Generalizability of Neural Network-based Identification of PV in Aerial Images

Joseph Ranalli[1] and Matthias Zech[2]

[1]Penn State Hazleton, Hazleton, PA, USA
[2]German Aerospace Center (DLR), Institute of Networked Energy Systems, Oldenburg, Germany

Abstract — **Identification of PV panels from aerial imagery is a potential strategy for building comprehensive behind-the-meter PV datasets. Several previous studies have utilized Convolutional Neural Networks with the goal of producing tools that can perform these identification tasks. Neural Network approaches rely on labelled data for training, with several aerial imagery datasets with labelled PV already available. This study aims to investigate generalizability of models trained on one set of labelled PV data to other datasets, to further understanding of how these models can be applied. Six different PV datasets were utilized, and test data results were compared. Overall, we find that generalizability suffers when models are presented with different data than they were trained on. We describe some dataset features that led to particularly poor generalization. This study highlights the need for further research to investigate strategies for improving generalizability of trained Neural Network models.**

I. INTRODUCTION

A multitude of academic and industrial contexts require knowledge of the location and configuration of distributed photovoltaic (PV) installations. Examples include evaluating renewable energy policy by tracking regional PV growth rates and providing technical data to support aggregate-level forecasts of distributed PV generation, necessary to simulate and operate climate-neutral energy systems. Generally, no large scale registry of distributed PV systems exists. Detailed data about behind-the-meter PV may be particularly difficult to obtain. Even in cases where data on PV installations exists, it may not be accessible to researchers or the general public. Instead, availability of data on these PV systems varies widely by jurisdiction and is subject to the policies of local stakeholders who directly collect and use this information.

Recent efforts to improve data availability for distributed PV have attempted to identify systems visually from satellite or aerial imagery. Progress has been made utilizing Deep Learning on these imagery sources, for instance for the United States [1], regions of Germany [2], [3], China [4] and even worldwide [5]. Global studies mostly rely on publicly available image sources with wide coverage, but low-detail spatial resolution. This typically limits detection to utility-scale PV systems with large capacity that are visible from satellites on rough spatial resolutions.

A suitable global PV inventory also needs to contain residential PV sites that are small, but numerous, making up more than 40\% of the global PV capacity [6]. The rough spatial resolution of satellite imagery may limit its usefulness towards this task. Instead, aerial imagery may be a more reliable source for high resolution PV detection. Aerial imagery is becoming increasingly open-access, but due to its origin from airborne flights, is usually only available to a limited spatial extent. Furthermore, its localized nature may lead to greater diversity in the imagery, both in terms of acquisition equipment and processing, and resulting from local visual diversity in the area being imaged. To derive a global PV registry which includes small-scale and large-scale PV systems, it is therefore likely that PV detection systems need the ability to deal with a range of different image sources with variable spatial resolution and availability.

Regarding neural network models, many academic studies are based upon performance using individual data sources. But the need for a flexibility in generating a global PV registry raises the question how well models trained on single-site data generalize to other datasets, particularly for higher resolution aerial imagery datasets. Studies investigating the generalizability [7], [8] have already shown that neural networks generalize poorly when trained and tested on different cities, even when the images originate from the same data source. Different local characteristics, as geography and population density [8] may impact generalizability. A study testing the generalizability more broadly across different local sites and data sources is still lacking.

This study aims to fill this gap by systematically testing how well models trained on single datasets perform on other sites, using six different labeled aerial image datasets, including Northern Germany, Southern France, the United States (sets from California and New York City). These six datasets were labeled using different methods, by different research groups and have different original data sources. This provides a realistic testbed for verification metrics when a model is applied outside its trained dataset without any modifications.

II. METHODOLOGY

A. Model Architecture

This study used a Fully Convolutional Neural Network architecture that had previously been applied for the purpose of identifying PV arrays in aerial images [3]. Specifically, the network architecture utilized was a u-net [9], which was designed for semantic segmentation tasks. A u-net architecture uses symmetric encoder and decoder paths, producing a structure that resembles the letter *u*. The implementation of u-net used in this study was based on the python library *segmentation_models* [10], which contained a version of u-net built using Tensorflow and Keras [11], along with pretrained weights from the *ImageNet* competition [12]. A backbone of ResNet-34 was used for the u-net model, as it was previously shown to produce similar quality results to more detailed backbones [3] for this task.

As the primary outcome of this study is the application of neural networks, rather than the neural network itself, a thorough optimization of the training process was not conducted. Rather, settings that produced reasonable and consistent results were used. Training was performed on a desktop computer with a single, consumer-grade GPU, and completed in a reasonable timeframe (2-3 hours per training dataset). During training, the encoder weights were frozen to their pre-trained values to reduce the number of weights being adjusted. Early stopping was employed to prevent overfitting, stopping the training after 10 consecutive epochs without a reduction in the loss. Data augmentation was used to simulate features of a larger dataset. The parameters used for augmentation were: rotation (up to 30°), zoom (factor of 0.2) and height and width shifts (factor of 0.1 each). The weights with the best value of validation loss observed during training were retained for evaluation.

B. Source Data

Data was utilized from six separate publicly available datasets consisting of labeled PV installations in aerial/satellite images: two cities in California, Fresno and Stockton (CA-F and CA-S) from [13], two datasets in France (FR-G and FR-I) from [14] using images from Google Earth and the French government (IGN) respectively, a dataset of Google Earth images from Germany (DE-G) used in a previous study by the authors [3] and a newly labelled dataset based on publicly available images for Queens, New York (NY-Q) [15].

In a few of the original datasets (CA-F, and CA-S and NY-Q) raw images were too large (e.g. 5000x5000 pixels) to be processed by the model as whole images. Thus, these large images were split into tiles of smaller size for the training and testing. Tiles that did not contain any labelled pixels were excluded from each dataset. The overall number of retained tiles and other details of the datasets are provided in Table I.

Due to the variable number of tiles in each dataset, a subset of 1000 tiles was randomly chosen from each to form the representative subset for that data source. This ensured that all models were trained on an equal quantity of data. Each set of 1000 images was randomly split into sets of 720 training, 80

validation and 200 test tiles. These split sizes were consistent across all datasets, and did not include any crossover. All images were resized to 576x576 pixels during processing to match the network architecture. Consequently, smaller images (e.g. FR-G and FR-I datasets) had an effective increase in scale of about 1.4, while the datasets with larger tiles were scaled down by a factor of around 0.9. The effective resolutions of each dataset after scaling are also given in Table I.

In addition to the 6 original datasets, a synthetic combination dataset was created by combining 133 of the processed tiles from each dataset (with two extras chosen from NY-Q to reach the correct total number of 1000) to compare training on the aggregate. After training, each trained model was evaluated on the test data for all the datasets and values for several performance metrics were recorded.

While fundamentally subjective in nature, an effort was made to describe contextual differences between the datasets. Five separate categories were created to describe the various types of characteristics found within the datasets: large structures and/or flat roofs (often commercial buildings), large open spaces (roughly more than 50\% of the image), agricultural (identified by visible patterns or rows), large bodies of water and utility scale PV. Images for each category from each dataset's 1000 tile subset were manually counted and rounded to the nearest ten to reflect the subjective nature of these judgments. A table summarizing these categories is in Table II. The remainder of images in each dataset primarily consisted of residential housing, for which the approximate number of structures per image is notated as well.

TABLE I
DATASETS

Dataset	Tot. Tiles	Tile Size	Resolution	Scaled Res.	Ref
CA-F	1,044	625x625	0.3 m/px	0.32 m/px	[13]
CA-S	4,192	625x625	0.3 m/px	0.32 m/px	[13]
FR-G	13,303	400x400	0.1 m/px	0.07 m/px	[14]
FR-I	7,865	400x400	0.2 m/px	0.14 m/px	[14]
DE-G	1,325	639x640	0.18 m/px	0.2 m/px	[3]
NY-Q[1]	1,007	625x625	0.15 m/px	0.16 m/px	[15]

[1] labeling of this dataset is still ongoing

TABLE II
CONTEXTUAL DIFFERENCES BY DATASET (APPROXIMATE)

Dataset	Large/Flat	Open Space	Agri.	Util. PV	# Bldg/Tile
CA-F	70	140	40	0	20-40
CA-S	70	80	10	0	20-40
FR-G	10	20	0	0	2-5
FR-I	20	90	20	0	5-10
DE-G	60	80	10	10	10-20
NY-Q	130	10	0	0	10-20

Despite the qualitative nature of these categorizations, a few generalizations can be made. All of the datasets primarily consisted of detached or semi-detached residential dwellings. The NY-Q dataset contained the lowest number of images that featured large open spaces and the highest number of large and flat-roofed structures (often commercial or industrial buildings in this case), which corresponds to its more urban character. Qualitatively, NY-Q also had less visible vegetation intermingled in the residential areas (i.e. trees) than the other datasets. CA-F had the highest frequency of images with large open areas, and many of these areas showed vegetation in rows/patterns that are presumed to indicate agricultural activity (and impacted the detection of PV as will be seen). The number of residential buildings per tile roughly corresponded to the scaled resolution of the images, with the two California datasets having the lowest zoom level (covering the widest area), FR-G having the highest zoom level, and FR-I, DE-G and NY-Q in the middle. It is also notable that the FR-G and FR-I datasets were both designed to have tiles centered on a positive PV system [14]. In part this explains some of the counts seen in Table II for FR-G; due to the high zoom level and smaller overall tile size, centering the tile on the PV array may have decreased the probability of observing some of the features present in some of the more randomly aligned datasets.

III. Results and Discussion

As stated, each of the seven trained models was evaluated using 200 test images associated with each of the six datasets. Precision and recall are especially useful metrics in interpreting the performance of these models. Precision can be interpreted as the percentage of predicted positive pixels that are correct, defined as in Eq. 1. Recall can be interpreted as the percentage of the labelled positive pixels that are identified by the model, defined as in Eq. 2. The Jaccard index, also called the intersection over union, or IoU, is defined in Eq. 3. IoU represents the match between the the actual and predicted PV regions in the image and is useful in representing the overall performance of the model. Results of precision and recall for the models are presented in Table III and Table IV respectively. Results for the IoU are given in Table V.

$$p = \frac{True\ Positives}{Predicted\ Positives} = \frac{TP}{TP+FP} \qquad (1)$$

$$r = \frac{True\ Positives}{Labeled\ Positives} = \frac{TP}{TP+FN} \qquad (2)$$

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} = \frac{TP}{TP+FP+FN} \qquad (3)$$

In conjunction, these metrics help us understand the model's performance. High values in both precision and recall indicate that the model identifies most ground truth pixels and does so correctly, while low values in both indicates that the model makes mostly incorrect positive identifications and fails to

identify many ground truth positives. The crossed value cases on the two metrics provide useful information as well. A high precision with low recall indicates that the model is usually correct when it predicts positives, but that it fails to identify many of the ground truth positives (i.e. is very selective about predictions). For example, the FR-I trained model is correct 95\% of the time when predicting pixels to be PV in the FR-G test dataset (precision), but identifies only 36\% of the overall PV pixels (recall). Conversely, a low precision with a high recall indicates that a model identifies most of the positive pixels in the ground truth, but also incorrectly identifies many pixels as PV. An example is the CA-F trained model, which predicts 59\% of the PV pixels from the DE-G ground truth dataset (recall), but is only correct about its positive PV predictions 7\% of the time (precision).

A. Individual Model-to-Test Case Performance

Generally, models performed better on their own test data than on other test datasets, which is unsurprising, but does highlight the challenges in generalization of models that has been described by the literature. Additionally, each model performed best of the group on its corresponding test data, with a few minor exceptions in the precision metric. The combination dataset had relatively good performance for all test datasets. This is unsurprising as its more diverse training set seems to have helped it to produce more generalizable results. Figures 1 - 6 show results on one of the better performance images for each test dataset. The images in these figures are all examples of cases where all models performed reasonably well, and were selected from the top 5\% of average IoU scores for each test set.

The average IoU performance of models on their own test data was a value of 0.71. As measured by the IoU score, most generalization from trained models to different test sets was quite poor, though a few examples of moderate performance were observed. The best two examples of generalization when judging by IoU were FR-I model's performance on NY-Q data and CA-S predicting data from CA-F. In part, the connection between the two California cases may result from their shared data source and overall locale. However, the lack of the reciprocal boost to performance of CA-F on the CA-S test data suggests that other factors may be at play.

TABLE III
PRECISION RESULTS BY DATASET

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
CA-F	0.87	0.46	0.36	0.48	0.07	0.25
CA-S	0.82	0.79	0.51	0.31	0.22	0.24
FR-G	0.10	0.03	0.91	0.76	0.41	0.52
FR-I	0.63	0.64	0.95	0.79	0.67	0.77
DE-G	0.70	0.65	0.83	0.91	0.77	0.82
NY-Q	0.59	0.66	0.90	0.87	0.75	0.90

Train datasets in rows, test datasets in columns

TABLE IV
RECALL RESULTS BY DATASET

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
CA-F	0.79	0.59	0.15	0.58	0.59	0.35
CA-S	0.62	0.72	0.13	0.47	0.59	0.37
FR-G	0.06	0.01	0.88	0.52	0.15	0.29
FR-I	0.15	0.23	0.36	0.84	0.37	0.67
DE-G	0.19	0.33	0.11	0.30	0.79	0.48
NY-Q	0.07	0.24	0.15	0.50	0.47	0.89

Train datasets in rows, test datasets in columns

TABLE V
IOU RESULTS BY DATASET

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
CA-F	0.71	0.35	0.11	0.36	0.06	0.16
CA-S	0.55	0.61	0.11	0.22	0.17	0.19
FR-G	0.03	0.00	0.81	0.45	0.13	0.26
FR-I	0.13	0.19	0.35	0.69	0.31	0.56
DE-G	0.18	0.29	0.11	0.29	0.63	0.44
NY-Q	0.07	0.22	0.15	0.47	0.40	0.81

Train datasets in rows, test datasets in columns

Investigating more deeply, the individual images with the best- and worst- IoU score performance were visualized for each of these combinations. In the case of FR-I predicting on NY-Q, the best performing images were dominated by large commercial rooftop systems (one shown in Fig. 6). The worst performing images were predominantly of residential housing, but no particular pattern was able to be discerned from qualitative inspection. In the case of CA-S predicting CA-F test data, the best performance was observed on large scale systems, as well as images with PV closely clustered, as opposed to scattered across the image. Several high performance images include instances of apparent agriculture, for which some models exhibited significant confusion (as shown in Fig. 7). No clear pattern was visible in the poorly performing images from CA-S to CA-F. Several ground mounted systems were missed, and many of the poor performing images did contain very small segments of PV. Though none of the models were a great success, the FR-I model was most generalizable overall, having the highest average IoU score across all other test sets (IoU = 0.31).

B. Overall Generalization

The worst individual example of generalization was the FR-G model, which had exceptionally poor performance on both of the CA test sets. Even FR-G's best performance on images within CA-S and CA-F showed essentially no skill at prediction. Since FR-G had the highest scaled detail level (0.07

m/pix) by a factor of two, this may suggest that the models are not adapting well to imagery at lower zoom levels. While FR-G generalized poorly overall, it showed its best test performance on the FR-I dataset, which in addition to having the closest scaled resolution, was also produced by the same labelers and for the same geographic region.

The average generalization performance on each metric was computed by excluding the results for each model's own test data and that for CMB-6. The highest IoU scores averaged across all models occurred for the FR-I test data (IoU = 0.36) and the NY-Q test data (IoU = 0.32), indicating that these test data were easiest for a general model to predict. It is important to note though that these numbers are well below the performance achieved by the specially-trained models. The easiest images for all models to predict in the FR-I test data were primarily relatively large systems, with very clear panel frames. The most predictable images in NY-Q were all large commercial-scale systems, also with clearly visible module frames.

The FR-G test data was the hardest for all other models to predict (average IoU = 0.17). However, most models had relatively good precision on this model, indicating that models were reluctant to predict positive values in FR-G, but were often correct when they did. One noticeable feature in the FR-G predictions is that most models tended to discretize the panels, rather than predicting overall areas of panels. This effect can be seen in Fig. 3. This may be a consequence of the higher zoom levels making the module frames appear to be gaps. FR-G's predictions on its own data did not exhibit this feature. This also points to how differences in labeling of module frames, and for larger systems, gaps between rows may impact the quality of predictions, especially when working with data sources at varying zoom levels. Despite the usage of different zooming levels in data augmentation, the models seem to be highly sensitive to the spatial resolution of the image source.

C. Combination Model Performance

As stated, the combination model (CMB-6) was trained on a random selection of 133 images from each other dataset. Besides representing a more generalized training source, this also provides the combination with a small sample of data from each test set it will be compared to. As seen in Tables III - V, this led to the combination model having relatively good precision across all test sets, meaning that it tended to avoid false positives. Except for the case of CA-F which showed exceptional generalization from CA-S, CMB-6 also had the best IoU score for each test dataset aside from specifically trained model for that dataset.

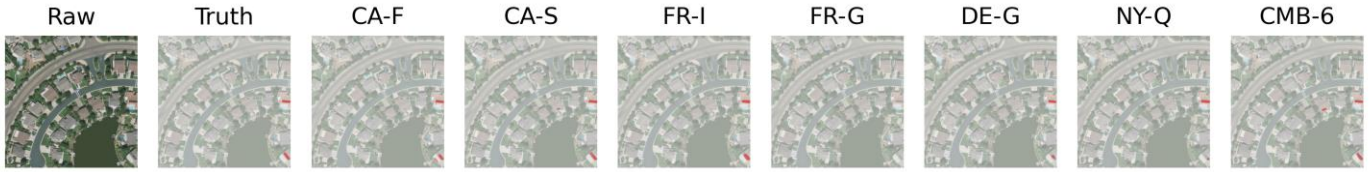Fig. 1. Tile from CA-F showing prediction by each model.

Fig. 2. Tile from CA-S showing prediction by each model.

Fig. 3. Tile from FR-G showing prediction by each model.

Fig. 4. Tile from FR-I showing prediction by each model.

Fig. 5. Tile from DE-G showing prediction by each model.

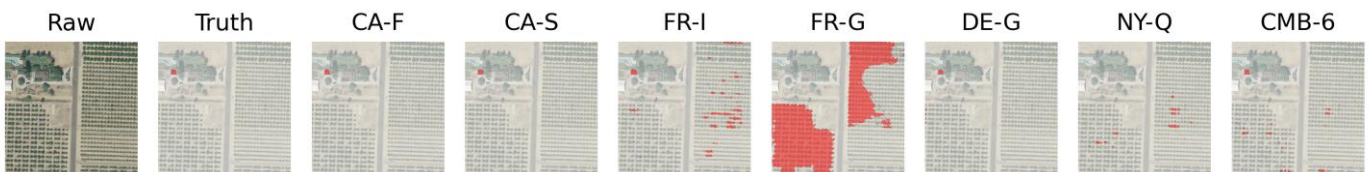Fig. 6. Tile from NY-Q showing prediction by each model.

Fig. 7. Tile from CA-F showing incorrect identification of agricultural rows by some models.

Generally speaking, the images for which CMB-6 had its worst performance were comparable to those already described for other datasets. This is also true for the case of CA-F, the test data that was most challenging for CMB-6. As mentioned, the precision remained relatively high for CA-F indicating that inclusion of more general training data improved its selectivity in making predictions for that data. CMB-6 experienced the largest benefit on FR-G when compared to predictions of other datasets. Including a small quantity of the hardest-to-predict data improved the model substantially for that case. Looking at image-wise performance, CMB-6 avoided the discretized module predictions that other models tended to experience for FR-G. Its worst performing images were characterized by a high number of false positives.

III. CONCLUSION

Identification of solar PV from aerial or satellite imagery has potential to improve the quality of data on distributed solar installations and to broaden access to such data. A growing number of labeled datasets are available to support training of neural network models for this task. However, generalization of the models across locales and image sources is a necessary technical hurdle to clear to permit these datasets to find broad application.

This study found that generalization of models trained on a single dataset is relatively challenging when applied to other datasets. This result is consistent with conclusions made by other investigators [7], [8]. We observed some limited cases of strong generalization between models, but without an ability to draw strong conclusions on the basis of qualitative interpretation of the datasets alone, however some inferences were possible. The worst performing model was trained on a dataset that had a significant difference in zoom level as compared to others, which is likely to have had an impact on its performance. Overall, commercial rooftops and large scale systems tended to be best predicted across all models. A model trained on a combination of data from each of the test sets showed adequate (but not exceptional) performance across all sets of test data. Where most models showed modest precision, albeit with low recall, the combination model struck a middle ground to improve performance: raising the recall at the cost of some precision.

We have described some of the strengths and weaknesses of generalization across six separate aerial imagery datasets applied for identification of PV. However, no hard conclusions were able to be reached on the basis of this data alone. Further work is needed to test methodologies that may improve the generalizability of the trained models and address the differences found within the source data. These steps may lead to the ultimate goal of producing application-ready tools for computer-based aerial identification of PV.

REFERENCES

[1] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States," *Joule*, vol. 2, no. 12, pp. 2605–2617, Dec. 2018, doi: 10.1016/j.joule.2018.11.021.

[2] K. Mayer *et al.*, "3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D," *Appl. Energy*, vol. 310, p. 118469, Mar. 2022, doi: 10.1016/j.apenergy.2021.118469.

[3] M. Zech and J. Ranalli, "Predicting PV Areas in Aerial Images with Deep Learning," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, Jun. 2020, pp. 0767–0774. doi: 10.1109/PVSC45281.2020.9300636.

[4] X. Hou, B. Wang, W. Hu, L. Yin, A. Huang, and H. Wu, "SolarNet: A Deep Learning Framework to Map Solar Plants In China From Satellite Imagery," in *Climate Change AI*, Climate Change AI, Apr. 2020. Accessed: Dec. 19, 2022. [Online]. Available: https://www.climatechange.ai/papers/iclr2020/6

[5] L. Kruitwagen, K. T. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, Art. no. 7882, Oct. 2021, doi: 10.1038/s41586-021-03957-7.

[6] S. Joshi, S. Mittal, P. Holloway, P. R. Shukla, B. Ó Gallachóir, and J. Glynn, "High resolution global spatiotemporal assessment of rooftop solar photovoltaics potential for renewable electricity generation," *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, Oct. 2021, doi: 10.1038/s41467-021-25720-2.

[7] R. Wang, J. Camilo, L. M. Collins, K. Bradbury, and J. M. Malof, "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)*, Oct. 2017, pp. 1–8. doi: 10.1109/AIPR.2017.8457965.

[8] W. Hu *et al.*, "What you get is not always what you see—pitfalls in solar array assessment using overhead imagery," *Appl. Energy*, vol. 327, p. 120143, Dec. 2022, doi: 10.1016/j.apenergy.2022.120143.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[10] P. Yakubovskiy, *Segmentation Models*. Github, 2019. [Online]. Available: https://github.com/qubvel/segmentation_models

[11] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Google, 2015. [Online]. Available: https://tensorflow.org

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[13] K. Bradbury *et al.*, "Distributed Solar Photovoltaic Array Location and Extent Data Set for Remote Sensing Object Identification." figshare, Oct. 02, 2018. doi: https://dx.doi.org/10.6084/m9.figshare.3385780.

[14] G. Kasmi *et al.*, "A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata." Zenodo, Jul. 20, 2022. doi: https://doi.org/10.5281/zenodo.7059985.

[15] "NYS Interactive Mapping Gateway." https://gis.ny.gov/gateway/mg/ (accessed Nov. 22, 2022).