Exploring Robustness of Stanford's DeepSolar Model to Distribution Shifts

Kerrie Wu (kerriewu@stanford.edu)   Julian Cooper (jelc@stanford.edu)   Andrea van den Haak (vandenhaak@stanford.edu)

Introduction & Background

- **In many countries, the rollout of rooftop solar has been ad hoc and largely untracked.** Being able to estimate rooftop solar adoption allows utilities to predict power output reasonably well and make the adjustments to grid management and investment.[2]
- In December 2018, Stanford's Deepsolar team produced an aerial imagery dataset and model which together largely solved this challenge for the United States [1]. However, even in its current form, the **Deepsolar model is not robust to "distribution shifts"**.
- In January 2023, Kasmi et al. published a new dataset for residential areas in France. As feared, Stanford's Deepsolar performed poorly: **classification F1 score fell from 0.91 to 0.24** and **segmentation IOU fell from 0.51 to 0.06**.

Problem Statement

Our goal is to adapt DeepSolar's model to new domains and improve its overall robustness to domain shifts while using minimal domain-specific labeled data.

We explore two data-efficient techniques for improving the DeepSolar model's performance on the French dataset: **regularized fine-tuning** and **interpolation-based data augmentation**.

Datasets & Pre-Processing

United States (Deepsolar, 2018) [1] Heavily skewed, larger dataset covering all parts of the contiguous United States.

- 46,090 images with solar panels, 366,467 images without
- 3 x 320 x 320 input image size
- Industrial and residential regions

France (Kasmi et al., 2023) [2] Class balanced, smaller dataset covering primarily residential areas of France.

- 13,303 images with solar panels, 15504 images without
- 3 x 400 x 400 input image size
- Residential regions only

Image pre-processing pipeline: Significant work was required to wrangle the French dataset into a state usable with selected components of the existing Deepsolar training pipeline.

- Data augmentation transforms including random 90-degree rotations and horizontal flips
- Resize images to 299 × 299 to match the Inception3 model's input API
- Normalize mean and standard deviation of each channel
- Split dataset into test (5000 images), validation (1000 images), and training (remainder)
- Create class-balanced finetuning sets with 100, 500, 1000, and 5000 images per class

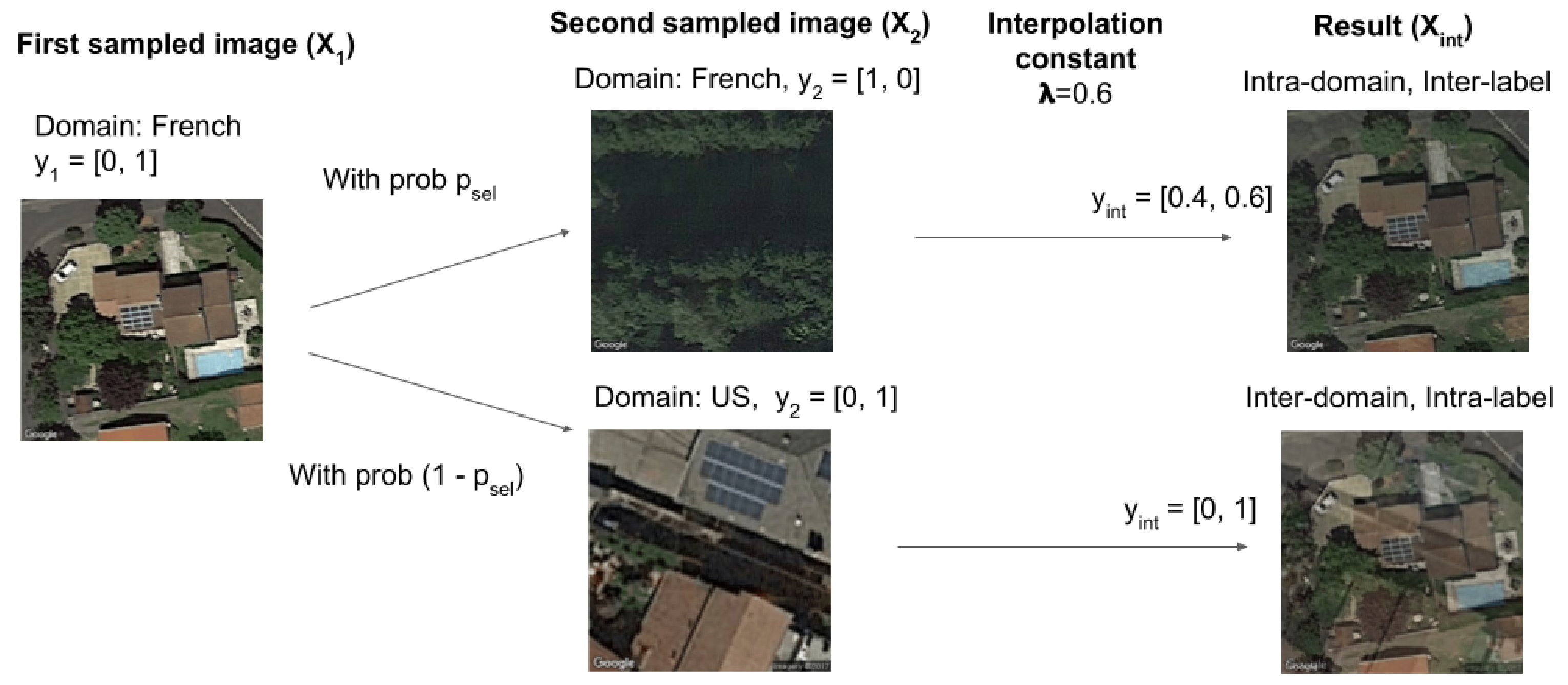
Method 1: Regularized Finetuning

L2 regularization: Use the Adam optimizer and tune weight decay and learning rate. Weight decay constant guarantees we do not take too large steps away from pretrained weights.

Adversarial regularization: When the input x is perturbed by a small amount, the output should not change much. To achieve this, Jiang et al. optimize loss $\mathcal{F}(\theta)$ using: $\min_\theta \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$, where $\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_{p \leq \epsilon}} l_s\left(f(\tilde{x}_i;\theta), f(x_i;\theta)\right)$.

Method 2: LISA Data Augmentation

LISA [3] is a data augmentation method that uses **linear interpolation between training examples** to improve model robustness to distribution shifts.


First sampled image (X_1)
Domain: French
y_1 = [0, 1]
With prob p_{sel}
With prob (1 - p_{sel})
Second sampled image (X_2)
Domain: French, y_2 = [1, 0]
Domain: US, y_2 = [0, 1]
Interpolation constant
λ=0.6
Result (X_{int})
Intra-domain, Inter-label
Inter-domain, Intra-label
y_{int} = [0.4, 0.6]
y_{int} = [0, 1]

1. Select a training example and label X_1, y_1 from the French finetuning dataset
2. Sample a λ value in the range [0, 1] from a $Beta(2,2)$ distribution
3. Select a second example and label X_2, y_2 for either inter-label or inter-domain interpolation
4. Resize both X_1 and X_2 to the input size for the model
5. Construct $X_{int} = \lambda X_1 + (1-\lambda)X_2$, and $y_{int} = \lambda y_1 + (1-\lambda)y_2$

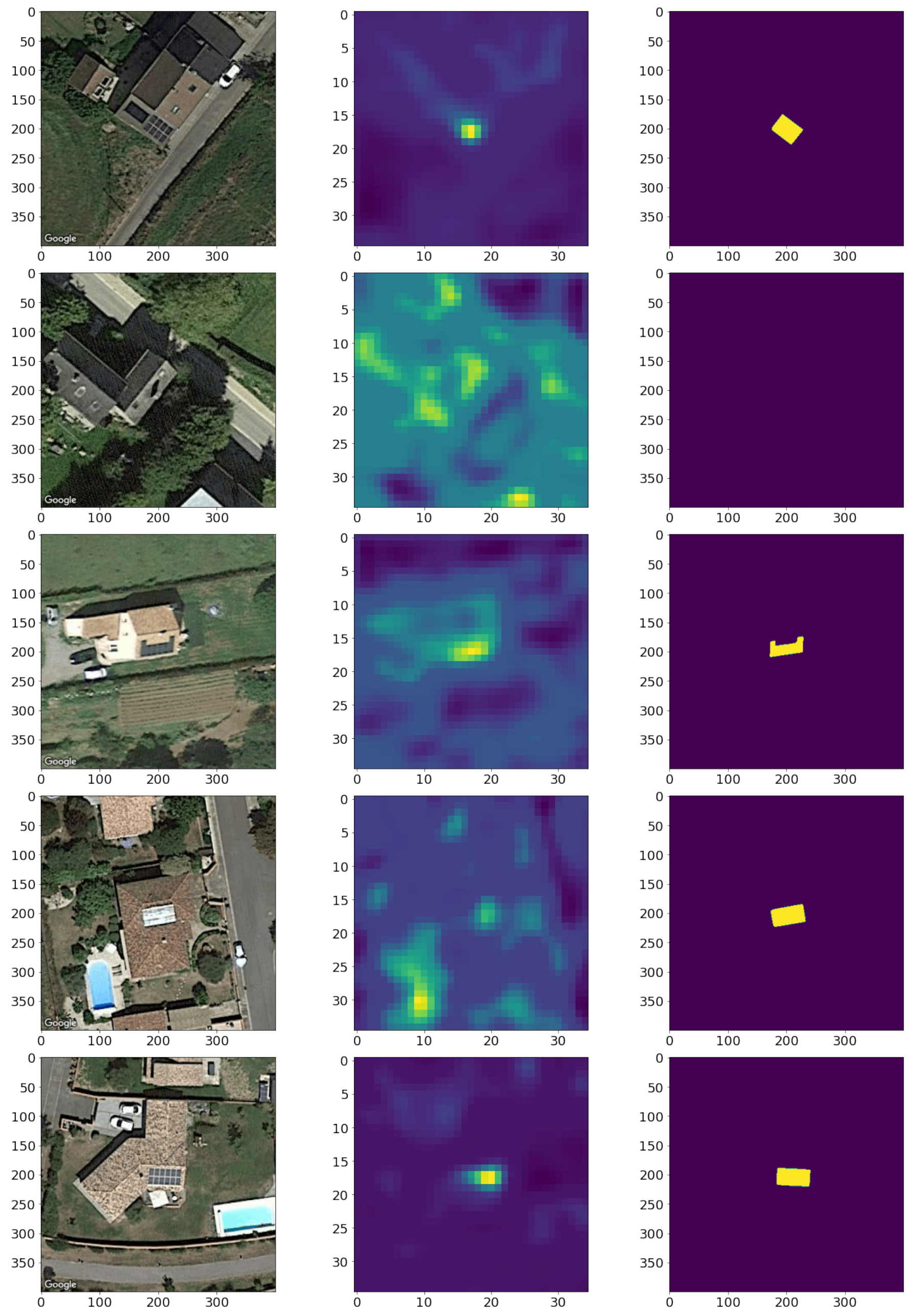
Experimental Results

Summary of our key experimental results, including classification and segmentation performance metrics for our finetuned model variants on the test French dataset.

			Classification				*Segmentation*	
Model	Eval Dataset	No. rows	Accuracy	F1 Score	Precision	Recall	Area Error	IOU
Deepsolar Baseline	USA	n/a	0.99	0.91	0.95	0.86	-0.08	0.51
Deepsolar Baseline	FR	n/a	0.57	0.24	0.96	0.14	-0.73	0.06
Regularized Finetuning	FR	100	0.87	0.87	0.86	0.89	0.10	0.23
Regularized Finetuning	FR	500	0.94	0.94	0.97	0.91	-0.24	0.48
Regularized Finetuning	FR	1000	0.97	0.96	0.96	0.97	0.10	0.52
Regularized Finetuning	FR	5000	0.98	0.98	0.97	0.99	-0.02	0.54
LISA Data Augmentation	FR	100	0.87	0.88	0.81	0.98	0.02	0.31
LISA Data Augmentation	FR	500	0.95	0.95	0.95	0.96	0.07	0.40
LISA Data Augmentation	FR	1000	0.97	0.97	0.96	0.98	-0.04	0.10
LISA Data Augmentation	FR	5000	0.97	0.97	0.96	0.98	-0.04	0.38

Discussion & Error Analysis

- **Image misclassification**: The most significant difference between the original Deepsolar baseline and our finetuned model variants (Fig, row 2). The baseline model achieved precision of 0.96 but recall of only 0.14. This means it was confident in the images it did select but misclassified many images that did in fact have solar panels. For those images, it predicted an empty segmentation!
- **Object confusion**: Initially, a large source of error was object confusion (Fig, row 4). As we add more data from the French dataset, which is more heavily weighted towards residential settings, our models improved at correctly identifying objects. For example, in row 5 we no longer confuse a swimming pool for a solar panel.
- **Diffuse prediction error**: Our last common source of error was placing high probabilities on area surrounding the solar panel (often rooftop). To address this, we tuned our CAM segmentation threshold (how confident we need to be for a pixel to be considered 1 vs 0) as an additional hyperparameter. This improved total area error but not IOU since our model will often put comparable weight on solar panels and other roof elements!



Example input images (left), CAM outputs (middle), and true segmentation masks (right). Rows: (1) good example of simple segmentation; (2) bad example of model misclassification; (3) bad example of overly diffuse prediction; (4) bad example of confusing objects; (5) good example of handling similar objects.

Conclusion and Future Work

- We are able to achieve **comparable classification and segmentation performance on a new domain** by finetuning the DeepSolar on only 500 per-class labeled samples from the target data distribution using either L2 regularization or LISA.
- **However, neither method improved model robustness** to distribution shifts in general. The finetuned models performed well on the French dataset, but poorly on the US dataset.
- Future work should include expand this investigation to include additional known distribution shift techniques such as Domain Adaptation and General Adversarial Networks.

Acknowledgements

We would like to thank Zhecheng Wang for providing us with initial guidance on directions for this project, introducing us to the DeepSolar Github repository, and providing us with the DeepSolar US dataset.

1. 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.
2. 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.
3. Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving outof-distribution robustness via selective augmentation, 2022