The effects of spatial resolution on rooftop photovoltaic detection in satellite imagery

Husnain Khurshid
National University of Sciences and Technology, Islamabad
husnainkhurshid360@gmail.com

September 23, 2026

Abstract

Rooftop photovoltaic detection is reported at ground sample distances spanning two orders of magnitude, but no study isolates resolution as a variable: comparisons change the sensor, the scene and the geography together, so no number from one paper is commensurable with another. We hold the scene fixed and vary only resolution, degrading 0.1 m aerial orthophotos of 645 rooftops through a point-spread-function sensor model to 0.3, 0.8, 2.5 and 10 m, and evaluating four detector families at every rung under identical spatially blocked cross-validation. We then acquire real Sentinel-2 over the same rooftops to test whether the simulation is faithful.

Four results follow. First, and most consequential for the field, random cross-validation inflates chip-level R^2 by a factor of 22 on this data (0.699 ± 0.047 random, 0.032 ± 0.440 spatial) because overlapping chips place the same rooftop in training and test; the evaluation protocol dominates every other design choice we measure. Second, 97.5% of arrays are smaller than one Sentinel-2 pixel, and three-band imagery at 10 m has no predictive power for panel fraction whether the three bands are simulated or real ($R^2 \approx 0$ in both cases). Third, spectral breadth substitutes for spatial resolution at 10 m: moving from three bands to twelve buys $+0.275\,R^2$, more than any change in resolution we measure. Fourth, spatial context becomes *more* valuable as resolution coarsens, not less — a small convolutional network matches per-pixel colour classifiers above 0.8 m but exceeds them by 0.05–0.06 F1 at 2.5 and 10 m.

We report detection floors as curves against array area rather than as a single threshold, alongside a matched chance baseline that the conventional per-array detection criterion fails to beat.

1 Introduction

Rooftop photovoltaic capacity is being installed faster than it is being recorded. Where registries exist they are incomplete — Germany's mandatory register implies that a majority of rooftop capacity sits below the size threshold at which open detection pipelines operate — and where they do not exist, satellite detection is the only practical inventory method. The question a planner therefore asks is not whether detection works, but at what imagery resolution it begins to work for the array sizes that dominate their building stock.

That question has no answer in the literature. Rooftop detection is reported at 0.1 m from aerial survey [Jiang et al., 2021], at 0.3 m from commercial basemaps [Yu et al., 2018, Lodhi et al., 2026], and at 10 m from Sentinel-2 [Kruitwagen et al., 2021], and every one of these studies uses a different sensor over a different geography with a different building stock. Performance differences between them confound resolution with everything else that changed. The consistent statement across the coarse-resolution work — that rooftop arrays are smaller than one pixel and are therefore missed — is asserted rather than quantified.

We measure it. Our design is a controlled degradation: one set of scenes, imaged once, resampled through a sensor model to five ground sample distances, with every other factor held constant. Resolution is then the only variable, and differences across rungs are attributable to it. The approach follows Shermeyer and Van Etten [2019], who used controlled resolution changes to study object detection in satellite imagery, and parallels the use of spatial validation in ecological mapping [Roberts et al., 2017, Ploton et al., 2020], from which we borrow the evaluation protocol that turns out to matter more than anything else we test.

Our contributions are:

1. A controlled resolution ladder for rooftop PV detection from 0.1 to 10 m in which the scene, annotation and evaluation are held fixed, together with the degraded dataset and code.
2. Quantification of spatial-autocorrelation leakage in this task: random cross-validation inflates chip-level R^2 by 22×. We argue this is a general hazard for PV detection datasets built from overlapping chips.
3. Detection floors reported as curves of detection probability against array area, per method and per resolution, with a matched chance baseline — and the demonstration that the conventional per-array detection criterion does not beat that baseline.

4. A direct comparison against real Sentinel-2 over the same rooftops, which both validates the degradation model and shows that spectral breadth substitutes for spatial resolution at 10 m.

2 Related work

Rooftop detection at high resolution. DeepSolar [Yu et al., 2018] established the task at sub-metre resolution over the United States. Subsequent work added installation geometry from 3D building models [Mayer et al., 2022] and crowdsourced annotation across providers [Kasmi et al., 2023]. Lodhi et al. [2026] report 122,596 installations across three Pakistani cities from 0.3 m basemap imagery. All of this work operates where an array spans hundreds of pixels.

Detection at Sentinel-2 scale. Kruitwagen et al. [2021] mapped global utility-scale photovoltaic assets from Sentinel-2 and SPOT, using a recurrent model over the image time series to filter false positives and date construction. Work at this scale consistently reports that rooftop-scale arrays fall below the detection limit; none quantifies where that limit lies as a function of array size.

Generalisation. Ranalli et al. [2024] report that diversifying training data does not improve the generalisability of neural PV detectors — a result without a stated mechanism. Our leakage finding supplies a candidate: if evaluation splits do not separate training and test spatially, they never measured generalisation in the first place, and data diversity cannot improve a quantity that was not being measured.

Distinction from the nearest prior work. Jiang et al. [2021], whose dataset we use, run *cross-application* experiments: train on one of their three datasets, test on another, optionally fine-tune. That design varies the dataset, and with it the sensor, the location, the building stock and the atmosphere simultaneously. It answers whether a model transfers between datasets. It cannot answer what resolution costs, because everything moves at once. Our design varies resolution on a fixed scene and nothing else. The two experiments are complementary and measure different quantities.

3 Data

3.1 Source imagery and annotation

We use the PV01 subset of Jiang et al. [2021]: 645 annotated 256 × 256 chips of 0.1 m aerial orthophoto over Jiangsu Province, China, stratified into three roof types

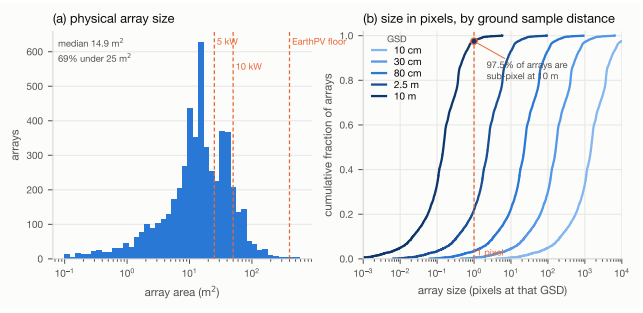


Figure 1: Array size in physical and sensor units. (a) The distribution of array areas, which no sensor choice changes: median 14.9 m², 69% below the 25 m² footprint of a 5 kW domestic system. (b) The same arrays expressed in pixels at each ground sample distance, as a cumulative distribution. At 10 m, 97.5% of arrays occupy less than a single pixel; at 2.5 m, 21.9% do. The paper's premise is visible here: at Sentinel-2 resolution this is a sub-pixel estimation problem, not a segmentation problem.

(flat concrete, 413 chips; brick, 138; steel tile, 94). Each chip covers 25.6 × 25.6 m. Imagery is three-band RGB; there is no near-infrared or shortwave channel, a constraint that shapes the experiment and which we return to in Section 6.4.

Connected-component analysis of the binary masks yields 5,238 distinct arrays of at least 4 pixels. Their size distribution (Figure 1a) is the premise of the paper: the median array is 14.9 m², 69.1% are smaller than 25 m² — roughly a 5 kW domestic system — and only 0.2% exceed 400 m². Essentially the entire dataset lies below the threshold at which open Sentinel-2 pipelines currently segment.

3.2 Simulation of sensor resolution

Resampling imagery with a generic interpolation kernel models neither the optics nor the detector of a coarser sensor: it either aliases, when no pre-filter is applied, or smooths by an amount unrelated to any real instrument. We therefore model the two stages explicitly. The scene is convolved with a Gaussian point spread function whose full width at half maximum equals the target ground sample distance, giving $\sigma = \mathrm{FWHM}/2\sqrt{2\ln 2}$ expressed in source pixels, and the result is then point-sampled at the target pitch. This corresponds to a well-sampled instrument with $Q \approx 1$.

Two alternatives are retained for ablation: a pure $s \times s$ box mean, which models detector integration without optics, and bilinear resampling, which models neither and is included because it is common practice.

From a 0.1 m source, every rung of our ladder is an exact integer scale factor ($\times 1, 3, 8, 25, 100$ for 0.1, 0.3, 0.8, 2.5 and 10 m), so no fractional resampling artefact is introduced anywhere. This property does not hold from

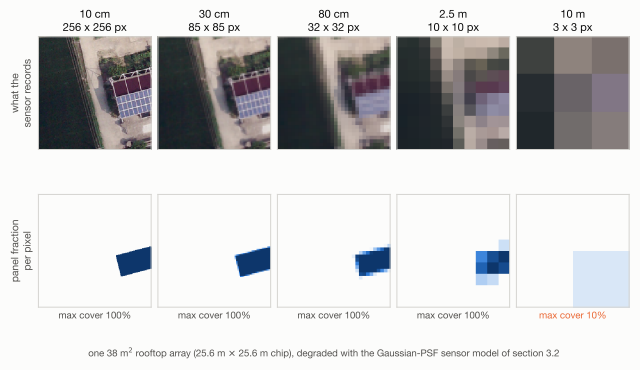


Figure 2: The resolution ladder applied to one 38 m^2 rooftop array. Top: the simulated sensor response at each rung, drawn with nearest-neighbour upsampling so the sampling grid remains visible. Bottom: the corresponding sub-pixel ground truth, the fraction of each target pixel covered by panel. Individual modules are countable at 0.1 m and the array is unambiguous at 0.8 m; by 2.5 m it is a coloured smudge, and at 10 m the entire chip is four pixels of which the most panel-rich contains 9.9% panel.



Figure 3: The degradation model against the sensor it imitates, for four rooftops spanning 5–30% panel cover. Row 2 is 0.1 m aerial imagery degraded to 10 m by the model of Section 3.2; row 3 is Sentinel-2 true colour over the same ground. The two are not photometrically identical — different instrument, date and atmosphere — and we do not claim they should be. What Section 6.4 shows is that they carry equivalent *information*: three bands at 10 m support no useful prediction from either source. Note also that neither row resolves anything a human could call a solar panel.

the 0.3 or 0.8 m sources in Jiang et al. [2021], which is why the ladder is built from PV01 alone and the coarser datasets are reserved for validation.

Labels are degraded separately and by area mean rather than by the sensor model: the fraction of a target pixel covered by panel is a property of the ground, not of the instrument observing it. This yields sub-pixel fractional ground truth at every rung, of which the binary label used for detection is a thresholding at 50% cover.

4 Detection methods

Because the source imagery is three-band, spectral indices requiring near-infrared or shortwave channels cannot be computed on the ladder, and pixel embeddings derived from Sentinel-1/2 time series are not defined for aerial frames. The ladder therefore compares methods that consume RGB, and Section 6.4 addresses the spectral question separately with real satellite data.

We evaluate four families, all predicting per-pixel panel presence:

Stratified prior. Predicts the training class distribution, ignoring the image. Any method that does not beat this has learned nothing at that rung.

Logistic regression on per-pixel colour features: the three channels, their sum, and two normalised colour ratios that separate dark bluish material from concrete, tile and vegetation. Features are standardised; Section 6 explains why this is not a cosmetic detail.

Gradient-boosted trees on the same features, capturing non-linear colour structure.

Convolutional network. Three 3×3 convolutional layers of width 32 with batch normalisation, then a 1×1 head: a receptive field of 7×7, with no pooling or downsampling. The absence of downsampling is deliberate. A U-Net cannot run on the 2×2 input the 10 m rung produces, and changing architecture between rungs would confound the model with the quantity being measured. This network runs unchanged from 256×256 to 2×2, so differences across rungs are attributable to resolution.

5 Evaluation protocol

5.1 Spatial blocking

The chips overlap. Median nearest-neighbour spacing between chip centres is 22.2 m against a 25.6 m chip, and 86.5% of chips lie within one chip-width of another. A randomly assigned split therefore places the same physical rooftop in both training and test partitions.

We instead grid the study area into 0.4 km blocks and assign whole blocks to folds, largest block first onto the currently smallest fold. This yields 53 blocks across five folds of exactly 129 chips each, with no rooftop straddling a split. Section 6.1 quantifies what this choice is worth.

Table 1: Chip-level panel-fraction regression at 0.1 m under two cross-validation schemes. Only the fold assignment differs.

Cross-validation	R^2	s.d. across folds
Random folds	+0.699	0.047
Spatial folds	+0.032	0.440

Table 2: Per-pixel detection F1 by method and ground sample distance, five-fold spatially blocked cross-validation. Best per column in bold.

Method	0.1 m	0.3 m	0.8 m	2.5 m	10 m
Prior	0.288	0.288	0.295	0.300	0.268
Logistic	0.635	0.672	0.674	0.593	0.500
Boosted	0.769	**0.784**	**0.755**	0.707	0.511
CNN	**0.786**	0.781	0.741	**0.768**	**0.561**

5.2 Detection criteria

Per-pixel scores do not answer the question a planner asks, which concerns arrays rather than pixels. Scoring per array requires a criterion, and the obvious one fails.

Under the conventional convention — an array counts as detected if *any* target pixel it overlaps is predicted positive — a random model flagging pixels at the trained model's own positive rate ($\approx 29\%$) achieves array recall of 0.999 at 0.1 m against the trained model's 0.995, and 0.804 against 0.821 at 2.5 m. Chance wins at three of five rungs. An array spans thousands of pixels at 0.1 m, so at least one positive prediction is near-certain; the criterion reports the positive rate rather than skill.

We therefore report two criteria, and a matched chance baseline for both. *Majority recall* requires most distinct target pixels an array overlaps to be predicted positive; chance falls to 0.001 at 0.1 m. It remains biased against large arrays, which are heterogeneous in shading and orientation and therefore harder to flag in majority. *Paired discrimination* scores each array by mean predicted probability over its footprint against a matched control footprint of equal pixel count drawn from panel-free pixels in the same chip; chance is exactly 0.5 by construction and size bias is impossible. Its limitation is that at 10 m only 611 of 5,238 arrays admit a control footprint, and that subset skews toward emptier chips.

Neither criterion alone is honest, so we report both.

6 Results

6.1 Spatial leakage dominates

Table 1 reports the same model, features and data under the two fold assignments.

Random assignment inflates R^2 by a factor of 22. It also makes the result appear stable: the standard deviation across random folds is 0.047, an order of magnitude tighter than the 0.440 obtained spatially. Per-fold spatial scores are $-0.725, +0.010, +0.079, +0.145, +0.649$ — the model does not transfer between neighbourhoods at all, and random splitting conceals this completely.

We emphasise that the magnitude is specific to this chip geometry. The direction should hold for any dataset of overlapping or spatially clustered chips, which describes most public PV detection datasets, but the factor of 22 should not be quoted as universal.

6.2 Performance against resolution

Table 2 and Figure 4 give the central measurement. Three observations.

Performance is approximately flat from 0.1 to 0.8 m and falls sharply between 2.5 and 10 m; the degradation is a cliff rather than a gradient. For the boosted baseline the fall is $0.707 \to 0.511$.

The peak is at 0.3 m rather than 0.1 m for both per-pixel methods. We suggest, without having isolated it, that within-panel texture — module frames, cell structure, specular glint — adds variance to a per-pixel colour classifier that 0.3 m averages away. Testing this would require a texture ablation we have not run.

Among per-pixel methods, the advantage of a stronger estimator collapses as resolution coarsens: the boosted model exceeds logistic regression by 0.133 F1 at 0.1 m and by 0.010 at 10 m. Spatial context behaves oppositely. The CNN is indistinguishable from the boosted baseline between 0.1 and 0.8 m ($+0.018$, -0.004, -0.015) and exceeds it at the two coarsest rungs ($+0.061$ at 2.5 m, $+0.050$ at 10 m). We had predicted the reverse, on the reasoning that a 2×2 chip offers no context to exploit. The measurement contradicts this. A plausible account is that at fine resolution a panel pixel is unambiguous and colour saturates the available signal, whereas at coarse resolution pixels are spectral mixtures and neighbourhood structure is what disambiguates them; context matters more as per-pixel information degrades. We state this as a hypothesis, not a demonstrated mechanism.

6.3 Detection floors

Figure 6 reports detection probability against array area. Under paired discrimination, overall scores are 0.995, 0.988, 0.979, 0.957 and 0.814 across the ladder, and the curves are monotone in array area as physical reasoning requires. Under majority recall the corresponding figures are 0.915, 0.871, 0.768, 0.650 and 0.365, against chance baselines of 0.001, 0.011, 0.045, 0.238 and 0.271.

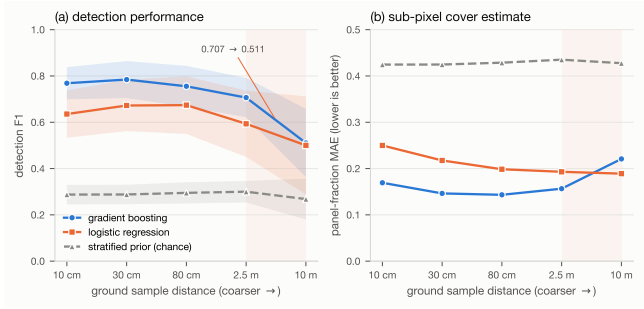


Figure 4: Detection performance and sub-pixel cover error against ground sample distance, with the axis running fine to coarse. Bands are ±1 standard deviation across the five spatial folds; the dashed line is the stratified prior. Performance is flat to 0.8 m and falls steeply between 2.5 and 10 m. Note the width of the bands: fold-to-fold variation is comparable to the differences between methods, which is itself a consequence of the spatial blocking.

The two criteria agree on the substantive point. At 10 m the signal is present — paired discrimination of 0.814 is far from the 0.5 of a coin — but operationally weak: majority recall of 0.365 sits only 0.094 above its matched chance floor, and no array size in this dataset reaches 80% recall at that resolution. Detectable in a paired comparison, and not reliable enough for unassisted inventory, are different statements, and only the second bears on whether a map can be made.

6.4 Real Sentinel-2 over the same rooftops

The ladder simulates a coarse sensor from fine imagery, which leaves two questions open: whether the simulation is faithful, and whether the three-band constraint of the source rather than resolution is what limits performance at 10 m. We address both by acquiring real Sentinel-2 Level-2A over the same 645 rooftops, via a public cloud-optimised archive requiring no credentials, mosaicking two tiles from near-cloud-free November 2023 scenes and resampling all twelve reflectance bands to 10 m. Reflectance is the archive digital number divided by 10^4 with no further offset: the products carry the baseline-04.00 radiometric offset already applied, and subtracting it a second time — an easy error, and one we made — drives 90% of blue-band pixels negative.

Panel-rich chips are darker than panel-poor ones in every band, and the separation is concentrated in the red edge and near infrared ($d' = -0.90$ to -1.19 for red-edge3, nir08 and nir09) rather than the visible (-0.15 to -0.25); Figure 7 shows both the spectra and the per-band separability. This is the expected signature of a surface engineered to absorb: the three visible bands average $|d'| = 0.21$ against 0.96 for the red edge and near infrared.

Table 3: Chip-level panel-fraction regression, identical spatial folds throughout. Fold standard deviations are 0.25–0.64; only the 12-band result exceeds its own spread.

Representation	R^2	MAE
Simulated 10 m, RGB (3 bands)	−0.002	0.171
Sentinel-2, visible (3 bands)	−0.049	0.182
Sentinel-2, red edge + NIR (6)	+0.052	0.163
Sentinel-2, all 12 bands	**+0.271**	**0.145**

Table 3 answers both questions. Simulated 10 m RGB and real Sentinel-2 visible perform equivalently and both at chance (−0.002 and −0.044), which is evidence that the degradation model is faithful at the level the comparison can resolve. And the limitation at 10 m is spectral rather than spatial: three bands carry no usable signal regardless of provenance, while twelve bands recover $R^2 = 0.271$. Spectral breadth buys more here than any resolution change we measure.

7 Discussion

What a planner should take from this. For arrays of domestic scale, sub-metre imagery is required for reliable per-array detection; at 0.8 m the boosted baseline still recovers 0.768 majority recall, and by 10 m that falls to 0.365 against a 0.271 floor. Where only 10 m imagery is available, the useful investments are spectral breadth and spatial context, not a better per-pixel classifier.

Evaluation. We consider the leakage result the most transferable finding here, and the most uncomfortable. It is also the cheapest to check: any group with an existing PV detection pipeline can re-run their own evaluation under spatially blocked folds in an afternoon. We would encourage this, and note that the result offers a mechanism for Ranalli et al. [2024]'s finding that data diversity does not improve generalisability.

Limitations. The evidence base is one dataset over a single 7 × 17 km area of Jiangsu, on a single Sentinel-2 date, comprising 645 overlapping chips of one building stock. The factor of 22 is specific to this chip geometry. Fold-to-fold variation is large enough to swamp several of the method differences we report, and we have deliberately not claimed significance for them; the differences we do emphasise — the leakage factor, the three-versus-twelve-band gap — are those that exceed their fold spread. Detection is evaluated against annotation from a single source, so annotator error is not characterised. Degradation is simulated, and while Section 6.4 validates it against a real sensor at 10 m in the

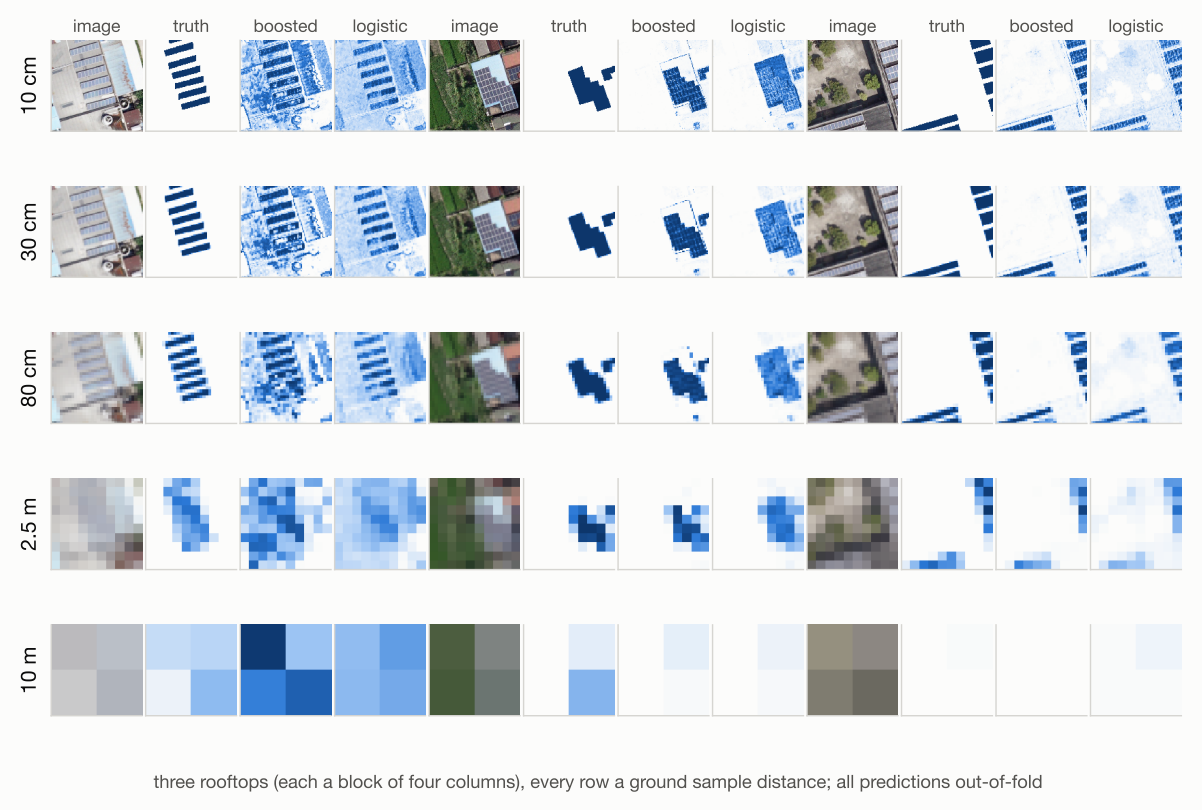


Figure 5: Out-of-fold predictions across the ladder for three rooftops, each occupying a block of four columns. Predictions track the annotation closely from 0.1 to 0.8 m, though the boosted model speckles false positives across bright roof material at 0.1 m. By 2.5 m the array is located but its shape is gone, and at 10 m two of the three rooftops produce almost no response at all. No model shown here ever saw the spatial block these chips belong to during training.

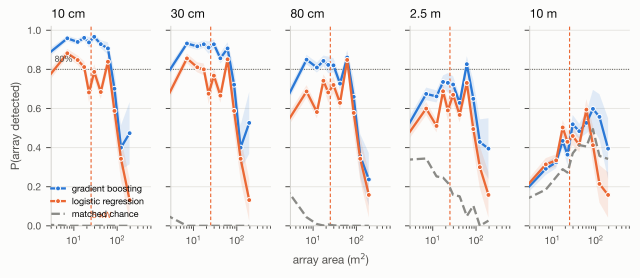


Figure 6: Probability of detecting an array as a function of its physical area, one panel per ground sample distance. The dashed grey line is a matched chance model flagging pixels at the trained model's own positive rate; at 10 m it reaches 0.271, so the apparent recall there is largely not skill. Shown for the majority criterion, which is biased against large arrays — see Section 4.2.

three-band case, we have not validated the intermediate rungs against real imagery at those resolutions.

Most importantly, the motivating application — rooftop PV inventory in countries without registries — concerns building stock quite unlike Jiangsu's. We make no claim that these floors transfer to other roof materials, array mounting conventions or atmospheric conditions, and regard establishing that as the natural next step.

8 Conclusion

Holding the scene fixed and varying only sensor resolution shows that rooftop PV detection degrades as a cliff between 2.5 and 10 m rather than gradually, and that 97.5% of arrays are sub-pixel at Sentinel-2 resolution. But the larger effect in this study is not resolution at all. Random cross-validation on overlapping chips inflates measured performance by a factor of 22 while simultaneously making it appear more stable, and spectral breadth at a fixed 10 m buys more than any resolution change we measured. The detection floor for rooftop photovoltaics is real, and it sits somewhat lower than the coarse-resolution literature assumes — but reaching it depends more on how a study evaluates and how many bands it reads than on how fine its pixels are.

Data and code. The degradation pipeline, ladder dataset and evaluation code are available at the project

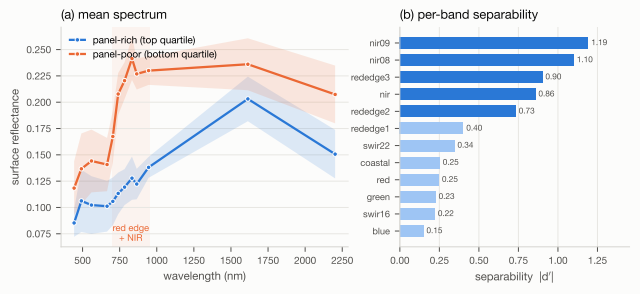


Figure 7: Why band choice dominates at 10 m. (a) Mean Sentinel-2 reflectance for chips in the top and bottom quartiles of panel cover, with 95% intervals. Panel-poor rooftops show the vegetation red edge and a strong near-infrared plateau; panel-rich chips suppress both. (b) Separability per band. The three visible bands average $|d'| = 0.21$; the red edge and near infrared average 0.96. A three-band visible sensor is reading the part of the spectrum where the two classes differ least.

repository. Source imagery is from Jiang et al. [2021] under CC-BY-4.0; Sentinel-2 is Copernicus data.

References

Hou Jiang, Ling Yao, Ning Lu, Jun Qin, Tang Liu, Yujun Liu, and Chenghu Zhou. Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery. *Earth System Science Data*, 13(11):5389–5401, 2021. doi: 10.5194/essd-13-5389-2021.

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*, 10:59, 2023.

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. doi: 10.1038/s41586-021-03957-7.

Muhammad Kamran Lodhi, Yumin Tan, Yang Li, Faisal Mumtaz, and Shahid Naeem. A weighted ensemble DL framework for granular RPV inventory and energy yield analysis in Pakistan. *Remote Sensing Applications: Society and Environment*, 42:102032, 2026.

Kevin Mayer, Benjamin Rausch, Marie-Louise Arlt, Gunther Gust, Zhecheng Wang, Dirk Neumann, and Ram Rajagopal. 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310: 118469, 2022.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11:4540, 2020.

Joseph Ranalli, Matthias Zech, and Hans Peter Tetens. Diversifying training data does not improve generalizability of neural network models for PV identification. *Journal of Renewable and Sustainable Energy*, 2024.

David R. Roberts, Volker Bahn, Simone Ciuti, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929, 2017.

Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

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.