Diversifying Training Data Does Not Improve Generalizability of Neural Network Models for PV Identification

Joseph Ranalli[1, a)] and Matthias Zech and Hendrik-Pieter Tetens[2]

[1)] *Penn State Hazleton, Hazleton, Pennsylvania, United States, 18202*

[2)] *German Aerospace Center (DLR) Institute of Networked Energy Systems, Oldenburg, Germany, 26129*

(Dated: 2 December 2024)

Data about behind-the-meter photovoltaic (PV) installations may be difficult to obtain for researchers. A number of investigators have considered Deep Learning as an attractive solution to this challenge, capable of directly identifying PV installations from aerial or satellite images. Deep Learning models are well known to experience challenges when working with data from sources that they have never been exposed to. This study investigated whether generalizability can be improved by diversifying training data across available labeled data sources. We assessed performance of models trained on a all possible combinations of six different labeled data sets of aerial PV imagery, with a fixed number of total training images. Unfortunately, our results indicate that no combination of model training data achieved generalized performance that approaches models trained on data from a target data source. This implies that generalized ResNet models cannot be developed simply by modifying the configuration of the training data. Consequently, researchers should expect that some degree of data labeling is likely to be necessary when adapting these models to new applications, but our results do indicate that significant performance improvements are possible with only small (~20%) introductions of target data. Future work may investigate alternative architectures, expanded training data sets, or ways to reduce the amount of labeled data necessary to adapt a model for a given application.

I. INTRODUCTION

A global transition away from carbon-intensive energy sources is underway, requiring growth of renewable, carbon-free forms of energy generation to meet societal energy needs. In particular, solar photovoltaics (PV) represent a renewable energy source that directly converts freely available solar irradiance into electricity. PV generation is growing quickly due to its comparatively low levelized cost of electricity, leading solar to represent a significant share of optimal energy scenarios for the United States[1].

The distributed placement of generation facilities is a common feature of most renewable energy systems, arising from the distributed nature of renewable energy resources. PV generation facilities vary widely in scale, ranging from large "utility-scale" generation facilities consisting of thousands of panels down to small scale systems on residential rooftops that may be made up of only a few panels. The largest of utility scale systems are rated to produce on the order of hundreds of megawatts at peak capacity, while residential systems may be rated for a few kilowatts. While this represents a scale difference as large as five orders of magnitude, residential systems are not negligible. To the contrary, small but numerous residential PV systems have been reported to represent more than 40% of global PV capacity[2].

In order for the energy transition to be planned for, studied and understood, access to information about PV installations is a strict necessity. For example, technical data about distributed PV installations is necessary

to forecast aggregate-level generation, needed to monitor and operate regional electricity grids. Generally speaking, no worldwide comprehensive inventory of PV installations exists, particularly when considering small-scale residential systems[3]. Where data exists, it is likely to vary significantly by jurisdiction in level of aggregation, comprehensiveness and availability. Nonetheless, there is at present a gap in accessible information about small-scale, behind-the-meter PV installations which remains of interest to researchers, policy-makers, utility operators and other stakeholders.

One potential solution to this problem lies in the use of computer vision systems operating on satellite or aerial imagery. These types of imagery offer visual indication of the presence of PV systems, and could potentially allow investigators to quantify details of PV installations[4]. Advances in computing technology have made implementation of Deep Learning-based computer vision techniques accessible for individual researchers, enabling this avenue of constructing a PV inventory. While studies have demonstrated this application for identifying PV installations[5], no ready-to-use system for general identification of PV installations exists, and a significant amount of ongoing research continues to advance the field. The most desirable system for the research community would be flexible and easy to access, allowing individual researchers to reliably process new image data sets as they become available so that data about PV installations can be used for further analyses. Thus, in order to serve as a tool for the research community, an ideal PV identification system would be available with low barrier to entry (e.g. open access, small and efficient, works on individual desktop computers) and could be generalized across unseen images. This study will examine the literature with

respect to this ideal and investigate the generalizability of common small scale models for identification of PV from image data.

II.   BACKGROUND

As stated previously, data on small-scale PV installations (e.g. location, capacity, characteristics) are needed for many research and operational activities, but are typically not openly available[3] or suffer from incongruities across data sources. Work by Yu et al.[5] leverages computer vision to identify PV installations and demonstrates the practical outcomes of using automated processes to build a database of PV installations. They apply the generated installation data for analysis of PV development measured against other geospatial and demographic variables. A followup study from the same group also considered similar analyses including time resolved effects[6]. Another example application comes from Perry and Campos[4], who demonstrate the ability of processing Google Earth imagery to verify metadata about PV installations.

A.   Architectures

Consequently, research into PV identification using computer vision on remote sensing imagery has been an area of interest from several investigators in recent years. Early efforts made use of generalized machine learning techniques[7], but most more recent efforts employ Deep Convolutional Neural Network (CNN) based approaches. For example, Yu et al. conducted an extensive study utilizing Deep Learning techniques[5]. Other groups have also demonstrated the use of neural networks to identify PV installations using data from Germany[8,9], China[10] and worldwide data[11,12].

Some studies have focused on advancement of different CNN architectures to improve performance at the task. Zhu et al. develop a network architecture specifically for identification of PV and perform transfer learning on a highly similar site with good success[13]. Guo et al. note the difficulty of class imbalance as a common issue for PV segmentation tasks, whereby more negative (i.e. background) than positive pixels are typically present in data sets[14]. They develop an architecture with characteristics that show favorable performance on a single data set when dealing with variation in resolution and accommodating class imbalance[14]. Zech and Ranalli[9] utilized a method that estimates uncertainty of the predictions as part of the characterization of PV identification task.

B.   Generalizability

Challenges remain in developing these techniques for more generalized applications. To make the most use of PV identification system, users need to be able to apply the system to inventory PV installations in new data as it becomes available. This is difficult, because studies have shown that CNN performance suffers when applied to data that is different from its training data[15]. Insufficient attention has been paid to the problem of generalization because for most research, models are often trained and evaluated on data from the same geographic location[15].

When it comes to generalization of CNN models, statistical similarity between training data and the target application location is important. Often this similarity is disturbed by geographic differences[3,15]. Differences in data may arise from the nature of the remote measurement (e.g. satellite vs. aerial imagery, use of orthographic rectification, different spatial resolution), differences in the sensor (e.g. type, sensitivity, calibration), differences in the site geography (e.g. urban vs. rural, construction similarities, types of features) or differences induced by temporal effects (e.g. images from summer vs. winter, varying atmospheric conditions, different shadowing)[3,16]. Data set quality is also a potential source of error due to unknown mistakes in the labeled "truth" data[3]. Openness of data, code and methods is important for research in terms of developing repeatable approaches that can also be practically applied[3]. Satellite data, which offers potentially the widest geographic coverage, may be limited by spatial resolution. Li et al. investigated varying pixel resolutions, observing that best performance occurred with resolutions finer than 0.3 m,[16] which does allow for some satellite data to be useful, but is close to current resolution limits.

A few studies have attempted to investigate the generalizability of CNN models. Wang et al. compare performance from two cities in California (Fresno and Stockton) from the same data source and observe poor generalization[15]. They determined that improved performance requires substantial quantities of local data, but less than the full training dataset. They were also able to use analysis of encoded data through a t-SNE algorithm to identify some of the limitations in performance[15]. Hu et al. conduct a comparison of predictions on data from the United States, comparing data from Connecticut and San Diego, California and observed significant difficulty in generalizing[3]. It is not yet known to what degree is is possible to deliberately create a generalized model based on existing training data.

C.   Contribution of this work

This study, an extension of a work presented at a recent conference[17], aims to fill a gap in research on generalization of CNN for PV identification tasks. Specifically, we investigate how well ResNet models trained on a single small data set can be applied to data from other locations and sources, and whether diversification of the training data can improve performance. We conducted a comprehensive evaluation of how results from a model trained

on a given data set generalize to other locations by incorporating data from six distinct aerial imagery data sets covering Northern Germany, Southern France, and the United States. These data sets were sourced from different labeling methodologies and acquisition modalities, making them indicative of the variety of imagery data that researchers may encounter when attempting to perform analyses on a new location. We investigate whether generalizability can be improved by incorporating more diverse training data from multiple locations to provide a realistic representation of how well models can be applied outside their initially trained context.

III. METHODOLOGY

A. Model architecture

Most modern image identification tasks make use of Fully Convolutional Neural Network architectures. U-net architectures, first introduced for the segmentation of biological imagery[18], have been applied for identification by several previous investigations[9,14]. In u-net architectures, the encoder and decoder have a symmetric configuration that resembles a *u*-shape. In this study, we utilized open-source python implementations of u-net from the *segmentation models* library[19], built upon Tensorflow and Keras[20]. All models in our study were trained beginning with pretrained weights from *ImageNet*[21], which were included with the library.

Multiple backbones were considered for the u-net model including ResNet-34, ResNet-50 and ResNet-101. Initial tests indicated that there was no significant difference in performance with increasing backbone complexity when comparing models trained and tested on the same data set. Results for this comparison are shown in Table I. This is consistent with prior results in a different study by the authors[9] for this task. Consequently, subsequent results reported here will come from the ResNet-34 based models, as those offered faster compute times. In addition to comparing ResNet backbones, we also ran each model for two different random seeds to ensure results were independent of the individual image subset selection. Using these comparisons, we also obtained information about the repeatability via values for the average standard deviation of the metrics. Standard deviations of the metrics were 0.04 for Intersection over Union (IoU) score, 0.05 for precision and 0.06 for recall. These levels will be used as indicators of the statistical significance of subsequent results.

Other architectures for neural network based image segmentation exist and have various advantages or disadvantages. We limited ourselves to ResNet architectures in this study because are common, well supported by existing open source implementations and meet the need of being easy to implement on desktop hardware by individual investigators. As our primary interest was in comparing the impact of diversifying the training data on

TABLE I. Backbone Performance Comparison

| Backbone | IoU score by Train/Test Data Set | | | | | |
	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
ResNet-34	0.71	0.61	0.81	0.69	0.63	0.81
ResNet-50	0.68	0.57	0.81	0.67	0.63	0.81
ResNet-101	0.67	0.60	0.80	0.68	0.63	0.83

generalizability of the resultant model, using a fixed architecture still provides the opportunity to explore training data set combinations. Nonetheless, we acknowledge the use of a fixed architecture as a limitation of this work, and leave exploration of the architecture space for future investigations.

B. Training and evaluation

The primary goal of this study was to compare the generalizability of trained models using across diversification of the training data sets. Consequently, to maintain consistency between the models, we used a fixed architecture (ResNet-34 as described previously), loss function and tuned set of hyperparameters for the DE-G data set that were identified by the authors in previous work on the topic[9]. The process focused on workflows achievable by consumer-grade desktop computing hardware (Xeon Silver 4210R processor with 32 GB RAM and NVIDIA RTX A2000 12GB GPU). Code was developed in Python and made use of the *segmentation models* library[19], which utilizes Tensorflow and Keras[20] as the basis for formulating the model architecture. Each model was trained individually from a common starting point. Training of each model required approximately 3 hours, which was a reasonable timeframe for iteration and improvement. Encoder weights were frozen throughout training and early stopping after consecutive 10 epochs without a reduction in the loss function was used to prevent overfitting, retaining the weights with the best validation loss performance. While the 1000 image data sets considered here are relatively small by computer vision standards, we utilized data augmentation with the following parameters to simulate the effects of a larger data set: rotation (up to 30°), zoom (factor of 0.2), and height and width shifts (factor of 0.1 each).

Several metrics were used to evaluate the performance of each individual model. These are based upon the four truth categories for the predictions: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). We relied on the intersection over union (IoU) measure to indicate the overall performance of the models, as it produces values that depend most closely on overall match between ground truth and predictions. Precision and recall are also useful metrics that provide other indications about model performance. Precision indicates the percent of positive predicted pixels that correspond to ground truth positives, while recall indicates the percentage of ground truth positives that were

predicted. Definitions of these metrics are given by the following three equations:

$$IoU = \frac{TP}{TP + FP + FN} \qquad (1)$$

$$p = \frac{TP}{TP + FP} \qquad (2)$$

$$r = \frac{TP}{TP + FN} \qquad (3)$$

C. Source Data

To compare the generalizability of the neural network models, we utilized six data sets that contained labeled PV installations. These data sets represented different sources, resolutions and labeling methodologies. All data sets were filtered to retain only images that contained PV. Data sets with large tiles were first sliced to a manageable size. On final processing, all images were scaled to have an image size of 576 x 576 for compatibility with the model workflow, which resulted in a corresponding scaling up or down of the resolution of each image. Descriptions of the data sets will continue in the following paragraphs, but a summary of each resultant data set is provided in Table II

Two data sets consisted of openly available labels and PV imagery data from nearby cities in California: Fresno (CA-F) and Stockton (CA-S). These were previously published by Bradbury et al.[22]. These data sets consist of 30 cm resolution aerial orthoimagery tiles obtained from the United States Geological Survey (USGS). Images were natively provided as 5000 x 5000 tiles, but were sliced into 625 x 625 tiles for processing by the network. Only tiles containing positive pixels (i.e. the presence of PV arrays) were retained.

Two data sets utilized openly available labels and PV imagery data from France, as described by Kasmi et al.[23]. One of these used data from Google Earth (FR-G) with resolution of 10 cm/pixel. The second used imagery from the French national institute of geographical and forestry information (IGN), with a resolution of 20 cm/pixel and is designated FR-I. These data sets were created using a crowdsourced labeling process, and were unique in that images containing PV were centered on the PV feature. Both data sets used tiles of 400 x 400 pixels.

A data set based on Google Earth imagery from northern Germany is designated DE-G. It was first described in a previous study by the authors[9]. The native tile resolution was 18 cm/pixel and tiles were 639 x 640 pixels. Labeling was conducted manually by visual inspection using the *labelme* software package[24].

The sixth data set consists of 2018 orthoimagery from New York City (specifically Queens) in the United States (designated NY-Q). Data are obtained from the New York GIS Clearinghouse[25]. The native data consist of 5000 x 5000 tiles with a resolution of 15 cm/pixel. As with the California data sets, the tiles were sliced to a size of 625 x 625 prior to processing. Labeling of this data set was conducted by manual inspection of images and construction of polygons around the PV installations using *labelme*[24], and is still ongoing. We hope to make it openly available when complete. The results in this study are based on an initial sampling of around 1000 positive tiles from the in-progress labeling.

TABLE II. Data sets[17]

Data set	Tot. Tiles	Tile Size	Resolution	Scaled Res	Ref
CA-F	1,044	625x625	0.3 m/pix	0.32 m/pix	22
CA-S	4,192	625x625	0.3 m/pix	0.32 m/pix	22
FR-G	13,303	400x400	0.1 m/pix	0.07 m/pix	23
FR-I	7,865	400x400	0.2 m/pix	0.14 m/pix	23
DE-G	1,325	639x640	0.18 m/pix	0.2 m/pix	9
NY-Q[a]	1,007	625x625	0.15 m/pix	0.16 m/pix	25

In order to study the generalizability across data sets, we worked with a fixed data set size of 1000 tiles, roughly corresponding to the number of images in data sets with the least number of positive tiles available (CA-F, DE-G and NY-Q). Holding the number of images fixed at 1000, while representing a relatively small amount of training data, allows us to compare across combinations of these data sets on a fixed quantity of training data. Tiles were chosen randomly from each data set, and a portion of the results were repeated for two random seeds to ensure no issues with statistical representation. The 1000 tile data sets were split for training, validation and test sets with a 72%, 8% and 20% split respectively. We held the images designated to each category to be fixed for all training and evaluation combinations, which ensured that all models were tested on the same data.

We conducted a manual subjective inspection of the 1000 tiles used for each data set in order to provide some representative description of their context[17]. We manually counted images that fit into five bins based on their characteristics: large structures/flat roofs (usually commercial buildings), large open spaces (making up 50% of the image), patterned or row-based agricultural, bodies of water, and utility scale PV. Images not containing one of these features were primarily residential housing. Counts, rounded to the nearest 10, of these images are listed in Table III. For the residential imagery, we also include a rough count of the number of structures that were observed in a typical tile to give an indication of the building density.

These observations are inherently qualitative, but they serve to help describe contextual differences between the data sets. The predominance of residential dwellings was common across all data sets. NY-Q was the most urban of the data sets, with the smallest number of open spaces and the largest number of flat-roofed structures

TABLE III. Contextual Differences by Data set (Approximate)[17]

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

(which were often commercial-scale buildings). NY-Q's urban character was also qualitatively indicated by fewer observable trees within its residential areas as compared to the other data sets. CA-F had the highest incidence of open areas and uniquely contained a significant number of tiles that appeared to indicate agricultural activity in the form of row- or pattern-based vegetative activity. As previously mentioned, the FR-G and FR-I data sets both uniquely centered the images on positively identified PV systems[23], which along with the resolution and tile size differences, explains the difference in building count. This may reduced the probability of PV systems potentially spanning a tile boundary for the FR-G and FR-I sets[16].

D. Models trained using the data sets

In order to assess the generalizability of neural network models, an exhaustive characterization of models trained utilizing the six data sources was performed. To establish baseline performance, we first trained six data set-specific models (one corresponding to each data set) using the sets of 800 training and validation images taken only from a single data set. Each of these six models were evaluated against the 200 images making up the test data associated with each of the six data sets, providing results that showed how well each custom trained model generalized outside of its training data.

In addition to the 6 original data sets, training was also performed using combinations of training data from multiple data sets. Combination data sets always maintained the total size of 800 tiles for training and validation. These tiles were pulled from the previously identified training subsets associated with each data set. Models trained on these combination data sets were used to assess how incorporating more diverse data influenced the performance of a model. All possible combinations of the data sets were considered, including composite training data sets made by selecting from 2, 3, 4, 5, and all 6 data sets. All combination data sets used equal numbers of tiles from their constituent components to the extent possible, with any tiles required to reach 800 selected from the final data set. For example, a data set made up of CA-F, CA-S and FR-G would be composed of 266 images from CA-F, 266 images from CA-S and 268 im-

ages from FR-G, totalling 800. The individual tiles used were chosen randomly from among the training and validation tiles belonging to that data set. When considering all possible combinations, a total of 57 additional models were trained. Each was evaluated individually against the test data associated with each data set.

Finally, we conducted tests that show performance results from including small fractions of training data from the same data source as the test data in the training set. These evaluations were conducted only for models using test data for the NY-Q data set as the target. For three of the data sets (CA-F, CA-S and FR-I), we trained models that replaced a fraction of the training data with varying levels of data from NY-Q (1, 2, 3, 4, 5, 10, 20, 30, 40 and 50%). As with the other conditions evaluated, these models were all trained on a fixed number of tiles (800 total for training and validation). In conjunction with the combination data set that includes for example both CA-F and NY-Q data (which represents 50% data from NY-Q), these allow us to determine exactly how small amounts of data from a target test set can improve the performance of a model.

IV. RESULTS AND DISCUSSION

A. Models Trained on a Single Data set

First, we describe results for the baseline of models trained on a single data set. Results for the IoU are given in Table IV. Results of precision and recall for the models are presented in Table V and Table VI respectively. Averages shown in these tables exclude tests on the data corresponding to a model's training data, in order to show the overall generalized performance without being skewed by differences in the absolute predictability of a set of test data.

We can make some inferences about the model performance based on the conjunction between these metrics. For example, when FR-I predicts on the FR-G data set, it is often correct but very selective. It has a high precision (95%) indicating that it is usually correct when making predictions, by a low recall (36%) indicating that it does not identify a large share of the ground truth pixels. An example of the converse is available for CA-F predicting DE-G, where performance is achieved by erroneous over-

TABLE IV. IoU Values by Data set[17]

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q	Avg
CA-F	0.71	0.35	0.11	0.36	0.06	0.16	0.21
CA-S	0.55	0.61	0.11	0.22	0.17	0.19	0.25
FR-G	0.03	0.00	0.81	0.45	0.13	0.26	0.17
FR-I	0.13	0.19	0.35	0.69	0.31	0.56	0.31
DE-G	0.18	0.29	0.11	0.29	0.63	0.44	0.26
NY-Q	0.07	0.22	0.15	0.47	0.40	0.81	0.26
Avg	0.19	0.21	0.17	0.36	0.21	0.32	

Train Data set in Rows
Test Data set in Columns
Averages exclude the diagonal to highlight the results for unseen data

TABLE V. Precision Values by Data set[17]

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q	Avg
CA-F	0.87	0.46	0.36	0.48	0.07	0.25	0.33
CA-S	0.82	0.79	0.51	0.31	0.22	0.24	0.42
FR-G	0.10	0.03	0.91	0.76	0.41	0.52	0.36
FR-I	0.63	0.64	0.95	0.79	0.67	0.77	0.73
DE-G	0.70	0.65	0.83	0.91	0.77	0.82	0.78
NY-Q	0.59	0.66	0.90	0.87	0.75	0.90	0.75
Avg	0.57	0.49	0.71	0.67	0.42	0.52	

Train Data set in Rows
Test Data set in Columns
Averages exclude the diagonal to highlight the results for unseen data

prediction of the incidence of PV. In this case, modest recall is observed (59%) which means that many ground truth pixels are identified, but at the cost of predicting a significant number of false positives, indicated by the 7% score in recall.

It is unsurprising to observe that models generally performed best when predicting their corresponding test data, where the average IoU score across all models was 0.71. From the tables above, we can also observe that with a few exceptions in the precision metric, generalization was relatively poor; a model trained on a given data set was the best performer on test data for that data set. A few examples of moderate skill at generalization was observed. The best examples occurred with FR-I predicting NY-Q test data and the model trained on CA-S predicting CA-F. In the case of California, this may arise somewhat from the shared data source, but it is likely that confounding factors are at play, because no performance boost to CA-F was seen when predicting CA-S.

The FR-I trained model had the most generalizable performance to other test sets with an average IoU of 0.31. The FR-G trained model showed the worst individual example of generalization, especially on CA-F and CA-S where it showed virtually no predictive skill. FR-G compared to the California sets had the greatest discrepancy in resolution, which may suggest that models do not adapt well to lower resolution imagery (i.e. lower zoom levels). We can also consider the difficulty of the task for an arbitrary model but averaging across models for a given test data set. The test sets with the highest

average IoU scores across all models were FR-I (IoU = 0.36) and NY-Q (IoU = 0.32), showing greatest ease for prediction by a general model. For all these cases, it is important to note that none of these examples of generalization achieved the performance of the model trained on the corresponding data set.

The difficulty of predicting across multiple resolutions may also been inferred via the fact that the FR-G test data was most difficult to predict by other models, with an average IoU of 0.17. Investigating the precision and recall shows that models were usually correct when predicting positive values, but were hesitant to do so, as indicated by the low recall. When observing the predictions on a detailed level, it is possible to notice that some models tended to discretize the individual panels on FR-G, indicating that they apparently interpreted the frames of the modules as gaps in the array, which was not true for FR-G's own predictions. This may be an indicator that when training models at lower zoom levels (for which the frames are generally not resolvable), a degree of confusion arises in the predictions when applied at higher zoom, because additional physical features can be resolved. An example of this effect is shown in Fig. 1. In this case, the frames of the panels (which were not always visible in the NY-Q images) resolve with a width of multiple pixels in the FR-G images. We note that this occurred despite the zoom augmentation in the training methodology.

TABLE VI. Recall Values by Data set[17]

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q	Avg
CA-F	0.79	0.59	0.15	0.58	0.59	0.35	0.45
CA-S	0.62	0.72	0.13	0.47	0.59	0.37	0.44
FR-G	0.06	0.01	0.88	0.52	0.15	0.29	0.20
FR-I	0.15	0.23	0.36	0.84	0.37	0.67	0.35
DE-G	0.19	0.33	0.11	0.30	0.79	0.48	0.28
NY-Q	0.07	0.24	0.15	0.50	0.47	0.89	0.28
Avg	0.22	0.28	0.18	0.47	0.43	0.43	

Train Data set in Rows
Test Data set in Columns
Averages exclude the diagonal to highlight the results for unseen data
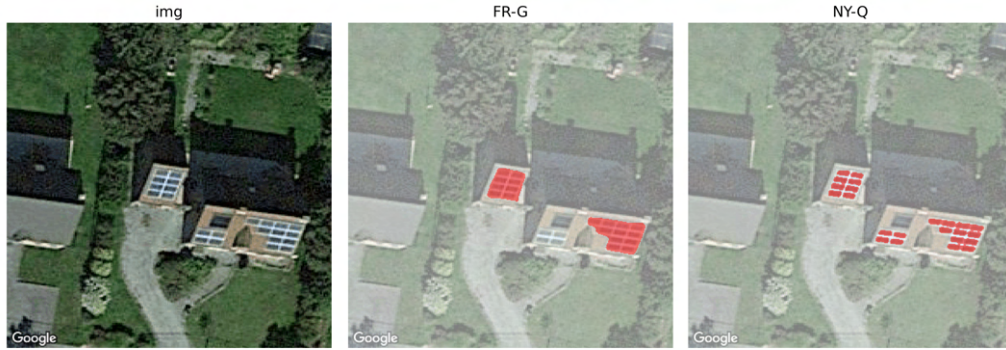
FIG. 1. Example of discretization of the individual array modules in an FR-G image when predicted by the NY-Q trained model.

B. Models Trained on Combined Data

Since no individual models generalized well across other data sets, we investigated how training models on data from multiple sources affects their performance at predictions. We will refer to these as models trained on combination data sets. Tests were made considering two modalities. First, we trained and tested combination models on data they had seen, that is, using a fraction of training data corresponding to the target test data set, paired with model from additional data sources. This investigation answers the question "Compared to training on exclusively the target data source, can exposure to more diverse data improve the performance of a model?" Second, we looked at combination models tested on completely unseen data, which is to say that no data from the test data's data source was used in training. In this case a three data source combination trying to predict NY-Q would never contain training data from NY-Q (e.g. could be trained on CA-F, FR-I and DE-G). This framing attempts to answer the question, "Does exposure to more diverse data improve the generalizability of a model for data from an unknown source?" In either case, we tested all possible combinations of training data sources that met the modality criteria. As stated previously, for combination models the total number of 800 training and validation images was maintained, and splits between the training data sources were as even as possible. Results

for both seen and unseen combination tests are discussed below.

1. Combination Models for Predicting Data from a Seen Data Source

Combination models for predicting seen data were trained using training data from the target test data source combined with data from an additional n number of data sources. That is to say, these combination models always contained some data from the same data source as the corresponding test data set. So a three data source combination model attempting to predict NY-Q test data might be trained on CA-F, FR-G and NY-Q data. Combinations were trained making use of data from 2, 3, 4, 5 or 6 total data sets (corresponding to 50, 67, 75, 80 and 83% training data from sources other than the test data source, respectively). Results on IoU score, precision and recall as a function of number of data sources are shown in Figs. 2 - 4. Results shown are normalized to performance of a model trained solely on data from the corresponding test data source (i.e. from Table IV) to better indicate trends when accounting for performance offsets. A tabulated form of Fig. 2 is shown for two source combinations only in Table VII.

Results show that adding additional diverse data to a model on average worsens its IoU score performance as

compared to models trained specifically on a given data source. This performance degradation is greater as more diverse data is included in the training set (i.e. less of the target data is used). This trend is universal across all test data sources, though some cases (e.g. CA-F) experience a greater degree of performance loss. Simply put, this result indicates that diversifying the fixed-size training data set never increased its performance over a model trained exclusively on the target data source, and thus simply diversifying the training data could not be recommended as a strategy for creating a more generalizable model. While the data does not allow a definitive conclusion to be drawn for this phenomenon, we hypothesize that it occurs due to displacement of the target data in the training data set (i.e. a combination of two data sources contains only half as many images from a given data source as each would individually). That is, the raw number of images from the target data appears to be more important than the diversity of the training data.

The results may be investigated in more detail by considering the effects on precision and recall. In the case of precision, we observed that introducing more diverse data may for some cases lead to an increase in the precision of models (i.e. make them less likely to make false positive predictions), by up to 7% relative to the baseline for the highest precision models, but did not reach the level of statistical significance based on the repeatability described previously. These are indicated by bars with values above 1.0 in Fig. 3. For cases where the precision was negatively impacted, performance reductions did not exceed more than 20% from the baseline. In the case of recall, very slight improvements in performance were observed when training on diverse data in a limited number of cases, but the potential negative impacts were quite significant (up to 40% reduction in performance). As in the case of IoU, negative impacts were more likely when including data from a greater number of different data sources. So we could conclude that creating combination data sets on seen data may make models more precise in predictions, but any benefits are outweighed by reductions in the recall leading to an overall degradation of performance.

In sum, these results indicate that as compared to custom-training for a model to make predictions on a given data source, adding a more diverse set of data does not improve overall model performance. Conversely, the IoU score performance was always reduced as compared to a baseline trained exclusively on data from the test data source. Reductions in performance seemed to be dominated by loss of recall, as a few cases actually exhibited increases in precision resulting from increased diversity in the training data. These models would be less likely to predict false positives, but always at the cost of predicting an increased number of false negatives. However, because there was no discernible pattern to the likelihood of improving precision, it would be difficult to deliberately produce a model with these characteristics
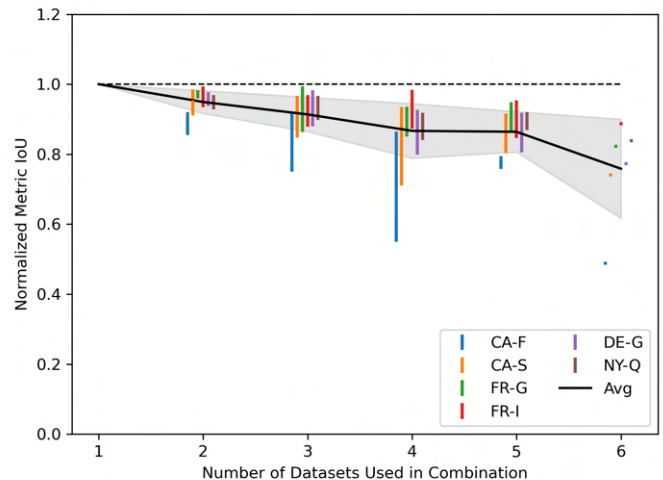

FIG. 2. IoU score performance on combination models tested on seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data (always including the test data in this case). Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.
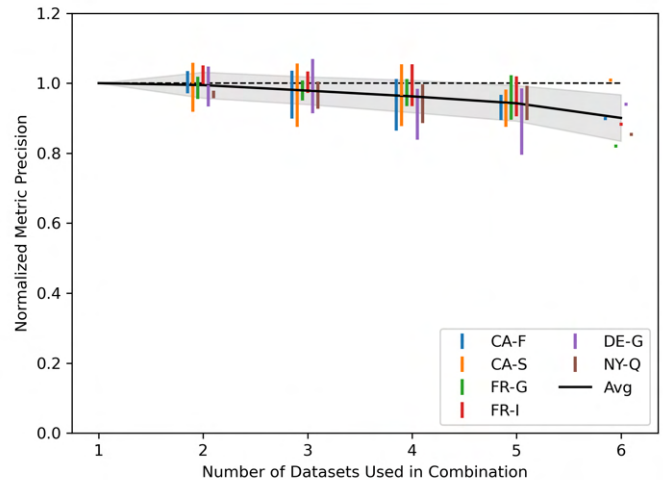

FIG. 3. Precision performance on combination models tested on seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data (always including the test data in this case). Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.

on an a priori basis. These results indicate that to get the best performing model on a given test data set, one should favor training on as much data from that data source as possible.
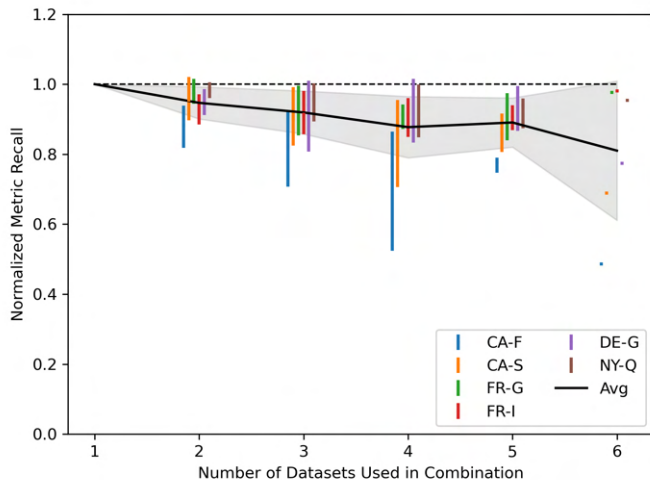
FIG. 4. Recall score performance on combination models tested on seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data (always including the test data in this case). Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.
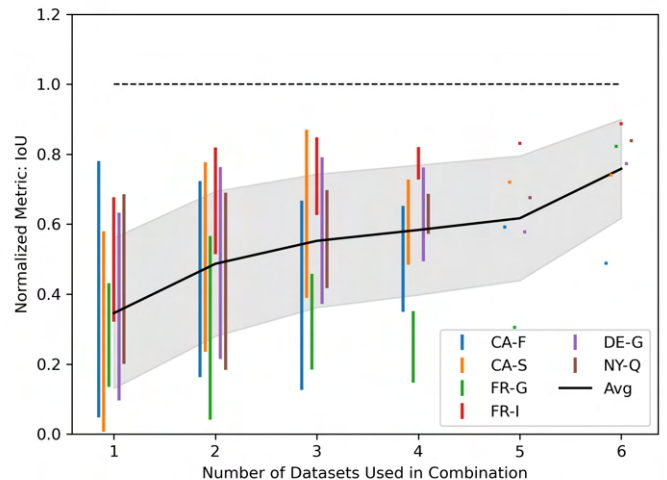
FIG. 5. IoU score performance on combination models tested on never seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data. Note that "6" case always includes seen data. Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.

TABLE VII. Two Set Seen Combination Normalized IoU Performance by Test Data set

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
CA-F & CA-S	0.92	0.95	-	-	-	-
CA-F & FR-G	0.86	-	0.99	-	-	-
CA-F & FR-I	0.89	-	-	0.99	-	-
CA-F & DE-G	0.89	-	-	-	0.96	-
CA-F & NY-Q	0.92	-	-	-	-	0.97
CA-S & FR-G	-	0.92	0.96	-	-	-
CA-S & FR-I	-	0.93	-	0.96	-	-
CA-S & DE-G	-	0.98	-	-	0.96	-
CA-S & NY-Q	-	0.97	-	-	-	0.97
FR-G & FR-I	-	-	0.98	0.94	-	-
FR-G & DE-G	-	-	0.96	-	0.93	-
FR-G & NY-Q	-	-	0.98	-	-	0.93
FR-I & DE-G	-	-	-	0.99	0.98	-
FR-I & NY-Q	-	-	-	0.93	-	0.97
DE-G & NY-Q	-	-	-	-	0.95	0.95

2. Performance of Combination Models on Unseen Test Data

We also tested a suite of combination models whose constituents never included data from the test data's source. These models were trained using 2, 3, 4 or 5 data sources, representing all combinations that did not include the test set. Data for a single unseen source correspond to those described in section IV A, but are included with these results as a baseline. Combination results for six data sources necessarily include the test data source and thus do not fully meet the unseen criteria, but those results are also included here for comparison purposes.

Results are shown for IoU in Fig. 5, for precision in Fig. 6 and recall in Fig. 7. A tabulated version of 5 for two source combinations only is given in Table VIII (note that this is essentially the complementary values to Table VII).

Results on IoU in Fig. 5 show that no combination models performed nearly as well as the custom trained models, however on average, performance in all three metrics did improve by diversifying the training data. A much higher degree of vertical spread was observed, indicating a much greater degree of variability in the performance of individual combinations. Inspecting results for the other metrics, we see a similar result to that for the seen models, in that it was common for combination models to exceed the precision of custom-trained models, and in this case, did so at statistically significant levels in the extreme cases. This increase in precision is never accompanied by improvement in the IoU score as compared to the custom trained models. We also may observe that the case of testing on FR-G data appears to be an outlier, for which no combination of data resulted in substantial performance improvements.

These results indicate that on average, including more diverse training data tends to improve the performance of models at identifying PV arrays. However, when considering the best individually performing model for any given test set, no universal pattern emerged. That is, some test sets experienced highest IoU score performance in a model using 2 or 3 data sets, while one case's best performing model used only a single data set. Thus, utilizing data from as many data sets as possible increases
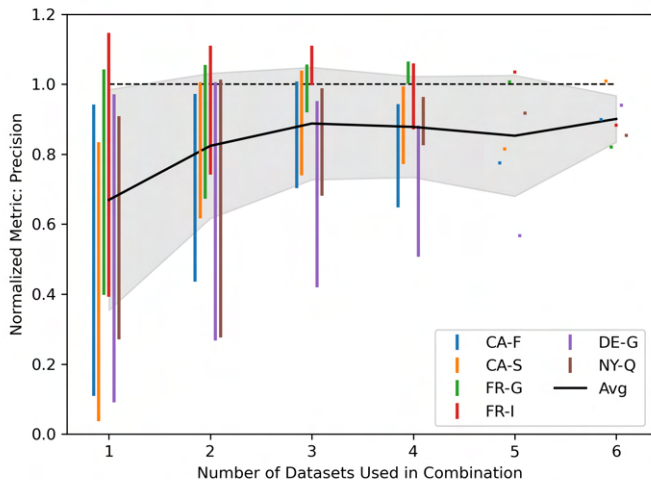
FIG. 6. Precision performance on combination models tested on never seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data. Note that "6" case always includes seen data. Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.
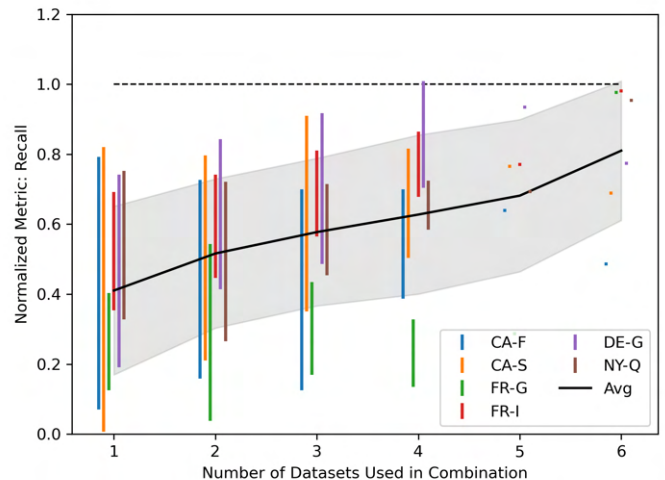

FIG. 7. Recall score performance on combination models tested on never seen data by test data set. Horizontal axis shows the number of individual data sets used in making up the training data. Note that "6" case always includes seen data. Markers indicate range of individual combination model performance. Line shows average performance at this combination size, while area shows first standard deviation range across all models.

the probability of achieving moderately accurate performance from a generalized model, but might not lead to the best model performance overall. The best way to ensure high performance is to include data from the target data source model in the training data, as seen by comparing the absolute values of IoU score between Figs 2 and 5.

It is worth emphasizing that Fig. 5 clearly shows that no combination model performed better (or even nearly as well) on IoU score as a model trained on data from the target data source. This suggests that seeking to produce truly generalized models may not yield the most satisfactory results regardless of the care taken in their selection of their makeup for training.

C. Image-wise Performance

To better understand some of the performance of the models, we visualized the individual images with the best- and worst- IoU score on an image-wise basis, shown in Figs. 8 and 9 respectively. In these cases, best and worst IoU scores were determined by averaging across the performance from the models trained on a single data set.

While the representations of best images are inherently anecdotal, a few comments can be made about their shared characteristics. In the images representing the best average predictions, PV systems tended to be large and rectilinear. In the case of NY-Q test data in particular, all five images show large commercial rooftop systems. CA-F is the possible outlier to these observations,

TABLE VIII. Two Set Seen Combination Normalized IoU Performance by Test Data set

	CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
CA-F & CA-S	-	-	0.13	0.53	0.21	0.18
CA-F & FR-G	-	0.69	-	0.82	0.46	0.38
CA-F & FR-I	-	0.70	0.35	-	0.41	0.63
CA-F & DE-G	-	0.74	0.06	0.65	-	0.53
CA-F & NY-Q	-	0.78	0.19	0.79	0.76	-
CA-S & FR-G	0.63	-	-	0.66	0.44	0.50
CA-S & FR-I	0.68	-	0.31	-	0.55	0.58
CA-S & DE-G	0.55	-	0.04	0.51	-	0.64
CA-S & NY-Q	0.72	-	0.30	0.77	0.74	-
FR-G & FR-I	0.16	0.24	-	-	0.43	0.63
FR-G & DE-G	0.31	0.45	-	0.66	-	0.53
FR-G & NY-Q	0.17	0.41	-	0.74	0.61	-
FR-I & DE-G	0.41	0.49	0.57	-	-	0.69
FR-I & NY-Q	0.16	0.41	0.26	-	0.62	-
DE-G & NY-Q	0.31	0.45	0.24	0.65	-	-

because four of the five best performing images appear to show small rooftop arrays. Additionally, the solar arrays in these images appear to be similarly colored. To attempt to quantify the preference for identifying large systems, we computed the Pearson correlation coefficient between the number of positive pixels in the labeled image (related to the overall array size) and the average IoU score for the image across all unseen test sets. All data sets except FR-G showed a small, albeit statistically significant imagewise relationship between number of pixels and IoU score ($p < 0.005$), with the level of association varying from $\rho = 0.21 - 0.49$. This indicates that most of these models do tend to perform better on larger sys-

tems. The lack of a relationship in the FR-G data may result from the relative scarcity of larger systems in that data set (as indicated in Table III).

The worst performing images were predominantly of residential housing. Again, despite the anecdotal nature of these data, a few observations can be made. Many of these images across all data sets contain examples where there is very little contrast between the roof and the array. In some cases from FR-G, we observe instances where the camera sensor seems to have saturated due to solar reflection, and these were also difficult to predict. The difficult to predict images from NY-Q are all examples of very small segments that are cropped by the edges of the frame. One of the poorly performing images from CA-F contains examples of agricultural rows (present in several other images from that data set as well). It was common for models not trained on CA-F data to predict false positives on the regularly spaced rows in this and similar images, which we hypothesize is due to their regular patterned structure that may resemble the rows of large-scale PV installations. An example is shown in Fig. 10.

D. How Much Target Data is Needed for Good Performance?

Given that no generalized model performed as well as the custom trained models, we investigated what quantity of training data corresponding to the test set could result in improvements to performance. Given the large number of possible combinations, these investigations were conducted only for test data corresponding to the NY-Q data set. Models were trained by combining each of the data sets with a fraction of data from NY-Q. Fractions of 1, 2, 3, 4, 5, 10, 20, 30, 40 and 50% were investigated. Results for testing on NY-Q data are shown in Fig. 11. As evident, including data from the target data source quickly improves the performance of the model. While there is a degree of subjectivity to interpretation, diminishing returns are reached around 20% of the NY-Q data.

We also looked at the extent to which incorporating the small quantities of NY-Q data affected their ability to predict their corresponding test data. Those results are depicted in Fig. 12. Overall, models remain at or above about 90% of their baseline IoU value. The degree of impact varies by data source, with CA-F being most affected. A few models obtain IoU score performance that exceeds the baseline with small amounts of NY-Q data. While this does suggest the possibility of introducing very small amounts of diverse data to improve the performance of a model, it is important to note that these increases are very small and none reached the level of statistical significance relative to average IoU standard deviation of 0.04 described in Section III A. In any event, the improvements seen are not universally beneficial, nor does introducing the data produce a consistent trend that

could be used to try to produce a more effective model.

Coupled with the previous results on combination models, these data suggest that there is no reliable method to produce a generalizable neural network model for segmentation of PV within images simply by diversification of the training data. Rather, the best method for producing reliable neural network models for this task is inclusion of some labeled data from the target data source in training. Fractions of at least 20% produced models that reached performance around 90% of the normalized level produced with full labeling of training data from the target, regardless of the initial quality of the model without NY-Q data. The choice of training effort ultimately requires a balancing decision between desired model quality and labeling effort.

V. CONCLUSION

Use of trained neural networks remains an attractive option for remote identification of PV systems for a variety of research and decision making tasks. We conducted a comprehensive study of training a ResNet–based neural network for this purpose and whether the possibility exists to generalize such a model based on diversification of the training data. The experimental design used a fixed size and architecture to control for the impact of those effects on the results. Our study may aid researchers in planning training approaches for development of models for identification of PV from aerial images, especially when considering the breadth of data sources currently available.

We obtained a negative result for the ability to improve performance of a custom trained model by diversifying its training data, while using a fixed total number of training images. We did not observe any combination models where incorporating training data from additional sources increased a model's ability to predict the location of PV in images relative to a model trained exclusively on the target data. We did observe very slight (1-2%) increases in IoU score corresponding to introduction of less than 10% NY-Q data that were not statistically significant relative to the repeatability of the IoU metric. This implies that when preparing to utilize a model with new data sources, achieving the best performance will require some degree of labeling of the new source, and researchers should plan to make balanced decisions regarding desired model performance against the invested labeling effort.

Our results on predicting unseen data showed agreement with previous studies that were based on fewer data sources. When investigating models that are tested on completely unseen data, it was not possible to create truly generalized models, regardless of the combinations of training data used at the fixed size of 1000 training images. While incorporating data from many data sources into training data had the potential to yield a more general model on average, these models did not approach the

FIG. 8. Five images for each test set with best averaged IoU score across all models trained on a single data set.
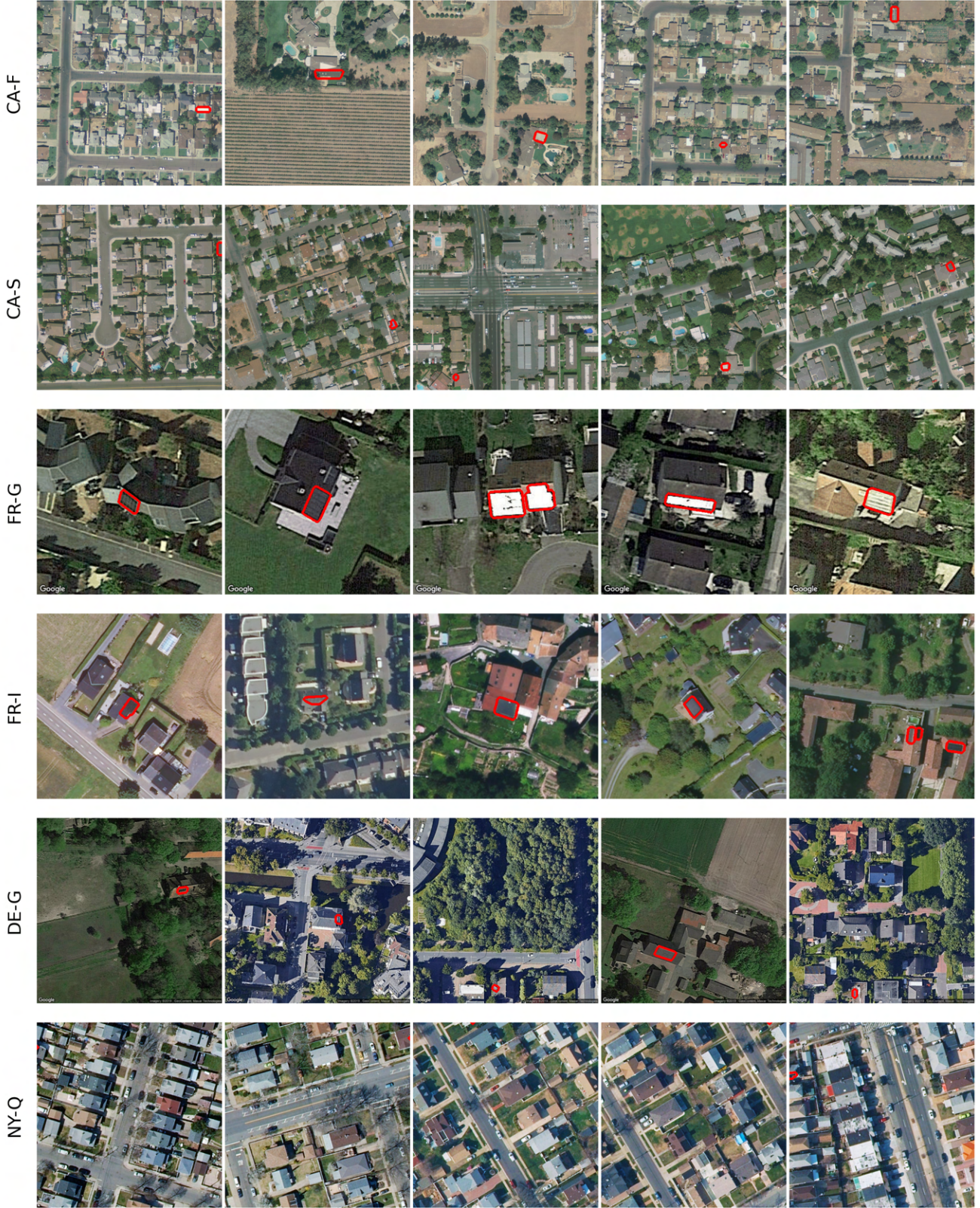
FIG. 9. Five images for each test set with worst averaged IoU score across all models trained on a single data set.

FIG. 10. Example of false predictions of PV by the FR-I trained model on the rows of agricultural activity seen in CA-F.
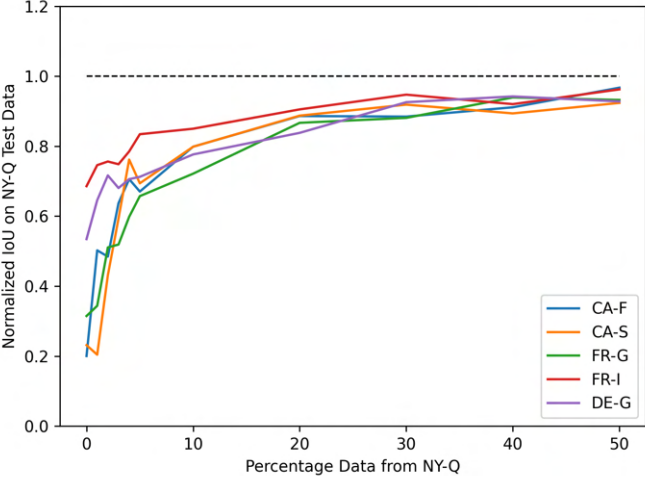

FIG. 11. IoU score results on NY-Q for inclusion of incremental amounts of NY-Q data to improve performance. Colored label corresponds to the data source paired with NY-Q in the combination.
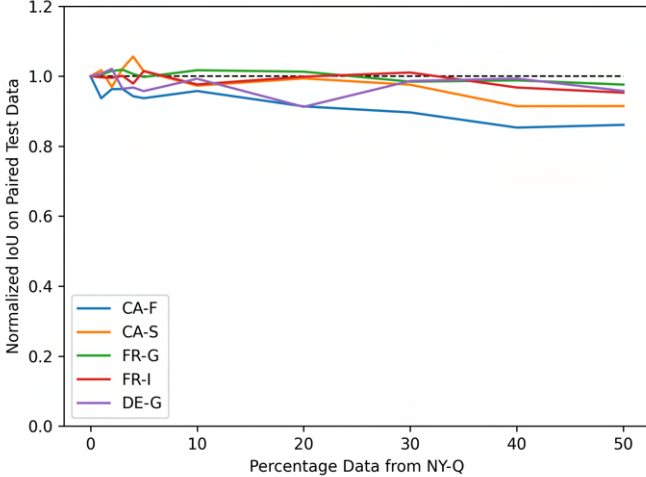
FIG. 12. IoU score results on corresponding test data for inclusion of incremental amounts of NY-Q data to improve performance. Colored label corresponds to the data source paired with NY-Q in the combination.

performance of models trained specifically on data from the target data source. When predicting unseen data, the average model trained on maximally diverse training data produced IoU score results around 60% as effective as models custom trained on the test data's source. As before, this result indicates to researchers planning to apply models for PV identification purposes on a new data source that it is necessary to include some labeled data on the new source to achieve successful models.

Achieving the best performance therefore necessitates the incorporation of some degree of labeled data from the target for the purpose of training. We investigated the degree to which small amounts of labeled data could improve performance, with the intention of allowing investigators to minimize the labeling effort. Our investigation indicates that the greatest gains come during the addition of up to 20% data from the target data set, after which the magnitude of returns diminish. Future work may consider more advanced methodologies for training to reduce the labeling effort required to gain these bene-

fits.

Results of this study are limited by the experimental methodology employed in this study, which utilized a fixed model architecture commonly used by many other investigators, and was limited to a fixed training data set size of 1000 images. Further research may be necessary to conclusively determine whether more sophisticated model architectures could overcome the limitations on generalizability seen here, or whether models trained on a substantially larger data set would produce more favorable results. However, the results obtained from this study provide insight to researchers who are hoping to apply neural networks for PV identification utilizing close to "off-the-shelf" approaches. When using publicly available data sets, architectures accessible through common open source packages and desktop hardware, researchers should plan for some labeling effort to apply PV identification models to unseen data sources.

We look forward to the opportunities that neural net-

work based PV identification models offer to help answer questions related to growth, access and affordability of distributed solar generation.

ACKNOWLEDGMENT

[1] Paul Denholm, Patrick Brown, Wesley Cole, Trieu Mai, Brian Sergi, Maxwell Brown, Paige Jadun, Jonathan Ho, Jack Mayernik, Colin McMillan, and Ragini Sreenath, "Examining Supply-Side Options to Achieve 100% Clean Electricity by 2035," Tech. Rep. NREL/TP-6A40-81644 (National Renewable Energy Laboratory, 2022).

[2] 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," Nature Communications **12**, 5738 (2021), number: 1 Publisher: Nature Publishing Group.

[3] W. Hu, K. Bradbury, J. M. Malof, B. Li, B. Huang, A. Streltsov, K. Sydny Fujita, and B. Hoen, "What you get is not always what you see—pitfalls in solar array assessment using overhead imagery," Applied Energy **327**, 120143 (2022).

[4] K. Perry and C. Campos, "Panel Segmentation: A Python Package for Automated Solar Array Metadata Extraction Using Satellite Imagery," IEEE Journal of Photovoltaics **13**, 208–212 (2023), conference Name: IEEE Journal of Photovoltaics.

[5] 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 **2**, 2605–2617 (2018).

[6] Z. Wang, M.-L. Arlt, C. Zanocco, A. Majumdar, and R. Rajagopal, "DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models," Joule **6**, 2611–2625 (2022).

[7] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," Applied Energy **183**, 229–240 (2016).

[8] K. Mayer, B. Rausch, M.-L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal, "3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D," Applied Energy **310**, 118469 (2022).

[9] M. Zech and J. Ranalli, "Predicting PV Areas in Aerial Images with Deep Learning," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)* (2020) pp. 0767–0774, iSSN: 0160-8371.

[10] 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, 2020).

[11] L. Kruitwagen, K. T. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," Nature **598**, 604–610 (2021), number: 7882 Publisher: Nature Publishing Group.

[12] Allen Institute for AI, "Satlas,".

[13] R. Zhu, D. Guo, M. S. Wong, Z. Qian, M. Chen, B. Yang, B. Chen, H. Zhang, L. You, J. Heo, and J. Yan, "Deep solar PV refiner: A detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery," International Journal of Applied Earth Observation and Geoinformation **116**, 103134 (2023).

[14] Z. Guo, Z. Zhuang, H. Tan, Z. Liu, P. Li, Z. Lin, W.-L. Shang, H. Zhang, and J. Yan, "Accurate and generalizable photovoltaic panel segmentation using deep learning for imbalanced datasets," Renewable Energy **219**, 119471 (2023).

[15] 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)* (2017) pp. 1–8, iSSN: 2332-5615.

[16] P. Li, H. Zhang, Z. Guo, S. Lyu, J. Chen, W. Li, X. Song, R. Shibasaki, and J. Yan, "Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning," Advances in Applied Energy **4**, 100057 (2021).

[17] J. Ranalli and M. Zech, "Generalizability of Neural Network-based Identification of PV in Aerial Images," in *2023 IEEE 50th Photovoltaic Specialists Conference (PVSC)* (2023) pp. 1–7.

[18] 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*, Lecture Notes in Computer Science, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Springer International Publishing, Cham, 2015) pp. 234–241.

[19] P. Yakubovskiy, "Segmentation Models," (2019).

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Tech. Rep. (Google, 2015).

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009) pp. 248–255, iSSN: 1063-6919.

[22] K. Bradbury, R. Saboo, J. Malof, T. Johnson, A. Devarajan, W. Zhang, L. Collins, R. Newell, A. Streltsov, and W. Hu, "Distributed Solar Photovoltaic Array Location and Extent Data Set for Remote Sensing Object Identification," (2018), https://dx.doi.org/10.6084/m9.figshare.3385780, publisher: figshare Type: dataset.

[23] G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus, "A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata," (Zenodo, 2022) type: dataset.

[24] Wada, Kentaro, "labelme," (2024).

[25] "NYS Interactive Mapping Gateway,".