Exploring Robustness of Stanford's DeepSolar to Distribution Shifts

Kerrie Wu
Stanford University
kerriewu@stanford.edu

Julian Cooper
Stanford University
jelc@stanford.edu

Andrea van den Haak
Stanford University
vandenhaak@stanford.edu

1. Introduction

We are interested in segmenting and classifying rooftop solar panels from satellite images, with a focus on residential areas. In many countries (including the United States), the rollout of rooftop solar has been ad hoc and largely untracked. Besides industrial sites and city buildings, electricity utilities are largely blind to where and how much solar neighborhoods have accumulated. Being able to estimate this (and its rate of increase) allows utilities to predict power output reasonably well (from meteorology forecasts) and adjust grid management and investment. [1]

2. Problem Statement

Stanford's DeepSolar team have largely solved this challenge for the United States [8] [4]. However there are two major ways that the model can be improved: (1) segmentation accuracy and (2) robustness to distribution shifts.

The original DeepSolar model has two branches, a classification branch that detects if there is a solar panel in an image, and a segmentation branch that produces a class activation map based on the image, which can be used to estimate the total area of solar panels in the image if the classification branch predicts positive [8]. The output class activation map is much lower in resolution than the input image size in the model, which means the segmentation is approximate and limited in detail. If the segmentation is more accurate and at the same resolution as the input, it could be used to predict the type of solar panel, and generate a more accurate estimate of the solar panel area. Therefore we seek to improve the resolution of the segmentation branch in our project.

Secondly, the DeepSolar model is not robust to "distribution shifts". This means that the model accuracy drops when tested on datasets from other regions with different satellite imagery to its USA-based training data. With the help of the DeepSolar team, we investigate different techniques for modifying the DeepSolar segmentation model in order to improve its robustness to distribution shifts. This will enhance the model's generalizability for use in countries other than the United States by requiring less (or ideally zero!) labeled data from other countries for finetuning the model.

3. Literature Review

We are looking into augmenting the segmentation branch of the model to produce more accurate segmentation outputs by incorporating Facebook's Segment Anything Model [2]. The Segment Anything Model is a foundational transformer vision model capable of zero-shot image segmentation, given an input image and point prompts of portions of the image to include, or exclude, in the mask. We anticipate that due to its extensive pretraining, we can use it to produce accurate segmentation masks for solar PV arrays that are the same size as the original input image without any finetuning.

In addition to incorporating the Segment Anything Model, we have also reviewed several papers on techniques for measuring robustness to and handling distribution shift.

- Taori, 2020 [5]: Major contributions included defining effective robustness and relative robustness metrics. The authors also concluded that robustness to synthetic data distribution shifts do not imply robustness to natural distribution shifts, and that a more diverse training set improved robustness.

- Yao, 2022 [7]: Describes a data augmentation method called LISA adds interpolations between original input-output training example pairs to the training data. The interpolation pairs can be selected in a targeted manner to improve robustness to domain shifts.

- Volpi, 2018 [6]: Describes how to adversarially augment the training data to achieve a more robust model. At each training iteration, the training examples fed to the model are augmented with examples that are considered difficult for the current model.

4. Data Set

We make use of two satellite imagery datasets for this project: (1) the original Deepsolar United States dataset, published in 2017 [8], and (2) a recently published dataset for France (released January 2023) [1].

The United States dataset is large and highly skewed. It includes 46,090 images with solar panels (positive) and 366,467 images without solar panels (negative). For positive images, we also have access to masks with 1 and 0 pixels indicating exactly where in the image the solar panel lies. The dataset covers most industrial and residential regions of the country. [8]

The France dataset provides ground truth segmentation masks for 13,303 images from Google Earth25 and 7,686 images from the French national Institute of Geographical and Forestry Information (IGN). It includes no negative satellite images (no ability to perform classification) and focuses on more rural and residential regions. [1] We will use this to test our model's ability to handle distribution shifts by training our model on United States data, but validating and testing our model on France data.

5. Technical Approach

Before starting our analysis, we needed to invest significant work in standing back up the DeepSolar model, including (1) rewriting some components from TensorFlow into PyTorch, (2) fixing bugs associated with outdated APIs, (3) ensuring code could natively make use of GPU resources, (4) rebuild partially broken image processing pipeline, and (5) reproduce baseline figures to convince ourselves that any fixes had been properly implemented.

To incorporate the Segment Anything Model [2], we adopt a similar data processing pipeline as the original DeepSolar Team [8]. We feed an input image to the original DeepSolar classification/segmentation model. If the DeepSolar model's classification branch predicts positive, then we use the segmentation branch's output class activation map (CAM) to generate point prompts to SAM: three positive (to include in the mask), and three negative (to exclude from the mask). As an initial approach, we normalize the CAM map to values between 0 and 1, and randomly select three points where the scaled CAM value is above 0.5 as positive points, and three points where the CAM value is below 0.1 as negative points. To evaluate performance and compare to baseline, we will use intersection over union and percent error in estimated solar PV area.

To handle the distribution shift between our United States and France datasets, we plan to experiment with interpolation-based data augmentation through LISA [7], regularized fine-tuning [3], and adversarial data augmentation [6]. To measure how well our model handles distribution shifts, we will use indsutry standard quantitative metrics such as effective robustness and relative robustness as defined by Taori [5]. (Note, we include percent error in predicted solar PV area to measure segmentation model performance irrespective of robustness.)

6. Intermediate/Preliminary Results

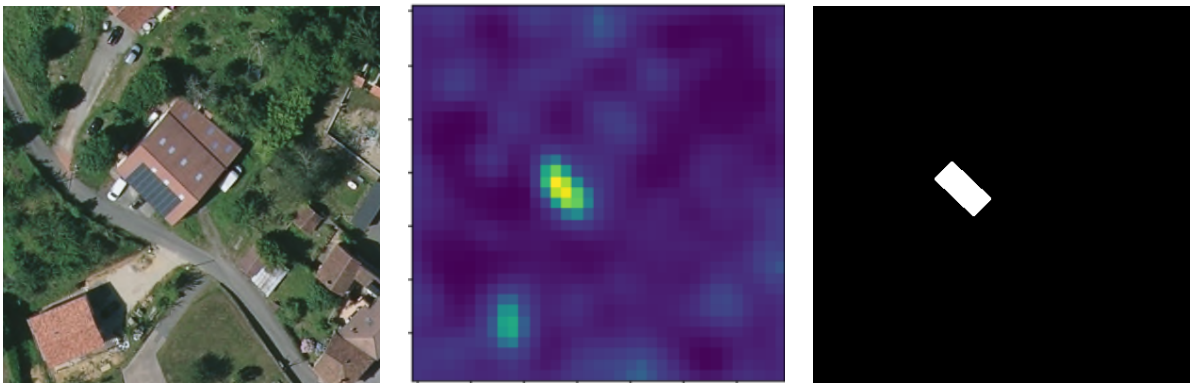
Table 1 summarizes our preliminary results as of the milestone deadline. The first row shows the baseline precision, recall, and area percentage error for both the United States and French datasets when using the pre-existing DeepSolar pretrained model. We confirm that for the United States test data, our results match those from the DeepSolar paper [8], with impressive classification results, but poor segmentation accuracy. This gives us confidence that our efforts to bring the codebase up-to-date in order to use modern packages was done correctly.

Baseline performance on our France dataset was worse than expected, especially for classification. This may indicate significant opportunity to improve DeepSolar's ability to handle distribution shifts (known to be an issue), but at this point we suspect there is still a bug in how we process and pass the France data into our model pipeline. That said, for the few images that were correctly classified (and segmentation attempted), we actually achieve very reasonable segmentation. See example CAM segmentation in Figure 1.

The second line of Table 1 illustrates our first results from trying to enhance the segmentation classifier of our DeepSolar model. While the results are not amazing, they are already a meaningful improvement from the baseline. See Figure 2 for visualizations of several example outputs. For improving usage of SAM, we can investigate different methods of picking point prompts. This can include changing the number of point prompts provided to SAM, and different selection methods (eg, tuning the CAM value thresholds, or sampling based on a probability distribution instead of randomly).

Model	US dataset				French dataset			
	precision	recall	area % error	IOU	precision	recall	area % error	IOU
Deepsolar US (baseline)	0.95	0.86	-8.0%	NA	1.00	0.001	0.47%	NA
Deepsolar US + SAM	0.95	0.86	47.0 %	0.24	TBD	TBD	TBD	TBD

Table 1. Preliminary results showing performance of the baseline model (Deepsolar US) on the original US dataset, compared to results with the same model on the French dataset. We also compare US dataset results to those of the original Deepsolar US model but using Facebook's Segment Anything Model for segmentation. Precision and Recall are both based on the image classification performance of the model (which are the same between the baseline and with SAM). Percent area error measures the percent difference between the predicted solar PV area and the true solar PV area as measured by the true segmentation masks. IOU stands for intersection over union and is calculated only for models using SAM because the CAM maps produced by the original Deepsolar model are lower resolution than the true segmentation labels, and therefore can't be used to calculate an IOU score. Future work includes adapting the DeepSolar Model to the French dataset and trying to use SAM on the French dataset to fill in the remaining TBD values in the table.



(a) Raw image (b) CAM segmentation (c) True segmentation

Figure 1. Example result from DeepSolar baseline model trained on United States data and asked to predict segmentation masks of France test image data. True segmentation masks (rightmost), CAM activations from the original DeepSolar model (middle), and the original input image (leftmost).

References

[1] Gabriel Kasmi, Yves-Marie Saint-Drenan, David Trebosc, Raphaël Jolivet, Jonathan Leloux, Babacar Sarr, and Laurent Dubus. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. Scientific Data, 10(1):59, Jan 2023. 1, 2

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1, 2

[3] Dongyue Li and Hongyang R. Zhang. Improved regularization and robustness for fine-tuning in neural networks, 2021. 2

[4] Kevin Mayer, Zhecheng Wang, Marie-Louise Arlt, Dirk Neumann, and Ram Rajagopal. Deepsolar for germany: A deep learning framework for pv system mapping from aerial imagery. pages 1–6, 2020. 1

[5] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. CoRR, abs/2007.00644, 2020. 1, 2

[6] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation, 2018. 2

[7] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation, 2022. 1, 2

[8] Jiafan Yu, Zhecheng Wang, Arun Majumdar, and Ram Rajagopal. Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states. Joule, 2(12):2605–2617, 2018. 1, 2

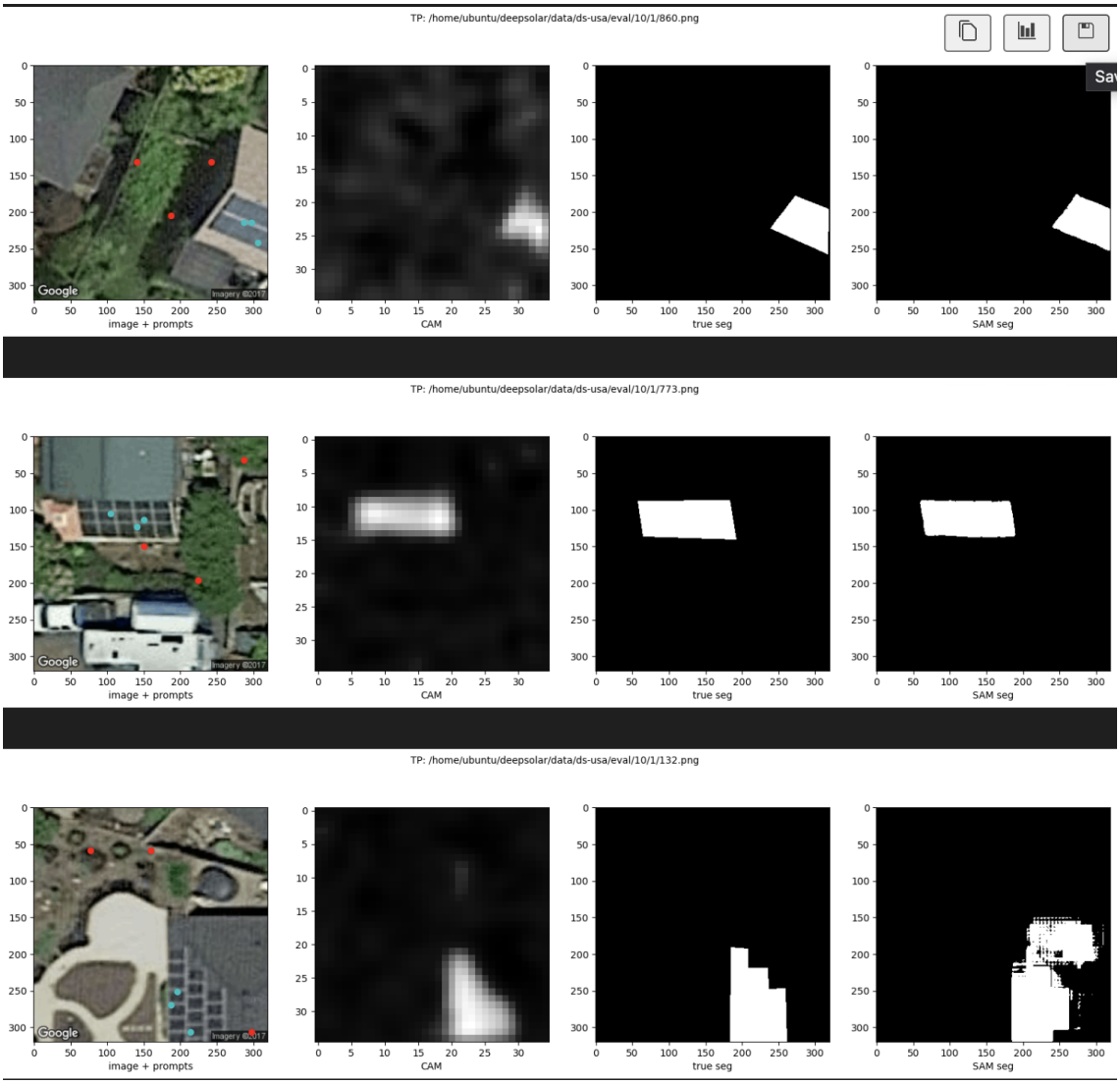


Figure 2. Preliminary segmentation of solar PV cells with SAM (rightmost column), compared to true segmentation masks (second from right), CAM activations from the original DeepSolar model (third from right), and the original input image (leftmost) with positive point prompts in cyan, and negative point prompts in red.