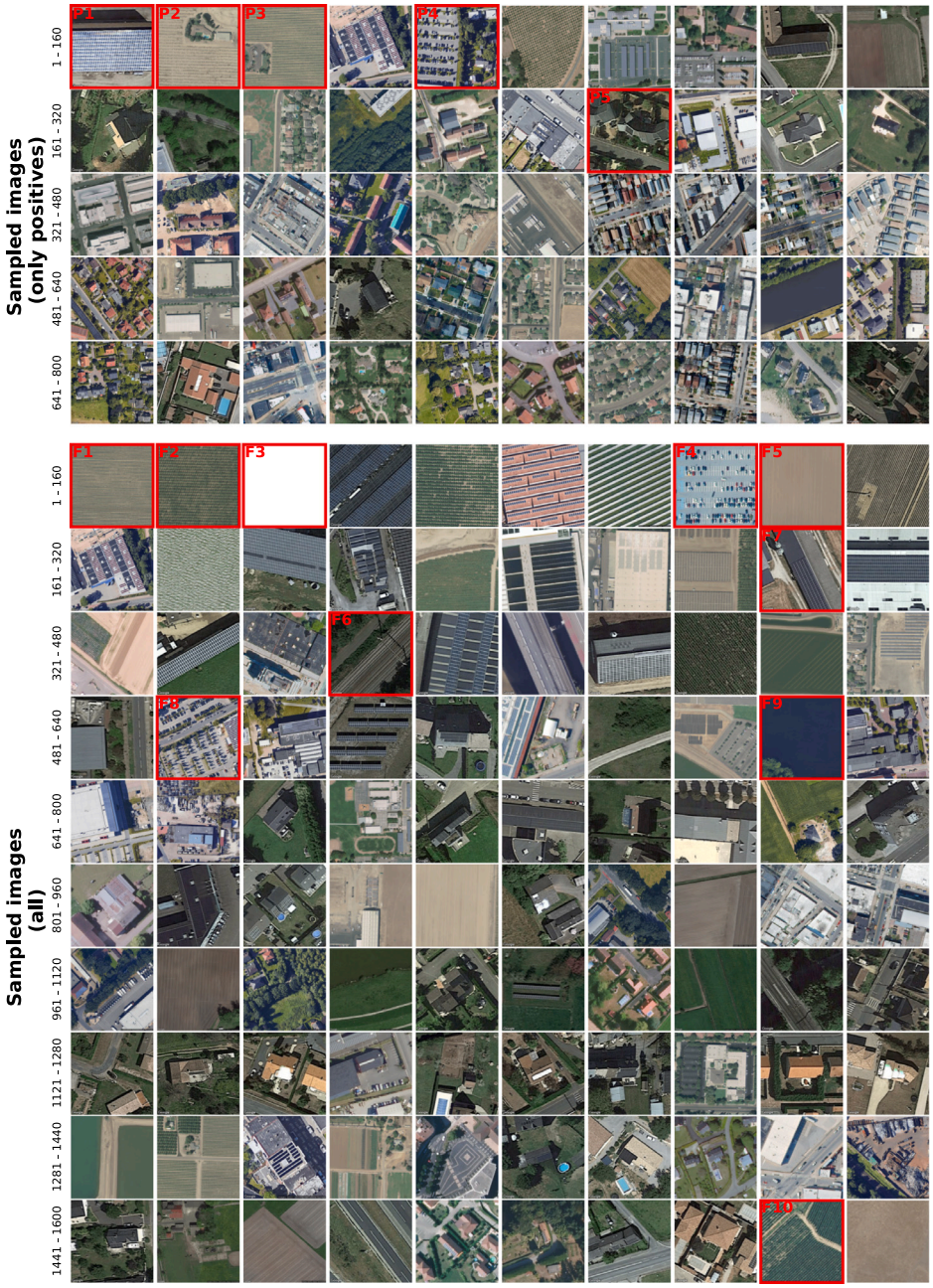



**Fig. 9.** Image with highest uncertainty for the first 50 queried rounds (first image entry of the respective batch) for Entropy. Annotated images with a red border and captions are mentioned in the main text. The annotations at the y-axis of the first image in each row describe how many images are represented by the respective rows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

FR-I), a Californian dataset (CA: CA-F, CA-S) and a German-American dataset (DENY: NY-Q, DE-G). For instance, this means that for the case of fine-tuning an existing model from France to California would result in fine-tuning an existing model trained on the full French dataset with the full Californian dataset. As previously, we train a model with the full dataset as a baseline for each location.

### 4.3.2. Performance

Fig. 11 shows the performance for the different remote sensing tasks. The rows indicate the datasets on which the base model is trained, while the columns represent the datasets used for fine-tuning or joint-learning. Similarly to training from scratch, Entropy shows much better performance than random sampling for both fine-tuning and joint-learning. Likewise, most improvements occur in early iterations, with fewer than 1000 images. In case of joint-learning with the additional CA dataset, the DeepAL model very quickly reaches the final performance level, while random sampling struggles to reach equal performance to the baseline. This shows that Entropy is able to select the difficult images from early on. In all cases, the model performs similarly to the baseline model, with some simulations even exhibiting higher test IoU scores. Conversely, random sampling performance plateaus at low test IoU scores for some cases (e.g. joint-learning DENY-FR/FR-DENY). These simulations show that Entropy not only works well when models are trained from scratch, but also when transferred to other

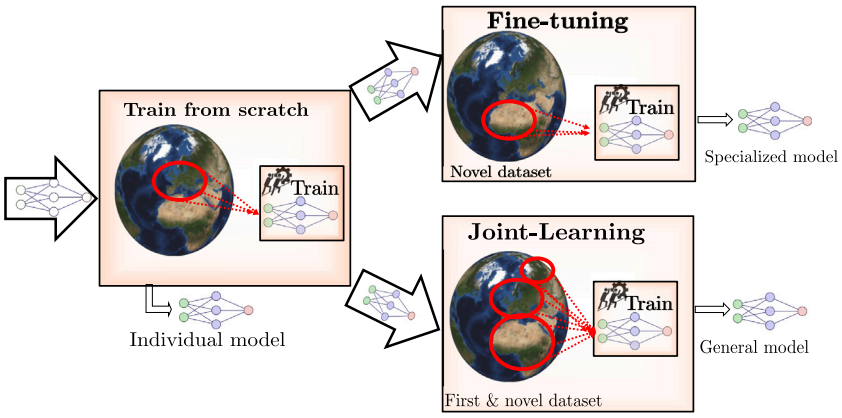



**Fig. 10.** Usage of DL models for different approaches to remote sensing model variants. Colored neural networks indicate trained neural networks, while the white neural network represents an untrained one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

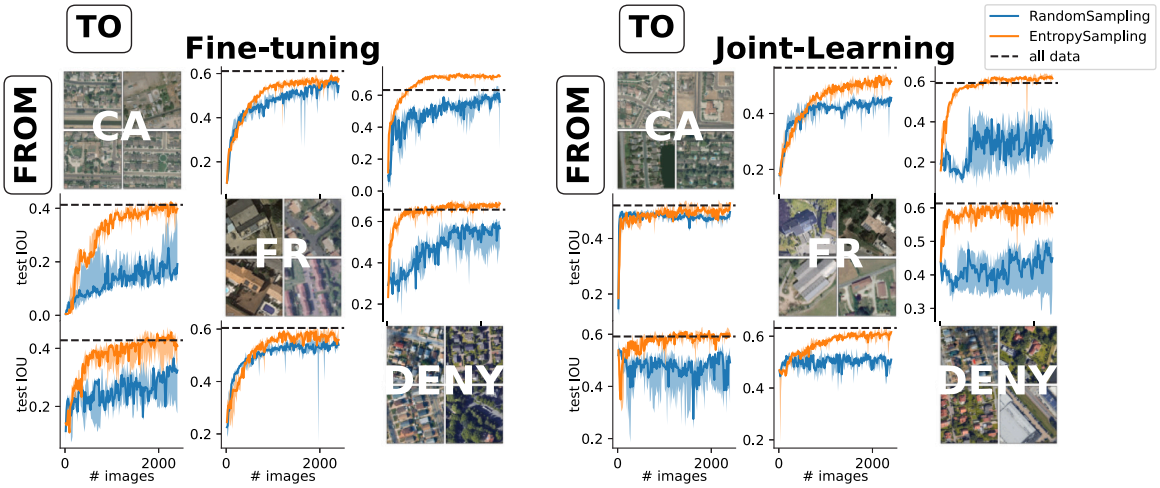



**Fig. 11.** Performance of different remote sensing model variants (fine-tuning and joint-learning) including negative samples. The rows indicate on which dataset the model is trained on, the columns indicate the dataset used to fine-tune or which is included in joint-learning.

model training variants. This makes it highly attractive for large-scale PV detection models.

## 5. Discussion

In this study, we demonstrate that DeepAL is highly promising for facilitating the training of supervised DL models for PV panel detection, by improving data efficiency and effectiveness. Through DeepAL, the number of labeled images required can be significantly reduced, as shown in several simulations. In the most realistic scenario, where there is no prior knowledge of whether an image contains a PV system, only 3% of the images are needed to achieve performance comparable to fully labeled datasets.

Beyond the considerable reduction in annotation effort, DeepAL provides reliable feedback to annotators. This is reflected by the minimal sensitivity of test IoUs to different model simulations and the nearly monotonically increasing performance trend as more images are annotated. For the simulations conducted, the performance of models trained with DeepAL plateaus after around 1000 to 1500 images, which represents a moderate labeling effort. We also show that DeepAL excels in the context of fine-tuning and joint-learning with novel datasets. When additional datasets contain limited new information, DeepAL can filter relevant images early, making it highly attractive for developing tailored and generalizable PV detection models. Further research is needed to determine whether DeepAL can similarly reduce labeling effort in larger-scale models, such as those with more than a million positive images. Nevertheless, the underlying mechanisms of DeepAL, as proposed in this study, should remain independent of the size of the model.

Additionally, we show that the choice of the acquisition function is crucial in DeepAL, as it determines the value of labeling an image for model training. Our results indicate that uncertainty-based methods outperform the investigated diversity-based method (Core-set) for segmenting PV panels. By examining the first two batches queried by the acquisition function and the sampling in the model feature space, we observe that Core-set tends to oversample from datasets with more information, which are not necessarily critical for improving model performance. In contrast, uncertainty-based methods excel at identifying challenging, low-contrast images, where PV panels are harder to distinguish from the background, early in the process. This is likely due to the characteristics of the aerial PV datasets, a condition under which uncertainty-based strategies outperform batch-based diversity sampling methods [67]. Our findings align with previous work in remote sensing, where uncertainty-based methods are shown to be highly effective [59]. Among these methods, Entropy stands out as the simplest to implement, requiring the least storage and computational resources, making it the most efficient method without sacrificing performance. Given that Entropy is linked to model uncertainty, a potential avenue for future research could be exploring how this uncertainty might also indicate when model training should stop or when retraining is necessary, further streamlining the training process.

In this study, we used a single DL model architecture, the U-Net architecture, to maintain comprehensibility. Therefore, a plausible question is whether DeepAL is also effective for different DL architectures. We repeated some experiments with the SegFormer [84] architecture, a modern transformer-based architecture with the results explained in Appendix B. Interestingly, for the case study of training from scratch with imbalanced data, the effectiveness of DeepAL was identical for both the U-Net and SegFormer architectures. The SegFormer achieved slightly higher test IoUs, likely due to its more sophisticated architecture, consistent with findings in the literature [36]. Similar observations were made for the case of joint-learning (see Appendix B). This confirms that DeepAL is at least effective for different DL model architectures, but with the caveat that a comprehensive study across all possible architectures was not feasible here.

An important consideration in the existing literature is whether, for large-scale model deployments, it is necessary to first classify relevant image regions before applying object segmentation as in [27,85,86]. The comparisons of performance on positive-only and full datasets demonstrate that DeepAL works for semantic segmentation with or without the need for initial classification. Its success at preferentially sampling positives in the highly imbalanced data for the full dataset cases indicates that it is implicitly capable of proposing which images show a PV panel. Furthermore, we believe that the results of this study support the idea that segmentation on the datasets with negative samples increases the segmentation accuracy, as the contextual information from similar-looking objects is also relevant for the segmentation of PV panels. However, this needs further academic research. For the sake of this study, it should be noted that the effectiveness of DeepAL does not rely on this question, the same algorithms could be applied as well for classification tasks.

Lastly, the strong separation of the different datasets in the feature space, as depicted in Fig. 2, shows that the inclusion of different datasets is highly important, not only in terms of their spatial resolution, but also in terms of their dataset characteristics. This is noticeable, as ignoring the DE-G and NY-Q dataset would not result in any images from the left-hand side of Fig. 2, despite there being datasets with a lower or higher spatial resolution. An open question remains whether generalizable models are achievable with extremely large datasets and how many different datasets would be required to achieve this. In principle, the framework of DeepAL can help in this regard by identifying critical datasets.

## 6. Conclusion

In this study, we demonstrate that Deep Active Learning can reliably reduce the labeling effort to reach equal or better model performance for PV segmentation. We showed that this applies to the case of using only positive images, the case of highly-imbalanced data, and for the model training variants of fine-tuning and joint-learning. We have identified Entropy-based sampling as the most favorable acquisition function for the case of PV detection. As this study simulated a human-in-the-loop setting, in which Deep Active Learning proposes relevant images to a human annotator in an iterative setting, the next step is to develop a software that implements this modality in an operational setting. Therefore, this article provides the groundwork for the development of a computational implementation toward the detection of global PV systems with frequent update times, and considering both utility-scale and rooftop-scale PV systems.

## CRediT authorship contribution statement

**Matthias Zech:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hendrik-Pieter Tetens:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Joseph Ranalli:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The labeled data is available online (except the DE-G dataset), the code including download and setup instructions will be published on github once accepted.

Code to run the experiments and create plots, will be published on git hub once accepted (Original data) (Github).

## Acknowledgment

M. Zech has been supported by the German Federal Environmental Foundation within the PhD scholarship (grant no. 20020/667-33/2).

## Appendix A. Indifference of summation or averaging of pixels in BALD algorithm

Assuming two images (x, y) with uncertainty estimates for each pixel $x_i, y_i$ which are provided by Monte Carlo Dropout. Let the uncertainty estimates for each pixel of image $y$ be larger than for image $x$, it follows that

$$\frac{\sum_i^{n^x} x_i}{n^x} \leq \frac{\sum_i^{n^y} y_i}{n^y} \tag{A.1}$$

Given this inequality and that both images have the same number of pixels ($n^y = n^x$), the inequality can be reformulated to

$$\frac{\sum_i^{n^x} x_i}{n^x} \leq \frac{\sum_i^{n^y} y_i}{n^y} \overset{n^x=n^y}{\equiv} \sum_i^n x_i \leq \sum_i^n y_i \tag{A.2}$$

which shows that summation and averaging are equivalent within the BALD method.

## Appendix B. DeepAL for a transformer-based model architecture

To investigate whether DeepAL works for different neural network architectures, we evaluated DeepAL on the SegFormer model architecture, which is based on a transformer.

Fig. B.12 illustrates that by using 1% of the data, DeepAL is already capable of reaching 0.5 in contrast to 0.3 for Random Sampling. Note that it takes more data to train the transformer, as it has not stopped training using the same number of additional images. This is possibly due to the more complex model architecture.

In Fig. B.13, DeepAL is tested on the SegFormer architecture for the case of joint-learning. As before, the results are highly similar to the U-Net model in Fig. 10. Thus, we conclude that the performance of DeepAL is not strongly dependent on the model architecture.

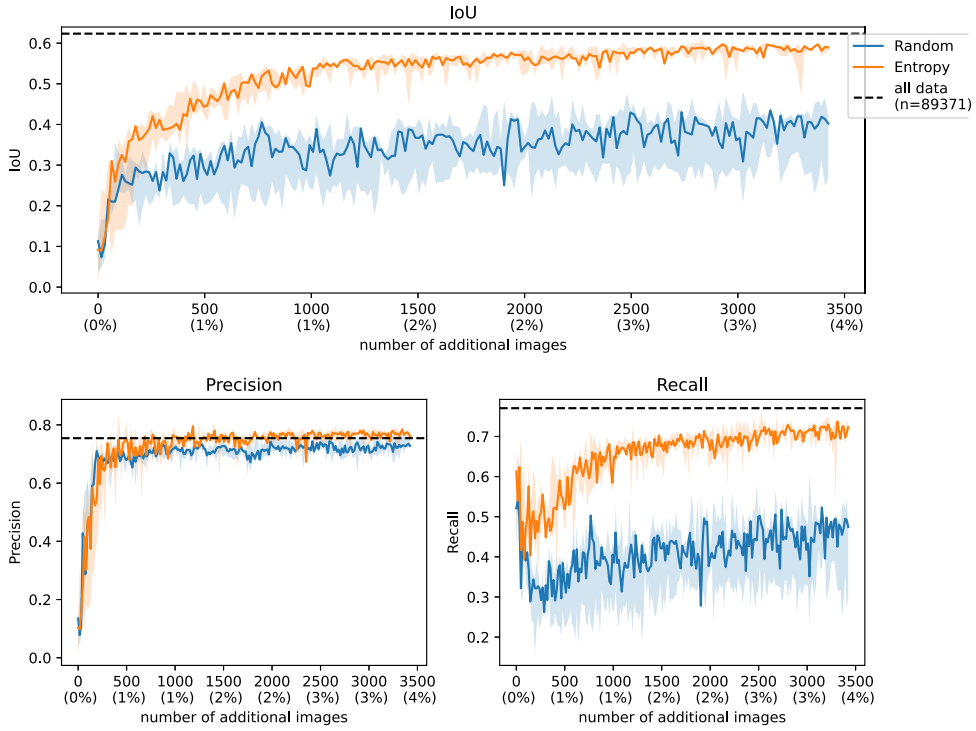



**Fig. B.12.** Same as Fig. 7 (Learning from scratch with negatives included) but for the transformer-based architecture.

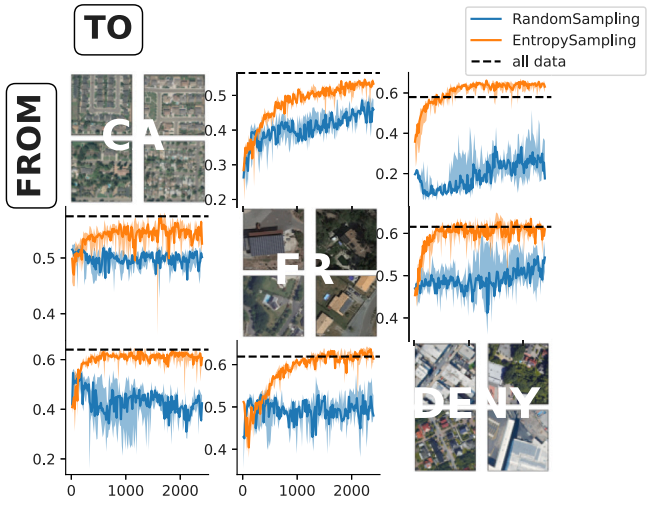



**Fig. B.13.** Same as Fig. 10 (Joint-Learning with negatives included) but for the transformer-based architecture.

## References


[1] Victoria M, Haegel N, Peters IM, Sinton R, Jäger-Waldau A, del Cañizo C, Breyer C, Stocks M, Blakers A, Kaizuka I, Komoto K, Smets A. Solar photovoltaics is ready to power a sustainable future. Joule 2021;5(5):1041–56. http://dx.doi.org/10.1016/j.joule.2021.03.005.

[2] Creutzig F, Agoston P, Goldschmidt JC, Luderer G, Nemet G, Pietzcker RC. The underestimated potential of solar energy to mitigate climate change. Nat Energy 2017;2(9). http://dx.doi.org/10.1038/nenergy.2017.140.

[3] Wilson C, Grubler A, Bento N, Healey S, De Stercke S, Zimm C. Granular technologies to accelerate decarbonization. Science 2020;368(6486):36–9. http://dx.doi.org/10.1126/science.aaz8060.

[4] Gernaat DE, de Boer HS, Dammeier LC, van Vuuren DP. The role of residential rooftop photovoltaic in long-term energy and climate scenarios. Appl Energy 2020;279(August):115705. http://dx.doi.org/10.1016/j.apenergy.2020.115705.

[5] Joshi S, Mittal S, Holloway P, Shukla PR, Ó Gallachóir B, Glynn J. High resolution global spatiotemporal assessment of rooftop solar photovoltaics potential for renewable electricity generation. Nature Commun 2021;12(1). http://dx.doi.org/10.1038/s41467-021-25720-2.

[6] Stowell D, Kelly J, Tanner D, Taylor J, Jones E, Geddes J, Chalstrey E. A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK. Sci Data 2020;7(1):1–15. http://dx.doi.org/10.1038/s41597-020-00739-0.

[7] PV.P.S. Task I. Regional Solar Power Forecasting 2020 Task 16 Solar Resource for High Penetration and Large Scale Applications PVPS. 2020, URL www.iea-pvps.org.

[8] Zech M, von Bremen L. End-to-end learning of representative PV capacity factors from aggregated PV feed-ins. Appl Energy 2024;361(February):122923. http://dx.doi.org/10.1016/j.apenergy.2024.122923.

[9] Yazdanie M, Orehounig K. Advancing urban energy system planning and modeling approaches: Gaps and solutions in perspective. Renew Sustain Energy Rev 2021;137(January 2020):110607. http://dx.doi.org/10.1016/j.rser.2020.110607.

[10] Alhamwi A, Medjroubi W, Vogt T, Agert C. GIS-based urban energy systems models and tools: Introducing a model for the optimisation of flexibilisation technologies in urban areas. Appl Energy 2017;191:1–9. http://dx.doi.org/10.1016/j.apenergy.2017.01.048.

[11] Dunnett S, Sorichetta A, Taylor G, Eigenbrod F. Harmonised global datasets of wind and solar farm locations and power. Sci Data 2020;7(1). http://dx.doi.org/10.1038/s41597-020-0469-8.

[12] Joshi B, Hayk B, Al-Hinai A, Woon WL. Rooftop detection for planning of solar PV deployment: A case study in Abu Dhabi. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8817, 2014, p. 137–49. http://dx.doi.org/10.1007/978-3-319-13290-7_11.

[13] Malof JM, Hou R, Collins LM, Bradbury K, Newell R. Automatic solar photovoltaic panel detection in satellite imagery. In: International Conference on Renewable Energy Research and Applications, ICRERA. IEEE; 2015, p. 1428–31. http://dx.doi.org/10.1109/ICRERA.2015.7418643.

[14] Yuan J, Yang HHL, Omitaomu OA, Bhaduri BL. Large-scale solar panel mapping from aerial images using deep convolutional networks. In: Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. 2016, p. 2703–8. http://dx.doi.org/10.1109/BigData.2016.7840915.

[15] Golovko V, Bezobrazov S, Kroshchanka A, Sachenko A, Komar M, Karachka A. Convolutional neural network based solar photovoltaic panel detection in satellite photos. In: Proceedings of the 2017 IEEE 9th international conference on intelligent data acquisition and advanced computing systems: technology and applications, IDAACS 2017. 1, IEEE; 2017, p. 14–9. http://dx.doi.org/10.1109/IDAACS.2017.8094501.

[16] Camilo J, Wang R, Collins LM, Bradbury K, Malof JM. Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery. 2018, http://dx.

doi.org/10.1109/IDAACS.2017.8094501, URL http://arxiv.org/abs/1801.04018. 10.48550.

[17] 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 2018;2(12):2605–17. http://dx.doi.org/10.1016/j.joule.2018.11.021.

[18] Castello R, Roquette S, Esguerra M, Guerra A, Scartezzini JL. Deep learning in the built environment: Automatic detection of rooftop solar panels using Convolutional Neural Networks. J Phys Conf Ser 2019;1343(1). http://dx.doi.org/10.1088/1742-6596/1343/1/012034.

[19] Hou X, Wang B, Hu W, Yin L, Wu H. SolarNet: A deep learning framework to map solar power plants in China from satellite imagery. 2019, URL http://arxiv.org/abs/1912.03685.

[20] Zech M, Ranalli J. Predicting PV Areas in Aerial Images with Deep Learning. In: Conference record of the IEEE photovoltaic specialists conference. June, Institute of Electrical and Electronics Engineers Inc.; 2020, p. 0767–74. http://dx.doi.org/10.1109/PVSC45281.2020.9300636.

[21] Zhuang L, Zhang Z, Wang L. The automatic segmentation of residential solar panels based on satellite images: A cross learning driven U-net method. Appl Soft Comput 2020;92:106283. http://dx.doi.org/10.1016/j.asoc.2020.106283.

[22] Kruitwagen L, Story KT, Friedrich J, Byers L, Skillman S, Hepburn C. A global inventory of photovoltaic solar energy generating units. Nature 2021;598(7882):604–10. http://dx.doi.org/10.1038/s41586-021-03957-7.

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

[24] Jiang H, Yao L, Lu N, Qin J, Liu T, Liu Y, Zhou C. Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery. Earth Syst Sci Data 2021;13(11):5389–401. http://dx.doi.org/10.5194/essd-13-5389-2021.

[25] Costa MVCVd, Carvalho OLFd, Orlandi AG, Hirata I, Albuquerque AOd, Silva FVe, Guimarães RF, Gomes RAT, Júnior OAdC. Remote sensing for monitoring photovoltaic solar plants in Brazil using deep semantic segmentation. Energies 2021;14(10):1–15. http://dx.doi.org/10.3390/en14102960.

[26] Kleebauer M, Horst D, Reudenbach C. Semi-automatic generation of training samples for detecting renewable energy plants in high-resolution aerial images. Remote Sens 2021;13(23):1–11. http://dx.doi.org/10.3390/rs13234793.

[27] Mayer K, Rausch B, Arlt M-L, Gust G, Wang Z, Neumann D, Rajagopal R. 3D-PV-locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. Appl Energy 2022;310:118469. http://dx.doi.org/10.1016/j.apenergy.2021.118469.

[28] Ortiz A, Negandhi D, Mysorekar SR, Nagaraju SK, Kiesecker J, Robinson C, Bhatia P, Khurana A, Wang J, Oviedo F, Ferres JL. An artificial intelligence dataset for solar energy locations in India. Sci Data 2022;9(1). http://dx.doi.org/10.1038/s41597-022-01499-9.

[29] Wang J, Liu J, Li L. Detecting Photovoltaic Installations in Diverse Landscapes Using Open Multi-Source Remote Sensing Data. Remote Sens 2022;14(24). http://dx.doi.org/10.3390/rs14246296.

[30] Chen Z, Kang Y, Sun Z, Wu F, Zhang Q. Extraction of Photovoltaic Plants Using Machine Learning Methods: A Case Study of the Pilot Energy City of Golmud, China. Remote Sens 2022;14(11). http://dx.doi.org/10.3390/rs14112697.

[31] Vlaminck M, Heidbuchel R, Philips W, Luong H. Region-based CNN for anomaly detection in PV power plants using aerial imagery. Sensors 2022;22(3):1–18. http://dx.doi.org/10.3390/s22031244.

[32] Plakman V, Rosier J, van Vliet J. Solar park detection from publicly available satellite imagery. GIScience and Remote Sens 2022;59(1):461–80. http://dx.doi.org/10.1080/15481603.2022.2036056.

[33] Arnaudo E, Blanco G, Monti A, Bianco G, Monaco C, Pasquali P, Dominici F. A Comparative Evaluation of Deep Learning Techniques for Photovoltaic Panel Detection From Aerial Images. IEEE Access 2023;11(May):47579–94. http://dx.doi.org/10.1109/ACCESS.2023.3275435.

[34] Guo Z, Zhuang Z, Tan H, Liu Z, Li P, Lin Z, Shang WL, Zhang H, Yan J. Accurate and generalizable photovoltaic panel segmentation using deep learning for imbalanced datasets. Renew Energy 2023;219. http://dx.doi.org/10.1016/j.renene.2023.119471.

[35] Ravishankar R, AlMahmoud E, Habib A, de Weck OL. Capacity estimation of solar farms using deep learning on high-resolution satellite imagery. Remote Sens 2023;15(1). http://dx.doi.org/10.3390/rs15010210.

[36] Tan H, Guo Z, Zhang H, Chen Q, Lin Z, Chen Y, Yan J. Enhancing PV panel segmentation in remote sensing images with constraint refinement modules. Appl Energy 2023;350. http://dx.doi.org/10.1016/j.apenergy.2023.121757.

[37] Kleebauer M, Marz C, Reudenbach C, Braun M. Multi-resolution segmentation of solar photovoltaic systems using deep learning. Remote Sens 2023;15(24):1–21. http://dx.doi.org/10.3390/rs15245687.

[38] Jianxun W, Xin C, Weicheng J, Li H, Junyi L, Haigang S. PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. Int J Appl Earth Obs Geoinf 2023;119(March):103309. http://dx.doi.org/10.1016/j.jag.2023.103309.

[39] Wang J, Chen X, Shi W, Jiang W, Zhang X, Hua L, Liu J, Sui H. Rooftop PV Segmenter: A Size-Aware Network for Segmenting Rooftop Photovoltaic Systems from High-Resolution Imagery. Remote Sens 2023;15(21). http://dx.doi.org/10.3390/rs15215232.

[40] Salama A, Hendawi A, Ali M, Al-Masri E, Franklin R, Deshpande A. SolarDetector: A transformer-based neural network for the detection and masking of solar panels. GIS: Proc ACM Int Symp Adv Geogr Inf Syst 2023. http://dx.doi.org/10.1145/3589132.3625649.

[41] Yang R, He G, Yin R, Wang G, Zhang Z, Long T, Peng Y, Wang J. A novel weakly-supervised method based on the segment anything model for seamless transition from classification to segmentation: A case study in segmenting latent photovoltaic locations. Int J Appl Earth Obs Geoinf 2024;130(February):103929. http://dx.doi.org/10.1016/j.jag.2024.103929.

[42] Olweus E, Mengshoel OJ. Detecting and Segmenting Solar Farms in Satellite Imagery: A Study of Deep Neural Network Architectures. In: 14th Scandinavian Conference on Artificial Intelligence SCAI 2024, June 10-11, 2024. 208, Jönköping, Sweden; 2024, p. 19–28. http://dx.doi.org/10.3384/ecp208003.

[43] Zhao Z, Chen Y, Li K, Ji W, Sun H. Extracting Photovoltaic Panels From Heterogeneous Remote Sensing Images With Spatial and Spectral Differences. IEEE J Sel Top Appl Earth Obs Remote Sens 2024;17:5553–64. http://dx.doi.org/10.1109/JSTARS.2024.3369660.

[44] Tan H, Guo Z, Lin Z, Chen Y, Huang D, Yuan W, Zhang H, Yan J. General generative AI-based image augmentation method for robust rooftop PV segmentation. Appl Energy 2024;368(February):123554. http://dx.doi.org/10.1016/j.apenergy.2024.123554.

[45] Ranalli J, Zech M, Tetens H-P. Deep Learning Models for PV Identification are Difficult to Generalize. In: accepted. 2024,

[46] García G, Aparcedo A, Nayak GK, Ahmed T, Shah M, Li M. Generalized deep learning model for photovoltaic module segmentation from satellite and aerial imagery. Sol Energy 2024;274(March):112539. http://dx.doi.org/10.1016/j.solener.2024.112539.

[47] Lodhi MK, Tan Y, Wang X, Masum SM, Nouman KM, Ullah N. Harnessing rooftop solar photovoltaic potential in Islamabad, Pakistan: A remote sensing and deep learning approach. Energy 2024;304(37):132256. http://dx.doi.org/10.1016/j.energy.2024.132256.

[48] Guo Z, Lu J, Chen Q, Liu Z, Song C, Tan H, Zhang H, Yan J. TransPV: Refining photovoltaic panel detection accuracy through a vision transformer-based deep learning model. Appl Energy 2024;355(November 2023):122282. http://dx.doi.org/10.1016/j.apenergy.2023.122282.

[49] Ksira Z, Blasuttigh N, Mellit A, Pavan AM. TinyML model for fault classification of photovoltaic modules based on visible images. In: International Conference on Artificial Intelligence in Renewable Energetic Systems. 2023, p. 373–80.

[50] Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J Photogramm Remote Sens 2019;152(November 2018):166–77. http://dx.doi.org/10.1016/j.isprsjprs.2019.04.015.

[51] Li P, Zhang H, Guo Z, Lyu S, Chen J, Li W, Song X, Shibasaki R, Yan J. Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning. Adv Appl Energy 2021;4:100057. http://dx.doi.org/10.1016/j.adapen.2021.100057.

[52] Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, Hoersch B, Isola C, Laberinti P, Martimort P, Meygret A, Spoto F, Sy O, Marchese F, Bargellini P. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ 2012;120:25–36. http://dx.doi.org/10.1016/j.rse.2011.11.026.

[53] Roy DP, Wulder MA, Loveland TR, C.E. W, Allen RG, Anderson MC, Helder D, Irons JR, Johnson DM, Kennedy R, Scambos TA, Schaaf CB, Schott JR, Sheng Y, Vermote EF, Belward AS, Bindschadler R, Cohen WB, Gao F, Hipple JD, Hostert P, Huntington J, Justice CO, Kilic A, Kovalskyy V, Lee ZP, Lymburner L, Masek JG, McCorkel J, Shuai Y, Trezza R, Vogelmann J, Wynne RH, Zhu Z. Landsat-8: Science and product vision for terrestrial global change research. Remote Sens Environ 2014;145:154–72. http://dx.doi.org/10.1016/j.rse.2014.02.001.

[54] Wang R, Camilo J, Collins LM, Bradbury K, Malof JM. The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: An empirical study with solar array detection. Proc - Appl Imag Pattern Recognit Workshop 2018;2017-Octob:1–8. http://dx.doi.org/10.1109/AIPR.2017.8457965.

[55] Tuia D, Volpi M, Copa L, Kanevski M, Muñoz-Marí J. A survey of active learning algorithms for supervised remote sensing image classification. IEEE J Sel Top Sign Proces 2011;5(3):606–17. http://dx.doi.org/10.1109/JSTSP.2011.2139193.

[56] Melgani F, Bruzzone L. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans Geosci Remote Sens 2004;42(8):1778–90. http://dx.doi.org/10.1109/TGRS.2004.831865.

[57] Crawford MM, Tuia D, Yang HL. Active Learning: Any Value for Classification of Remotely Sensed Data? In: Proceedings of the IEEE. 2013, p. 1–31.

[58] Mitra P, Uma Shankar B, Pal SK. Segmentation of multispectral remote sensing images using active support vector machines. Pattern Recognit Lett 2004;25(9):1067–74. http://dx.doi.org/10.1016/j.patrec.2004.03.004.

[59] Lenczner G, Chan-Hon-Tong A, Le Saux B, Luminari N, Le Besnerais G. DIAL: Deep interactive and active learning for semantic segmentation in remote sensing. IEEE J Sel Top Appl Earth Obs Remote Sens 2022;15:3376–89. http://dx.doi.org/10.1109/JSTARS.2022.3166551.

[60] Ren P, Xiao Y, Chang X, Huang PY, Li Z, Gupta BB, Chen X, Wang X. A Survey of Deep Active Learning. ACM Comput Surv 2022;54(9). http://dx.doi.org/10.1145/3472291.

[61] Wu M, Li C, Yao Z. Deep Active Learning for Computer Vision Tasks: Methodologies, Applications, and Challenges. Appl Sci (Switzerland) 2022;12(16). http://dx.doi.org/10.3390/app12168103.

[62] Zhan X, Wang Q, Huang K-h, Xiong H, Dou D, Chan AB. A comparative survey of deep active learning, preprint arxiv 2203.13450. 2022, URL http://arxiv.org/abs/2203.13450.

[63] Roy N, McCallum A. Toward Optimal Active Learning through Montle Carlo Estimation of Error Reduction. willamstown: ICML; 2001.

[64] Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J 1948;27(4):623–56. http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x.

[65] Settles B. Active Learning Literature Survey. University of Wisconsin–Madison; 2009.

[66] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: 34th International Conference on Machine Learning, ICML 2017. 3, 2017, p. 2130–43.

[67] Mittal S, Niemeijer J, Schäfer JP, Brox T. Best Practices in Active Learning for Semantic Segmentation. In: German Conference on Pattern Recognition (GCPR). 2023, URL http://arxiv.org/abs/2302.04075.

[68] Houlsby N, Huszár F, Ghahramani Z, Lengyel M. Bayesian active learning for classification and preference learning. 2011, arXiv preprint arXiv:1112. URL http://arxiv.org/abs/1112.5745.

[69] Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data. In: International conference on machine learning. 2017, p. 1183–92.

[70] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. 2016.

[71] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach. In: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. 2018, URL http://arxiv.org/abs/1708.00489.

[72] Kasmi G, Saint-Drenan YM, Trebosc D, Jolivet R, Leloux J, Sarr B, Dubus L. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. Sci Data 2023;10(1):1–12. http://dx.doi.org/10.1038/s41597-023-01951-4.

[73] Bradbury K, Saboo R, Johnson TL, Malof JM, Devarajan A, Zhang W, Collins LM, Newell RG. Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. Sci Data 2016;3(December):1–9. http://dx.doi.org/10.1038/sdata.2016.106.

[74] Ronneberger, Fischer, Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing; 2015, p. 234–41.

[75] Deng Z, Sun H, Zhou S, Zhao J, Lei L, Zou H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J Photogramm Remote Sens 2018;145(June):3–22. http://dx.doi.org/10.1016/j.isprsjprs.2018.04.003.

[76] Salberg A-Br. Detection of seals in remote sensing images using features extracted from deep convolutional neural networks. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2015. (0373):2015, p. 1893–6.

[77] Hernandez-Sequeira I, Fernandez-Beltran R, Pla F. Transfer Deep Learning for Remote Sensing Datasets: A Comparison Study. In: International Geoscience and Remote Sensing Symposium (IGARSS). 2022-July, Institute of Electrical and Electronics Engineers Inc.; 2022, p. 3207–10. http://dx.doi.org/10.1109/IGARSS46834.2022.9884667.

[78] Iakubovskii P. Segmentation models pytorch. 2019, GitHub repository.

[79] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press 2017;521(7553):785. http://dx.doi.org/10.1016/B978-0-12-391420-0.09987-X.

[80] Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 2012;28(19):2520–2. http://dx.doi.org/10.1093/bioinformatics/bts480.

[81] Pop R, Fulop P. Deep ensemble Bayesian active learning : Addressing the mode collapse issue in Monte Carlo dropout via ensembles. 2018, p. 1–15, URL http://arxiv.org/abs/1811.03897.

[82] Li Z, Hoiem D. Learning without Forgetting. IEEE Trans Pattern Anal Mach Intell 2018;40(12):2935–47. http://dx.doi.org/10.1109/TPAMI.2017.2773081, URL http://arxiv.org/abs/1606.09282.

[83] Cheng G, Xie X, Han J, Guo L, Xia GS. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. IEEE J Sel Top Appl Earth Obs Remote Sens 2020;13:3735–56. http://dx.doi.org/10.1109/JSTARS.2020.3005403.

[84] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Adv Neural Inf Process Syst 2021;15:12077–90.

[85] Wang Z, Arlt ML, Zanocco C, Majumdar A, Rajagopal R. DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models. Joule 2022;6(11):2611–25. http://dx.doi.org/10.1016/j.joule.2022.09.011.

[86] Kasmi G, Dubus L, Blanc P, Saint-Drenan YM. Towards Unsupervised Assessment with Open-Source Data of the Accuracy of Deep Learning-Based Distributed PV Mapping. CEUR Workshop Proceedings 2022;3343.