Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Transfer Learning: Evaluating the Trustworthiness of Pretrained GeoAI Models for Rooftop Solar PV Mapping

Tahsin Hossain
Queensland University of Technology
Tan Yigitcanlar
tan.yigitcanlar@qut.edu.au

Queensland University of Technology
Niklas Tilly
Queensland University of Technology
Filip Biljecki
National University of Singapore

Research Article

Keywords: geospatial artificial intelligence, GeoAI, rooftop solar photovoltaics, solar panel detection, transfer learning, deep learning, GeoAI-based analysis

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10465298/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Geovisualization and Spatial Analysis on October 29th, 2026. See the published version at https://doi.org/10.1007/s41651-026-00293-0.

Beyond Transfer Learning: Evaluating the Trustworthiness of Pretrained GeoAI Models for Rooftop Solar PV Mapping

Abstract: The increasing availability of pretrained Geospatial Artificial Intelligence (GeoAI) models has accelerated the operational adoption of deep learning within geographic information systems. However, models developed in one geographic region frequently experience performance degradation when deployed elsewhere because of geographic domain shift. Although transfer learning is widely used to improve cross-regional model performance, it remains unclear whether improved predictive accuracy alone provides sufficient evidence for trustworthy deployment. This study investigates this question using rooftop solar photovoltaic mapping in Brisbane, Australia, as an operational case study. Esri's pretrained US Solar Panel Detection model was locally calibrated through transfer learning and evaluated using both development and geographically independent test datasets. Local calibration substantially improved detection performance, increasing recall from 38.5% to 84.1% on the development dataset while maintaining strong performance on the independent test dataset, demonstrating effective adaptation to geographic domain shift. However, independent evaluation and qualitative error analysis also revealed persistent false positives, missed detections, and geometric segmentation errors, indicating that residual uncertainty remained following transfer learning. These findings demonstrate that although transfer learning substantially improves robustness, reliability, and model generalisation, predictive performance alone is insufficient to establish trustworthy deployment. By integrating complementary evidence relating to robustness, reliability, generalisation, transparency, and residual uncertainty, this study provides an operational approach for evaluating the trustworthiness of pretrained GeoAI models prior to deployment in geographically distinct environments.

Keywords: geospatial artificial intelligence; GeoAI; rooftop solar photovoltaics; solar panel detection; transfer learning; deep learning; GeoAI-based analysis

1. Introduction

The rapid advancement of Geospatial Artificial Intelligence (GeoAI) is transforming the way earth observation data are analysed and translated into actionable information for environmental monitoring, urban planning, disaster management, infrastructure management, and climate resilience (Jiao & Sun, 2018; Choi, 2023; Hossain et al., 2025). By integrating artificial intelligence with Geographic Information Systems (GIS) and remote sensing, GeoAI enables the automated extraction, interpretation, and analysis of increasingly large volumes of high-resolution geospatial data. Recent advances in deep learning, together with the growing availability of pretrained models embedded within operational GIS platforms, have substantially lowered the technical barriers to applying artificial intelligence in geospatial workflows (Alhawsawi et al., 2025). Rather than developing models from scratch, practitioners can now deploy pretrained GeoAI models for a wide range of mapping tasks, accelerating operational decision-making across government, industry, and research (Hu et al., 2024; Epasto, 2025; Shi et al., 2025).

However, the widespread availability of pretrained GeoAI models has also shifted an important challenge from model development to model deployment (Choi, 2023; Esri, 2025a; X. Zhang et al., 2026). Although pretrained models offer considerable practical advantages, they are increasingly transferred to geographic environments with different landscape characteristics, urban morphology, environmental conditions, and image acquisition properties, where predictive performance achieved during model development may not be maintained (Wu & Biljecki, 2021; Haynes et al., 2024). As GeoAI increasingly supports infrastructure investment, environmental management, disaster response, and public policy, ensuring that pretrained models can be deployed robustly, reliably, and transparently has become an important scientific and operational challenge (Wenwen Li et al., 2023; Shaamala et al., 2025; Viswambharan & Singh, 2025; X. Zhang et al., 2026).

One of the principal threats to trustworthy GeoAI deployment is geographic domain shift (Manso-Callejo et al., 2023; Jagatheesaperumal et al., 2024; Huang et al., 2026). Deep learning models frequently learn region-specific visual representations that reflect local building architecture, roofing materials, vegetation, illumination conditions, image characteristics, and other place-dependent features (Wu & Biljecki, 2021; Boulmay, 2022; Haynes et al., 2024; Kasmi et al., 2025). When transferred to geographically distinct environments, these learned representations may no longer adequately characterise the target landscape, resulting in systematic false positives, missed detections, spatial inconsistencies, and reduced model reliability. Such behaviour introduces uncertainty into downstream spatial analyses and may ultimately influence planning and policy decisions derived from GeoAI outputs. Recognising these limitations, developers of operational GeoAI models, including Esri, recommend local calibration or retraining before deployment in new geographic environments (Viswambharan & Singh, 2020; Singh & Viswambharan, 2021; Szukalski, 2024). However, it remains unclear

whether improved predictive performance following transfer learning is, by itself, sufficient to justify trustworthy deployment in previously unseen geographic contexts.

In this study, trustworthy deployment is viewed as extending beyond predictive accuracy alone. A GeoAI model should also demonstrate robustness to geographic domain shift, reliable performance on independent datasets, transparent understanding of model behaviour and failure modes, and explicit acknowledgement of residual uncertainty. Together, these characteristics provide stronger evidence that a model can support operational decision-making under geographically distinct conditions.

Rooftop Solar Photovoltaic (PV) mapping provides an appropriate real-world application through which to investigate these challenges (Yu et al., 2018; Wu & Biljecki, 2021; Yasmeen et al., 2022). Accurate rooftop PV inventories support distributed energy planning, electricity network management, renewable energy policy, and climate resilience by enabling more reliable estimation of installed generation capacity and its spatial distribution (Kaya et al., 2019; Arribas et al., 2021; Yasmeen et al., 2022; Manso-Callejo et al., 2023). They also provide an independent means of monitoring renewable energy adoption where administrative records are incomplete, inconsistent, or unavailable (Peronato et al., 2018; Taminiau & Byrne, 2020; Long et al., 2025). Because these inventories increasingly inform infrastructure investment and public policy, systematic detection errors may propagate into broader planning decisions, making rooftop PV mapping an appropriate operational case study for evaluating the deployment of pretrained GeoAI models (Tayal, 2017; Parsaeifar et al., 2025; Sadeq et al., 2025).

Australia provides a particularly relevant context for this assessment. By 2024, approximately one in three Australian households had installed rooftop PV systems, representing one of the highest rates of residential solar adoption globally (Narayanan et al., 2021; Energy AU, 2024). This rapid growth has created increasing demand for timely and spatially explicit rooftop PV inventories to support network planning, hosting capacity assessments, infrastructure investment, and broader urban sustainability initiatives (Zander, 2020; Best & Chareunsy, 2022; Best et al., 2023; Ghahremani et al., 2025). Although administrative datasets maintained by organisations such as the Clean Energy Regulator provide valuable information on installation numbers and capacity, they generally lack the building-level spatial precision required for operational analysis. Recent advances in GeoAI have substantially improved automated rooftop PV detection, and commercial GIS platforms now provide pretrained object detection models that can be readily integrated into operational geospatial workflows (Boulmay, 2022; Ghahremani et al., 2025; Parsaeifar et al., 2025). Among the most widely accessible examples is Esri's pretrained US Solar Panel Detection model, which has also served as the foundation for regional adaptations such as the New Zealand Solar Panel Detection model distributed through ArcGIS Living Atlas (Esri, 2023, 2025a). These developments illustrate the growing maturity of operational GeoAI while simultaneously raising an important practical question: can pretrained GeoAI models be trusted when deployed beyond the geographic environments for which they were originally developed?

Although transfer learning is widely recognised as an effective strategy for adapting pretrained models to new datasets, existing GeoAI research has focused predominantly on improving predictive performance following local calibration (Weiss et al., 2016; Boulmay, 2022; Franklyn, 2023; Nguyen et al., 2025). By comparison, the trustworthy AI literature emphasises broader considerations such as robustness, reliability, transparency, uncertainty, and accountability for operational deployment (OECD, 2019; NIST, 2023; Ghamisi et al., 2025). However, relatively few studies have explicitly connected these two perspectives by examining whether improvements achieved through transfer learning also provide sufficient evidence for trustworthy deployment under geographic domain shift. Consequently, an important gap remains between improving model performance and establishing confidence in operational model deployment.

To address this gap, this study investigates the deployment of a pretrained GeoAI model through the case of rooftop solar PV mapping in Brisbane, Australia. Specifically, Esri's pretrained US Solar Panel Detection model was locally calibrated using transfer learning and subsequently evaluated using both a development dataset and an independent test dataset excluded entirely from model development. The study uses an operational GeoAI workflow to examine how geographic domain shift influences model behaviour and whether transfer learning alone provides a sufficient basis for trustworthy deployment in a geographically distinct urban environment. Accordingly, the study addresses the following objectives:

- To evaluate how geographic domain shift affects the operational performance of a pretrained GeoAI rooftop PV detection model when transferred from the United States to Brisbane, Australia;
- To evaluate whether local calibration through transfer learning provides a sufficient basis for trustworthy deployment by comparing model performance on both development and independent evaluation datasets; and
- To examine the residual errors, uncertainties, and operational implications that remain after local calibration, thereby informing the responsible deployment of pretrained GeoAI models.

The principal contribution of this study is not the development of a new rooftop PV detection algorithm or the proposal of a new trustworthiness framework. Instead, it provides an empirical assessment of whether transfer learning alone provides sufficient evidence for the trustworthy deployment of pretrained GeoAI models under geographic domain shift. Using rooftop solar PV mapping as an operational case study, the research makes three contributions. First, it empirically evaluates the extent to which transfer learning improves the operational deployment of a pretrained GeoAI model when transferred to a geographically distinct environment. Second, it demonstrates how established trustworthy AI principles can be operationalised through complementary evidence relating to robustness, reliability, generalisation, transparency, and residual uncertainty within a practical GeoAI workflow. Third, it presents an evidence-based operational perspective that extends beyond conventional accuracy assessment and may assist researchers and practitioners in evaluating pretrained GeoAI models prior to operational deployment.

The remainder of this paper is organised as follows. Section 2 reviews the literature on operational GeoAI, geographic domain shift, transfer learning, and trustworthy artificial intelligence, from which an operational perspective for evaluating pretrained GeoAI deployment is established. Section 3 describes the study area, dataset preparation, transfer learning workflow, and evaluation methodology. Section 4 presents the empirical evaluation across the operational trustworthiness dimensions, including robustness, reliability, generalisation, transparency, and residual uncertainty. Section 5 discusses the broader implications of the findings for the responsible deployment of pretrained GeoAI models. Finally, Section 6 concludes the paper by summarising the principal findings, outlining practical recommendations, and identifying directions for future research.

2. Literature Background

2.1. Operational GeoAI, Geographic Domain Shift, and Transfer Learning

Recent advances in deep learning have fundamentally transformed the analysis of earth observation data. Convolutional Neural Networks (CNNs), instance segmentation models, and, more recently, transformer-based architectures have demonstrated remarkable capability in extracting complex geographic features from high-resolution aerial and satellite imagery (Yu et al., 2018; Chen et al., 2023; García et al., 2024; Ahmed et al., 2025). These developments have accelerated the adoption of GeoAI across a wide range of applications, including land-cover classification, building extraction, infrastructure monitoring, disaster assessment, environmental management, and rooftop solar PV detection (Best et al., 2023; Clark & Pacifici, 2023; Ghahremani et al., 2025). At the same time, the growing availability of pretrained deep learning models embedded within operational geospatial platforms has fundamentally changed the way GeoAI is applied. Rather than requiring practitioners to develop complex neural networks from scratch, modern GIS environments such as ArcGIS Pro enable pretrained models to be deployed and locally adapted within established geospatial workflows (Boulmay, 2022; Esri, 2025a). ArcGIS Living Atlas, for example, provides pretrained models for a range of remote sensing applications, illustrating a broader transition from research-oriented GeoAI towards operational AI-supported spatial analysis.

Despite these advances, pretrained GeoAI models frequently experience performance degradation when deployed outside the geographic environments on which they were originally trained. This phenomenon, commonly referred to as geographic domain shift, arises because deep learning models learn visual representations that are closely associated with the characteristics of their training data (Wu & Biljecki, 2021). Variations in building architecture, rooftop materials, vegetation, illumination conditions, sensor characteristics, image acquisition settings, and urban morphology can substantially alter the appearance of geographic features across regions, leading to systematic false positives, missed detections, and inconsistent model behaviour (Boulmay, 2022; Haynes et al., 2024; Coglan et al., 2025; Kasmi et al., 2025). Recent studies further suggest that such performance degradation reflects not only technical limitations but also embedded geographic biases, whereby models implicitly learn place-specific characteristics that do not generalise uniformly across different cities or countries (Hou et al., 2025). Consequently, understanding and mitigating geographic domain shift has become an increasingly important challenge for operational GeoAI deployment.

Transfer learning has emerged as one of the most widely adopted strategies for addressing geographic domain shift. Rather than developing entirely new models, transfer learning adapts pretrained feature representations using comparatively small quantities of locally annotated data, thereby reducing computational requirements, annotation effort, and model development time (Pan & Yang, 2010; Weiss et al., 2016). Numerous studies have demonstrated that transfer learning can substantially improve object detection performance across remote sensing applications, including rooftop PV mapping (Boulmay, 2022; Franklyn, 2023; Nguyen et al., 2025). Commercial GIS platforms have further accelerated this approach by providing accessible workflows

that enable pretrained models to be fine-tuned within operational environments, making local calibration increasingly feasible for practitioners without requiring advanced machine learning expertise.

However, improved predictive performance following transfer learning should not automatically be interpreted as evidence of trustworthy deployment. A model may achieve higher precision or recall after local calibration while still exhibiting residual geographic bias, inconsistent performance on previously unseen data, or persistent uncertainty in complex urban environments. Furthermore, many transfer learning studies evaluate performance using datasets that contribute to model development, providing only limited evidence regarding operational generalisation. From a responsible GeoAI perspective, evaluating pretrained models therefore requires broader evidence than predictive accuracy alone, including robustness to geographic domain shift, reliable performance on independent datasets, transparent understanding of model behaviour, and explicit acknowledgement of residual uncertainty. These considerations motivate the operational trustworthiness perspective adopted in this study.

2.2. Towards Trustworthy Deployment of Pretrained GeoAI Models

The increasing operational adoption of artificial intelligence has shifted attention beyond predictive performance towards broader questions of trustworthiness and responsible deployment. International frameworks, including the OECD Principles on Artificial Intelligence (OECD, 2019), the NIST Artificial Intelligence Risk Management Framework (NIST, 2023), the ISO/IEC 23894 standard for AI risk management (ISO, 2023), and more recently the European Union Artificial Intelligence Act (European Union, 2024), emphasise that trustworthy AI should exhibit characteristics such as robustness, reliability, transparency, accountability, and appropriate management of uncertainty. Although these frameworks were developed for artificial intelligence more broadly, their underlying principles are increasingly recognised as relevant to geospatial applications, where AI-derived outputs frequently support planning decisions with long-term environmental, economic, and societal consequences.

Within GeoAI, trustworthiness presents additional challenges because model performance is inherently influenced by geographic context (Epasto, 2025; Shi et al., 2025; Andorful et al., 2026). Unlike many conventional computer vision applications, Earth observation models are routinely transferred across locations that differ in landscape characteristics, built environments, environmental conditions, and image acquisition properties. Consequently, evaluation based solely on benchmark accuracy metrics may provide an incomplete assessment of operational performance. Recent studies have therefore highlighted the need to complement predictive accuracy with broader assessment of model robustness, geographic bias, uncertainty, transparency, and generalisation when deploying GeoAI models in real-world settings (Ghamisi et al., 2025; Hou et al., 2025). From this perspective, trustworthy deployment is not determined by how accurately a model performs on a single evaluation dataset, but by whether users can have justified confidence in its behaviour when applied to previously unseen geographic environments.

Building upon these principles, this study adopts an operational trustworthiness perspective for evaluating pretrained GeoAI deployment. Rather than proposing a new theoretical framework, we adapt established trustworthy AI principles from the broader AI governance literature to the operational evaluation of pretrained GeoAI models under geographic domain shift. Figure 1 illustrates the conceptual assessment framework adopted in this study. The framework conceptualises trustworthy deployment as a sequential process beginning with a pretrained GeoAI model developed in one geographic context, followed by deployment into a new environment where geographic domain shift may occur. Transfer learning provides a mechanism for local calibration; however, trustworthy deployment is not assumed following calibration alone. Instead, deployment should be supported through complementary evaluation of robustness, reliability, generalisation, transparency, and residual uncertainty before operational adoption.

Figure 1 illustrates how trustworthy deployment extends beyond model calibration. A pretrained GeoAI model first encounters geographic domain shift when transferred to a new operational environment. Although transfer learning can improve predictive performance through local calibration, responsible deployment requires additional evidence regarding model robustness, reliability on previously unseen data, generalisation beyond calibration datasets, transparency of model behaviour, and explicit acknowledgement of remaining uncertainty. Collectively, these dimensions provide the conceptual basis through which the empirical findings of this study are interpreted.

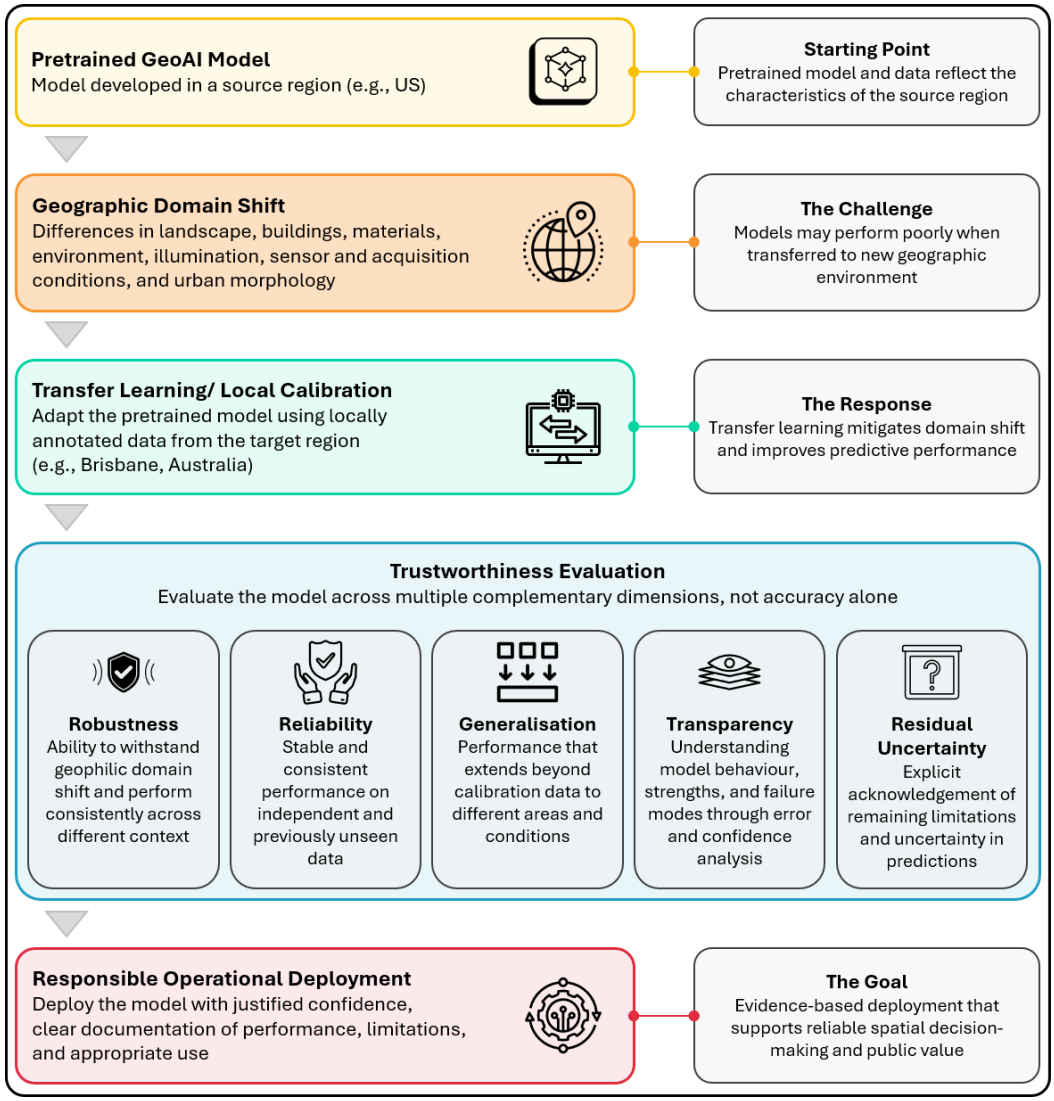


Figure 1: Conceptual trustworthiness assessment framework adopted in this study for evaluating pretrained GeoAI models under geographic domain shift

Table 1 operationalises the conceptual assessment framework presented in Figure 1 by linking each trustworthiness dimension to measurable evidence generated within the Brisbane case study. Rather than treating trustworthiness as an abstract principle, each dimension is evaluated using quantitative model performance, independent validation, qualitative error analysis, and practical deployment considerations.

Table 1: Operational trustworthiness dimensions evaluated in this study

Trustworthiness dimension	Operational interpretation	Evidence used in the study	References
Robustness	Ability to maintain performance under geographic domain shift	Cross-regional deployment from the United States to Brisbane	(Jagatheesaperumal et al., 2024; Kasmi et al., 2025; Lu et al., 2025; Huang et al., 2026)
Reliability	Stable performance on previously unseen data	Independent 10-grid test dataset	(Lu et al., 2025; Shi et al., 2025)
Generalisation	Consistent performance beyond calibration data	Comparison between development and independent datasets	(Manso-Callejo et al., 2023; Waqas & Humphries, 2024; Wu et al., 2024)
Transparency	Understanding why the model succeeds and fails	Confidence score analysis, false positives, false negatives, geometric error assessment	(Wenda Li et al., 2023; Andorful et al., 2026; Huang et al., 2026)

Residual uncertainty	Explicit acknowledgement of remaining model limitations	IoU sensitivity analysis and segmentation quality assessment	(Hu et al., 2024; W. Zhang et al., 2026)
Responsible deployment	Evidence supporting operational adoption	Practical recommendations derived from empirical findings	(Wenda Li et al., 2023; Shi et al., 2025)

Collectively, the conceptual framework (Figure 1) and its operationalisation (Table 1) establish the analytical lens adopted throughout the remainder of this paper. Rather than evaluating transfer learning solely on the basis of improved predictive accuracy, the study investigates whether local calibration provides sufficient empirical evidence to support the trustworthy operational deployment of pretrained GeoAI models in geographically distinct urban environments.

3. Research Design

Consistent with the operational trustworthiness assessment framework introduced in Section 2 (Figure 1), this study was designed to evaluate whether transfer learning provides sufficient evidence for the trustworthy deployment of a pretrained GeoAI model in a geographically distinct urban environment. Rather than focusing solely on predictive performance, the methodological workflow generates complementary evidence relating to robustness under geographic domain shift, reliability on independent data, model generalisation beyond calibration datasets, transparency through error analysis, and explicit assessment of residual uncertainty. Rooftop solar PV mapping in Brisbane serves as the operational case study through which these trustworthiness dimensions are empirically evaluated.

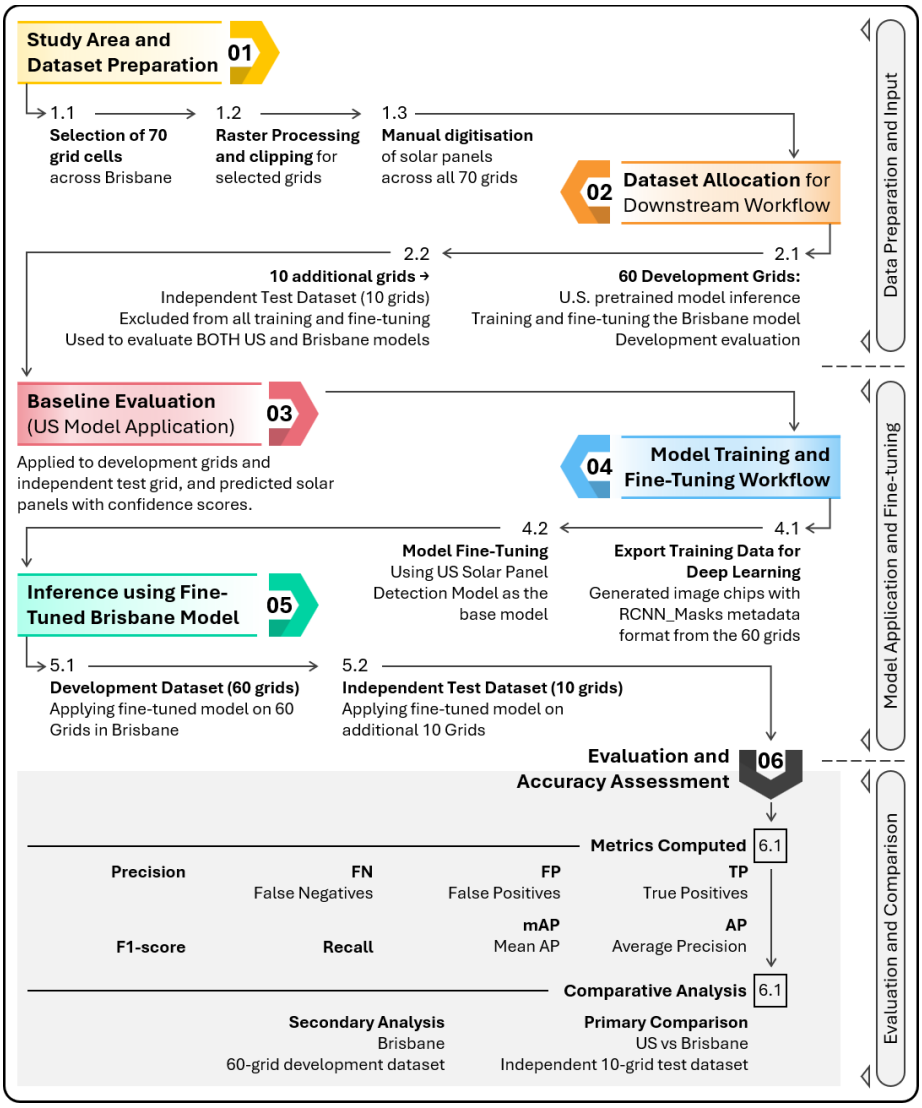


Figure 2: Methodological workflow adopted to operationalise the trustworthiness assessment framework

The methodology comprised three phases: (a) data preparation and dataset construction, (b) model application and local calibration, and (c) evaluation and comparative analysis. The first phase involved selecting representative grid cells across Brisbane, manually digitising rooftop PV installations to create reference data, and allocating the dataset into a development dataset and an independent test dataset. The second phase focused on applying the pretrained US model, fine-tuning it using locally annotated imagery from the development dataset, and subsequently performing inference using both the pretrained and fine-tuned models.

Although both the US and New Zealand solar panel detection models are available through ArcGIS Living Atlas, this study adopts the US model as the baseline for transfer learning. The New Zealand model represents a region-specific fine-tuned adaptation of the US model and is distributed as a packaged Deep Learning Package optimised for New Zealand imagery conditions. Furthermore, model documentation indicates that fine-tuning workflows are primarily designed to extend base pretrained models rather than further fine-tune already localised derivatives (Esri, 2023). Accordingly, the US model provides a clearer and more reproducible foundation for evaluating geographic domain shift and subsequent local calibration within the Australian context.

The final phase involved evaluating both the pretrained US model and the locally calibrated Brisbane model using manually digitised reference data and a consistent object detection evaluation protocol. To ensure a fair assessment of operational transferability, both models were evaluated on the same independent test dataset that was excluded entirely from model development. Complementary analyses on the development dataset were undertaken to evaluate the learning behaviour of the fine-tuned model and to provide a controlled comparison with the pretrained model under identical data and evaluation conditions. Figure 2 summarises the complete methodological workflow adopted in this study, with each phase described in detail in the following sections.

3.1. Study Area and Dataset Preparation

3.1.1. Grids Selection

Brisbane, located in South East Queensland (SEQ), was selected as the study area because of its high rooftop PV adoption and highly varied urban landscape (Simin & Michael James, 2024; Hou et al., 2025). The city is administratively divided into 190 suburbs and is home to over 2.5 million residents (Hossain et al., 2025). In the Australian context, a suburb refers to a locally defined residential or mixed-use district used for planning and infrastructure management (Farahani et al., 2022). Brisbane's urban fabric includes everything from high-density inner-city areas to low-density outer suburbs and newly developed estates (Doan et al., 2021; Li et al., 2024).

To evaluate the operational transferability of the pretrained US rooftop PV detection model and the effectiveness of local calibration, a regular fishnet grid was generated across the Brisbane Local Government Area (LGA). This process produced 6,086 grid cells, each measuring 500 m × 500 m. A total of 70 representative grid cells were selected using a purposive stratified sampling strategy. Of these, 60 grid cells formed the development dataset used for model fine-tuning and controlled performance evaluation, while the remaining 10 grid cells were reserved as an independent test dataset for evaluating operational transferability. The sampling strategy was informed by previous research on rooftop PV adoption in Brisbane and expert knowledge of the local urban environment to ensure broad representation of rooftop configurations, building densities, and urban typologies.

To guide the purposive stratified sampling strategy, six urban zones were identified, and representative grid cells were selected within or intersecting 36 suburbs across the Brisbane LGA (Table 2). Grid selection aimed to capture the distinctive structural and environmental characteristics of each urban zone, including dense roofscapes and image glare in inner-city areas, extensive rooftop PV arrays on industrial buildings, and tree-shaded or sloping rooftops in hilly suburbs. Capturing this diversity was considered essential for developing a locally calibrated model representative of Brisbane's heterogeneous urban environments and for evaluating the operational transferability of the pretrained model across diverse urban contexts. The spatial distribution of the selected grid cells, together with representative imagery and examples of manually digitised rooftop PV installations, is presented in Figure 3.

Table 2: Overview of selected grid cells used in this study

Urban zone	Representative suburbs	No. of grid cells	Urban characteristics
Inner-city	Bulimba, Fortitude Valley, Hamilton, Kangaroo Point, Spring Hill, South Brisbane, West End, Brisbane City	17	High-density rooftops, frequent glare, smaller rooftop PV arrays

Urban zone	Representative suburbs	No. of grid cells	Urban characteristics
Suburban Residential	Carindale, Chelmer, Graceville, Indooroopilly, Jindalee, Mount Gravatt, St Lucia, Toowong, Yeronga	15	Predominantly pitched tile and metal roofs
Industrial / Commercial	Eagle Farm, Rocklea, Acacia Ridge, Archerfield, Moorooka	12	Large flat roofs, extensive panel arrays
Coastal / Bayside	Wynnum, Sandgate, Brighton, Manly, Shorncliffe	10	Glare issues, potential colour distortion due to salt exposure
Leafy / Hilly	Kenmore, Chapel Hill, The Gap	7	Tree shading, mixed roof orientations
Outer Urban / New Estates	Rochedale, Forest Lake, Heathwood, Larapinta, Everton Park, Kuraby	9	Relatively uniform rooftop designs and PV installation patterns

3.1.2. Raster Processing and Digitisation

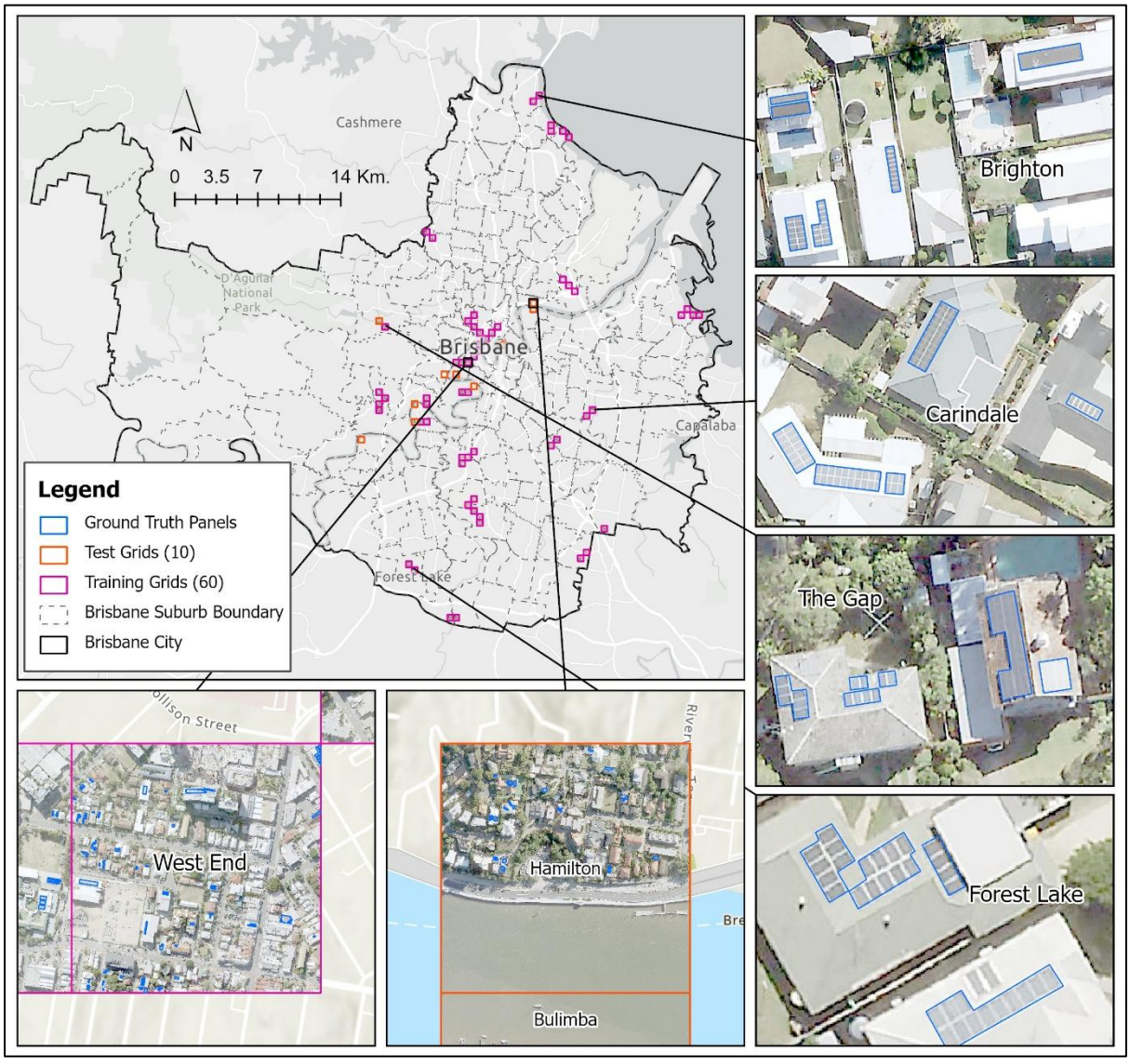


Figure 3: Study area and examples of manually digitised rooftop PV installation features showing variation in polygon representation, ranging from single panels to grouped installations, derived from the high-resolution aerial imagery (Brisbane_2020_10cm_SISP_LGA), sourced from the Queensland SISP

To extract input data for model execution, we sourced high-resolution aerial imagery tiles from the *Brisbane_2020_10cm_SISP_LGA* dataset, part of the Queensland Spatial Imagery Services Program (SISP) (Queensland Government, 2025). The program, coordinated by the Department of Natural Resources and Mines (Queensland Government), systematically acquires and maintains high-resolution aerial imagery across Queensland under the Queensland Spatial Imagery Acquisition Plan. Imagery captured under this program for three years or more is made publicly accessible following the Queensland Government's open data policy, facilitating research and planning applications.

The dataset comprises orthorectified RGB aerial imagery with a 10 cm ground sampling distance (GSD), including three spectral bands (red, green, blue). The high spatial resolution and spectral configuration align with the input specifications of Esri's pretrained US Solar Panel Detection model, which is tuned for high-resolution RGB aerial imagery (5–15 cm GSD). This compatibility enabled direct application of the pretrained model and subsequent local calibration using ArcGIS Pro's *Train Deep Learning Model* tool.

Each grid cell of the study area was clipped individually using a customised workflow built in ArcGIS *ModelBuilder* (Appendix A). The workflow iteratively selects each grid from the Brisbane polygon layer, extracts the matching extent from the imagery, and exports a clipped raster file named according to the grid ID. This automated process ensured consistent preparation of input rasters for each defined spatial unit, streamlining the data preprocessing required for model development and inference.

A total of 4,637 polygon features were digitised across the selected grid cells, representing observed rooftop PV installation features. This corresponds to an average of approximately 67 digitised features per grid cell. It is important to clarify that this number does not directly correspond to the number of individual solar panels, but rather to grouped digitised features based on visual interpretation of aerial imagery. This annotation strategy is consistent with the instance segmentation paradigm adopted by the pretrained Mask R-CNN model, where contiguous rooftop PV installations are treated as individual detection objects rather than individual photovoltaic modules.

As illustrated in Figure 3, some polygons may represent a single panel, while others correspond to paired or arrayed installations. This method aligns with annotation approaches adopted in previous deep learning-based solar panel mapping studies in the US and New Zealand, where manual digitisation prioritised spatial accuracy and operational consistency over strict one-to-one object representation. Such grouping is considered appropriate for model training and evaluation in urban-scale remote sensing contexts, where fine-grained delineation may not be practical or necessary for object detection performance (Esri, 2023, 2025a).

During accuracy assessment, predicted and reference polygons were matched using ArcGIS Pro's Compute Accuracy for Object Detection tool based on IoU thresholds. Consequently, grouped reference annotations and corresponding predictions were evaluated using the same spatial matching criteria, ensuring that the reported performance metrics reflected detection of rooftop PV installation features rather than individual solar panels.

3.2. Dataset Allocation

The 70 selected grid cells were divided into two datasets according to their role in the experimental design. A development dataset comprising 60 grid cells was used for model development and controlled performance comparison. Specifically, these grids were used to (i) establish the baseline performance of the pretrained US rooftop PV detection model under identical conditions, (ii) locally calibrate the pretrained model through transfer learning, and (iii) compare the performance of the pretrained and locally calibrated models using the same imagery and manually digitised reference data.

The remaining 10 grid cells formed an independent test dataset that was excluded entirely from model development. Manual digitisation was undertaken for these grids to provide reference data for accuracy assessment. Following model development, both the pretrained US model and the locally calibrated Brisbane model were applied to this independent dataset to evaluate operational transferability under identical conditions. The independent test dataset therefore provides the primary basis for evaluating operational transferability and the effectiveness of local calibration. By contrast, results obtained from the development dataset are presented separately to demonstrate model learning behaviour following local calibration and should not be interpreted as independent measures of model generalisation.

The separation between the development and independent test datasets is central to the operational trustworthiness perspective adopted in this study. Whereas the development dataset enables evaluation of model learning following transfer learning, the independent test dataset provides evidence regarding model reliability and generalisation in previously unseen geographic environments. This distinction reduces the risk of

overestimating model performance and enables assessment of whether improvements observed during local calibration are maintained under operational deployment conditions.

3.3. Operational Trustworthiness Evaluation

To establish the baseline performance and operational transferability of the pretrained US rooftop PV detection model, the model (*SolarPanelDetection_USA.dlpk*) was applied to each of the 60 raster tiles in the development dataset using the *Detect Objects Using Deep Learning* tool in ArcGIS Pro. The process was automated through a batch workflow developed in *ModelBuilder* (Appendix B), enabling efficient inference across the dataset. Each raster tile produced a classified output raster, which was subsequently converted to vector format. These outputs were merged into a single shapefile representing predicted rooftop PV installation polygons across the study area. The resulting dataset provided the baseline results against which the effectiveness of local calibration through transfer learning was evaluated.

The pretrained US model was executed using a tile size of 400 pixels, 100-pixel padding, and a confidence threshold of 0.9, consistent with the model's original training configuration. Bounding boxes and test-time augmentation were disabled to focus exclusively on mask-based instance segmentation. Inference was performed using an NVIDIA GeForce RTX 3050 GPU to maximise computational efficiency.

Following inference on the development dataset, the pretrained US model was also applied to the independent 10-grid test dataset using identical inference parameters. Applying the same pretrained model to both datasets ensured that comparisons with the locally calibrated model were conducted under identical experimental conditions, using the same imagery specifications, manually digitised reference data, and object detection evaluation protocol.

3.4. Transfer Learning and Local Calibration Workflow

3.4.1. Export Training Data for Deep Learning

To facilitate local calibration of the pretrained US rooftop PV detection model, the *Export Training Data for Deep Learning* tool in ArcGIS Pro 3.6 was used to prepare the training samples. This tool allows for the generation of image tiles (commonly referred to as "chips") paired with annotation metadata suitable for a range of deep learning frameworks such as PyTorch, TensorFlow, and Keras.

The development dataset comprised the 60 grid cells allocated for local calibration and controlled model evaluation. Each grid cell contained high-resolution (10 cm) aerial imagery together with manually digitised rooftop PV installation polygons that served as reference annotations. Image chips were generated by combining these raster and vector inputs using fixed-size windows and consistent geospatial parameters. The export parameters were as follows: 400 × 400 pixel tile size, no stride (i.e., no overlap), and exclusion of tiles without features. Output image format was TIFF, and the metadata format was set to RCNN_Masks, a format required for instance segmentation tasks using models such as Mask R-CNN. This process yielded a total of 2,287 labelled image chips; each associated with its corresponding segmentation mask.

The locally calibrated model retained the Mask R-CNN architecture used by the pretrained model, an extension of the Faster R-CNN framework (Ghimire et al., 2022; Parsaeifar et al., 2025). As illustrated in Figure 4, Mask R-CNN adds a parallel branch to the Faster R-CNN architecture to generate pixel-wise segmentation masks in addition to bounding boxes and class predictions. This structure allows the model to delineate individual solar panel instances with greater spatial precision, particularly important in complex urban environments where panels may be small, closely spaced, or partially occluded (Singh & Viswambharan, 2021; Esri, 2025b; Viswambharan & Singh, 2025).

By employing the RCNN_Masks format, the training dataset retained both spatial detail and object-level instance differentiation. These characteristics were critical for achieving accurate instance segmentation in Brisbane's diverse rooftop landscape.

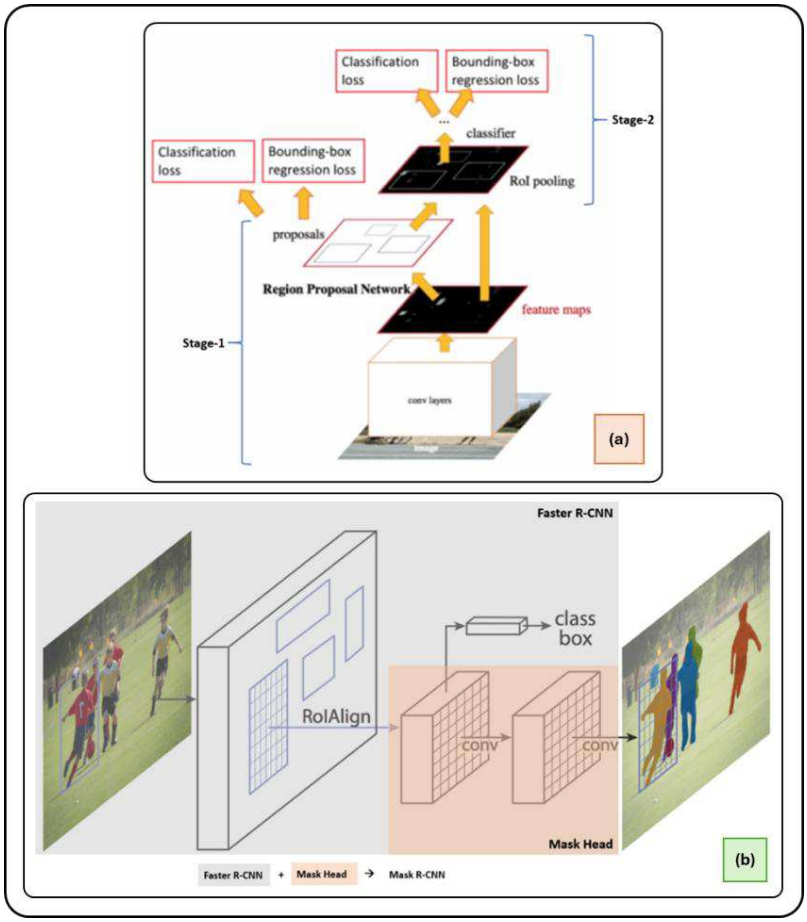


Figure 4: Overview of the Mask R-CNN architecture used for instance segmentation. (a) The Faster R-CNN framework, which consists of a two-stage process: a region proposal network and classification with bounding-box regression. (b) The Mask R-CNN extension, which introduces a mask prediction branch (Mask Head) and replaces RoI pooling with RoIAlign to preserve spatial accuracy in the segmentation outputs. Together, these enhancements enable accurate pixel-level detection of rooftop PV installation instances, derived from Esri Developer (2026)

3.4.2. Fine-Tuning the US Model on Brisbane Data

Local calibration was performed using the *Train Deep Learning Model* tool available in ArcGIS Pro 3.6 (*Image Analyst Extension*). The pretrained US rooftop PV detection model (*SolarPanelDetection_USA.dlpk*) served as the starting point, while the development dataset described in the previous section provided the imagery and manually digitised reference annotations used for transfer learning.

Model calibration was performed on a local workstation equipped with an NVIDIA GeForce RTX 3050 GPU, allowing for GPU-accelerated computation. The pretrained US model (*SolarPanelDetection_USA.dlpk*) was used as the pretrained backbone. Key training parameters included a batch size of 4, a chip size of 400 × 400 pixels, and a maximum of 100 epochs. A validation split of 20% was applied to monitor performance and prevent overfitting. Early stopping was enabled to terminate model calibration when validation loss ceased to improve, thereby reducing the risk of overfitting. Default data augmentation settings were used, incorporating transformations such as cropping, zooming, brightness adjustment, and contrast changes to enhance generalisability. The FREEZE_MODEL setting was applied to retain the pretrained feature extraction weights of the backbone while updating only the segmentation-specific layers.

The underlying architecture remained Mask R-CNN with a Feature Pyramid Network (FPN) and ResNet-101 backbone. This instance segmentation approach allows for pixel-level detection and delineation of solar panel footprints, which is particularly advantageous in complex urban settings with varied rooftop structures and high solar panel density. The training process involved iterative backpropagation using stochastic gradient descent, with performance evaluated based on validation loss. Following local calibration, the adapted model was subsequently used for inference on both the development and independent test datasets.

3.5. Inference Using Locally Calibrated Model

Following local calibration through transfer learning, inference was conducted using the locally calibrated Brisbane rooftop PV detection model (*SolarPanelDetectionBrisbane.dlpk*). The inference procedure closely mirrored that used for the pretrained US model, ensuring a controlled comparison between the two models.

Inference was performed on two datasets serving different purposes. First, inference was conducted on the 60-grid development dataset to examine model learning behaviour following local calibration. Second, the model was applied to the independent 10-grid test dataset, which provided the primary assessment of operational transferability and model generalisation. Using identical inference parameters for both datasets ensured that any performance differences reflected model capability rather than processing configuration.

Inference on both datasets was automated using the same *ModelBuilder* workflow previously developed for the pretrained US model. The only modification involved replacing the pretrained US model with the locally calibrated Brisbane model (*SolarPanelDetectionBrisbane.dlpk*) as the input model. All other parameters, including tile size (400 × 400 pixels), padding (100 pixels), batch size (4), and a confidence threshold of 0.9, were kept constant to ensure direct comparability across experiments. Model outputs comprised instance segmentation masks representing predicted rooftop PV installation features. These outputs were converted to vector polygons and merged into a single shapefile for each dataset. The resulting prediction layers were subsequently compared with the manually digitised reference data using the object detection evaluation protocol described in Section 3.6.

3.6. Operational Trustworthiness Evaluation

The model evaluation extended beyond conventional object detection accuracy. While predictive performance remains an essential requirement for GeoAI applications, trustworthy deployment also requires evidence that a model performs robustly across geographic contexts, generalises beyond the data used for calibration, behaves transparently under operational conditions, and retains known limitations that can be explicitly communicated to end users. Accordingly, this study evaluates the pretrained and locally calibrated models using multiple complementary analyses that collectively assess the trustworthiness of pretrained GeoAI deployment under geographic domain shift.

Model performance was first quantified using the Compute Accuracy for Object Detection tool available in ArcGIS Pro 3.6 (Esri, 2026). This tool compares predicted rooftop photovoltaic (PV) installation features with manually digitised reference annotations using a standardised object detection evaluation framework. Both the pretrained US model and the locally calibrated Brisbane model were evaluated using identical processing parameters and reference datasets to ensure fair and reproducible comparison throughout the study.

The evaluation framework combines spatial overlap measures with confidence-based performance metrics widely adopted in object detection research (Hui, 2018; Padilla et al., 2020; Esri, 2026). Predicted rooftop PV polygons were compared with manually digitised reference polygons using Intersection over Union (IoU), which measures the degree of overlap between predicted and reference objects. A prediction was classified as a True Positive (TP) when its IoU exceeded the specified threshold. Predictions without sufficient overlap were classified as False Positives (FP), while reference rooftop PV features that remained undetected were classified as False Negatives (FN). These object-level matches formed the basis for calculating the evaluation metrics summarised in Table 3.

Table 3: Evaluation metrics used for object detection accuracy assessment

Metric	Description	Formula	Range
True Positive	Number of predicted rooftop PV features that sufficiently overlap with reference rooftop PV features		
False Positive	Predicted rooftop PV features that do not sufficiently overlap any reference rooftop PV feature		
False Negative	Reference rooftop PV features missed by the model (no matching prediction)		
Precision	Proportion of correct predictions out of all predicted rooftop PV features	$TP/(TP + FP)$	0 to 1
Recall	Proportion of detected reference rooftop PV features out of all reference rooftop PV features	$TP/(TP + FN)$	0 to 1

F1 Score	Harmonic mean of precision and recall; balances both metrics.	$\frac{2 \times (Precision \times Recall)}{Precision + Recall)}$	0 to 1
Average Precision	Area under the Precision–Recall curve at a single IoU threshold (typically 0.5).	Area under interpolated PR curve	0 to 1
Mean Average Precision	Mean of AP scores over multiple IoU thresholds (e.g., 0.5 to 0.95, step 0.05).	Average of APs across IoU thresholds	0 to 1

Although these metrics provide quantitative measures of detection accuracy, they alone are insufficient for assessing trustworthy deployment. Accordingly, the object detection results were interpreted through the operational trustworthiness framework introduced in Section 2. Specifically:

- Robustness was evaluated by comparing model performance before and after transfer from the United States to Brisbane, thereby quantifying the effect of geographic domain shift.
- Reliability was assessed using an independent 10-grid test dataset that was excluded entirely from model calibration and internal validation.
- Generalisation was evaluated by comparing performance on the development dataset with performance on the independent test dataset to determine whether improvements obtained during local calibration were maintained on previously unseen imagery.
- Transparency was examined through analysis of confidence score distributions together with detailed visual assessment of TPs, FPs, FNs, and geometric segmentation errors.
- Residual uncertainty was evaluated by comparing performance across multiple IoU thresholds (0.50, 0.75, and 0.95), enabling assessment of model sensitivity to increasingly stringent spatial matching criteria.

Performance evaluation was therefore undertaken using two complementary datasets with distinct purposes. The 60-grid development dataset was used to establish the baseline performance of the pretrained US model, support local calibration through transfer learning, and examine model learning behaviour following fine-tuning. Because these grids contributed to model development, the resulting evaluation provides evidence of model learning rather than independent model performance.

The principal assessment of trustworthy deployment was conducted using the independent 10-grid test dataset, which was excluded entirely from model development, calibration, and internal validation. Both the pretrained US model and the locally calibrated Brisbane model were evaluated under identical inference conditions using this dataset. Consequently, comparisons between the two models on the independent dataset provide the primary evidence for assessing operational robustness, reliability, and generalisation following transfer learning.

Finally, quantitative evaluation was complemented by qualitative error analysis to identify the types of detection failures that remained after local calibration. Representative examples of FPs, FNs, geometric segmentation errors, and successful detections were examined to better understand model behaviour under diverse urban conditions. This combined quantitative and qualitative evaluation enables transfer learning to be assessed not only in terms of improved predictive accuracy, but also in terms of whether it provides sufficient evidence for the trustworthy operational deployment of pretrained GeoAI models in geographically distinct environments.

4. Analysis and Results

4.1. Local Calibration of the Pretrained Model

To locally calibrate the pretrained US rooftop PV detection model for the Brisbane context, transfer learning was performed using manually annotated rooftop PV installation features from the 60-grid development dataset. The training process ran for 32 epochs, utilising a one-cycle learning rate policy, with rates ranging from approximately 1.58×10^{-5} to 1.58×10^{-4}. This scheduling technique facilitates faster convergence and assists in avoiding suboptimal local minima during optimisation. Importantly, the *Train Deep Learning Model* tool in ArcGIS incorporates an early stopping mechanism, which automatically terminates training when validation performance no longer improves, thereby preventing overfitting and reducing computational overhead.

Figure 5 presents the progression of key training metrics over the full training cycle, including training loss, validation loss, AP, and epoch duration. The training loss decreased steadily from 0.88 at Epoch 1 to 0.41 by Epoch 28, reflecting the model's ability to minimise error on the training set. Validation loss followed a similar downward trend, from 0.74 to around 0.41, with minor fluctuations in later epochs, a typical occurrence

attributable to variance in the validation sample or mild overfitting. However, the alignment between training and validation loss trajectories suggests stable generalisation performance.

The AP score, a summary metric representing the area under the precision–recall curve, improved consistently from 0.43 to a peak of 0.86 by Epoch 24, before plateauing. This trajectory indicates progressive enhancement in detection accuracy and suggests that additional training beyond this point yielded diminishing returns. Epoch durations ranged between 25 to 38 minutes, reflecting variability in computational load and data processing across epochs. Upon completion, the trained model was exported in Esri's standard *.dlpk* (Deep Learning Package) format, enabling seamless integration with other ArcGIS tools for inference, sharing, or further deployment across different environments.

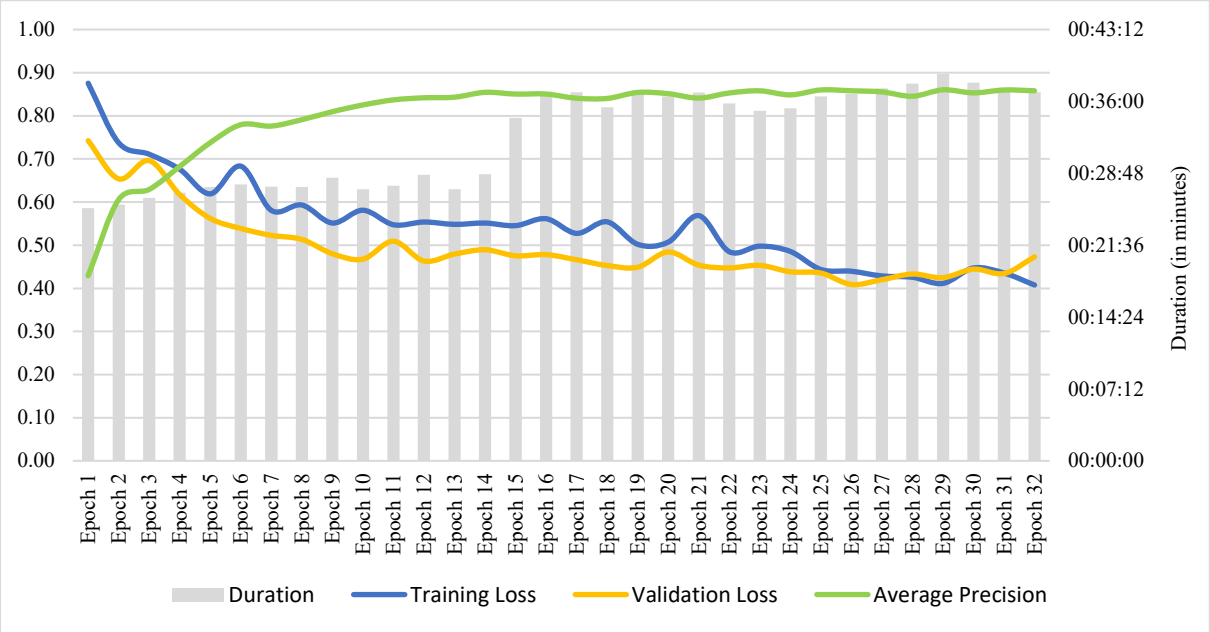


Figure 5: Evolution of key training metrics across all epochs, including training loss, validation loss, average precision, and epoch duration

Overall, the training results demonstrate that the fine-tuned model converged effectively and achieved a balanced performance without evidence of underfitting or overfitting. The steadily increasing AP highlights the model's growing confidence and precision in solar panel detection, a critical requirement for its application in complex and diverse urban settings such as Brisbane.

4.2. Robustness under Geographic Domain Shift

To evaluate the robustness of the pretrained GeoAI model under geographic domain shift, the pretrained US model and the locally calibrated Brisbane model were first evaluated using the 60-grid development dataset. Although these data contributed to model development and therefore do not represent independent validation, they provide useful evidence regarding how effectively transfer learning adapts a pretrained model to a new geographic environment. This section therefore addresses the robustness dimension of the operational trustworthiness framework introduced in Section 2, while the following sections evaluate reliability, generalisation, transparency, and residual uncertainty using independent evidence.

Figure 6 compares the confidence score distributions produced by the pretrained US model and the locally calibrated Brisbane model on the development dataset. Both models generated highly confident predictions, with mean confidence values exceeding 97% and a large proportion of detections assigned confidence scores above 98%. Interestingly, local calibration produced only minor changes in the confidence distribution despite substantially improving detection performance. This finding demonstrates that prediction confidence alone should not be interpreted as evidence of trustworthy model behaviour. A model may produce highly confident predictions while still exhibiting systematic omission or commission errors after deployment in a new geographic environment. Consequently, confidence scores require independent validation before they can be interpreted as evidence of trustworthy deployment.

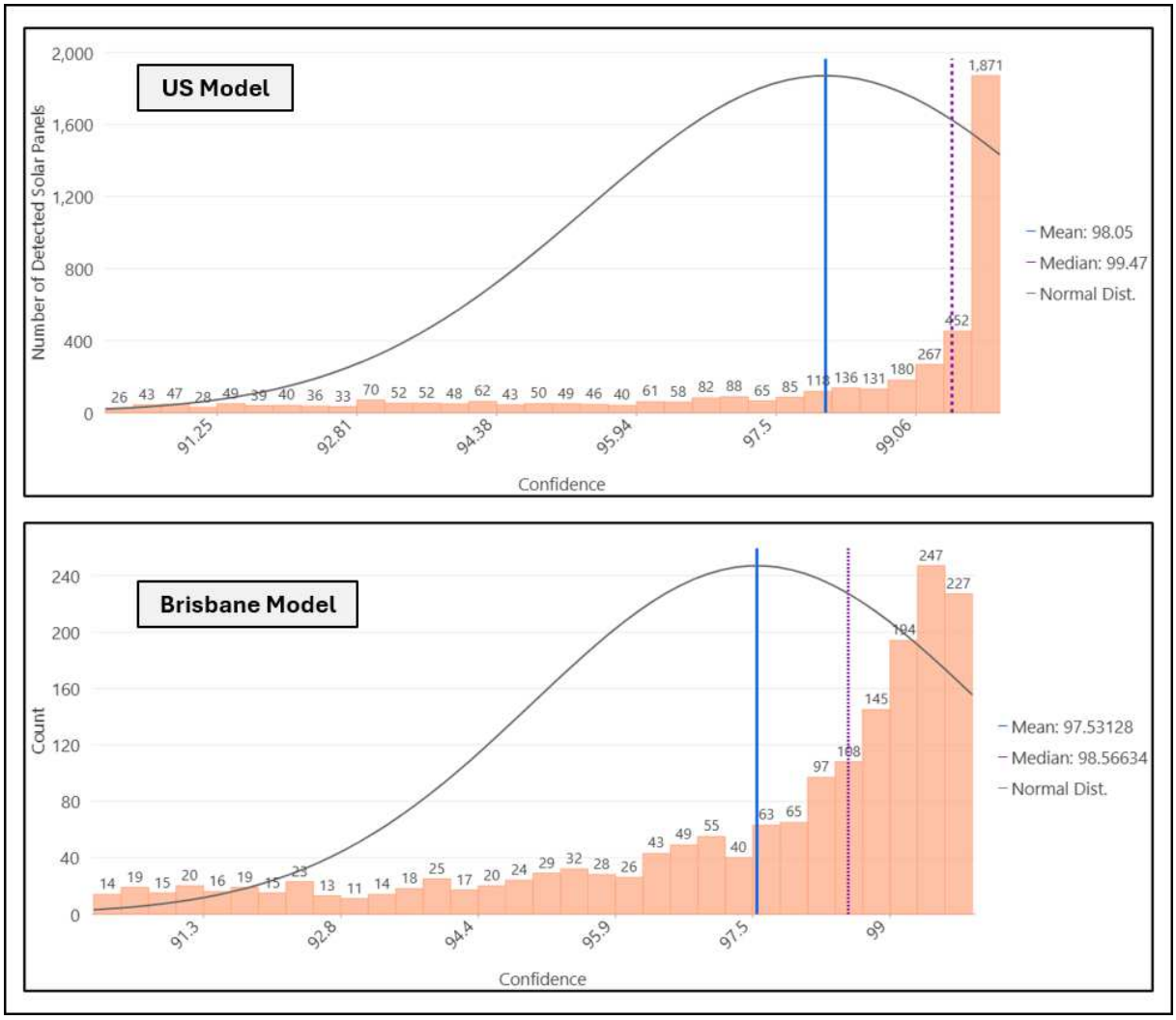


Figure 6: Comparison of confidence score distributions produced by the pretrained US model and the locally calibrated Brisbane model on the development dataset

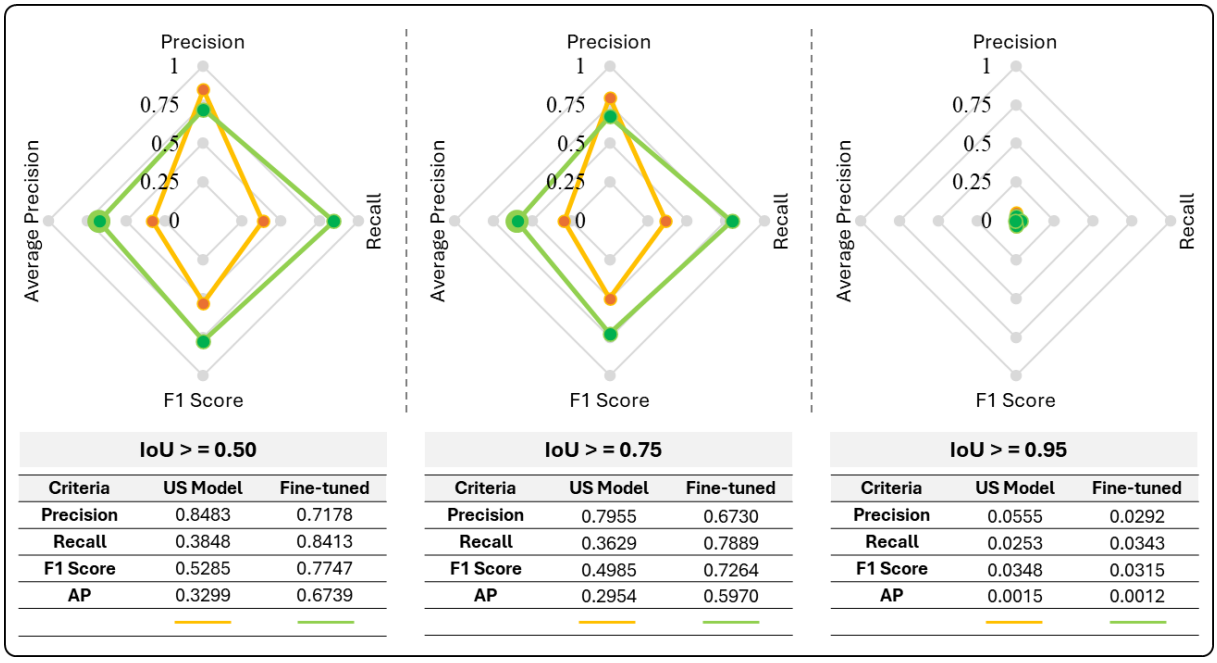
Table 4 provides quantitative evidence that local calibration substantially improved robustness under geographic domain shift. At IoU ≥ 0.50, TP increased from 1,460 to 3,192, while FNs declined from 2,334 to only 602, indicating that transfer learning substantially improved the model's ability to detect rooftop PV installations within the Brisbane environment. Comparable improvements were observed at IoU ≥ 0.75. However, both models experienced a pronounced reduction in performance at IoU ≥ 0.95, suggesting that although transfer learning improves detection capability, accurate pixel-level delineation of rooftop PV arrays remains challenging under strict geometric matching criteria.

Table 4: Comparison of detection outcomes for the pretrained US model and the locally calibrated Brisbane model on the development dataset

IoU	Model	True Positive	False Positive	False Negative
IoU > = 0.50	US Model	1,460	271	2,334
	Brisbane Model	3,192	1,255	602
IoU > = 0.75	US Model	1,377	354	2,417
	Brisbane Model	2,993	1,454	801
IoU > = 0.95	US Model	96	1,635	3,698
	Brisbane Model	130	4,317	3,664

Figure 7 demonstrates that local calibration substantially improved detection performance at operationally meaningful IoU thresholds (0.50 and 0.75). The most substantial improvement was observed in recall, which increased from 38.48% to 84.13% at IoU ≥ 0.50, while Average Precision more than doubled. Although

precision decreased moderately, the resulting increase in F1-score indicates a considerably better balance between omission and commission errors. These findings suggest that transfer learning substantially enhanced the robustness of the pretrained model when transferred to the Brisbane environment.



IoU > = 0.50

Criteria	US Model	Fine-tuned
Precision	0.8483	0.7178
Recall	0.3848	0.8413
F1 Score	0.5285	0.7747
AP	0.3299	0.6739

IoU > = 0.75

Criteria	US Model	Fine-tuned
Precision	0.7955	0.6730
Recall	0.3629	0.7889
F1 Score	0.4985	0.7264
AP	0.2954	0.5970

IoU > = 0.95

Criteria	US Model	Fine-tuned
Precision	0.0555	0.0292
Recall	0.0253	0.0343
F1 Score	0.0348	0.0315
AP	0.0015	0.0012

Figure 7: Comparative performance of the pretrained US model and the locally calibrated Brisbane model across three IoU thresholds, illustrating the effect of transfer learning on robustness under geographic domain shift

At IoU ≥ 0.95, performance declined substantially for both models, reflecting the inherent difficulty of achieving near-perfect geometric agreement for rooftop PV segmentation. This behaviour suggests that although transfer learning successfully adapts pretrained feature representations, residual uncertainty remains in the delineation of object boundaries. From the perspective of trustworthy GeoAI, these results indicate that improved predictive performance alone should not be considered sufficient evidence of trustworthy deployment, particularly when applications require highly precise spatial delineation.

Because these results were obtained using the same dataset employed for transfer learning, they should be interpreted as evidence of successful model adaptation rather than independent evidence of trustworthy deployment. Demonstrating trustworthiness additionally requires evidence that performance improvements generalise beyond the calibration data. The following section therefore evaluates both models on an independent test dataset to examine reliability and generalisation under operational conditions.

4.3. Reliability and Generalisation on an Independent Test Dataset

To determine whether the improvements achieved through transfer learning generalised beyond the development dataset, both the pretrained US model and the locally calibrated Brisbane model were evaluated using an independent 10-grid test dataset that was excluded entirely from model development. This dataset provides the primary evidence for assessing the reliability and generalisation dimensions of the operational trustworthiness framework introduced in Section 2 because it represents previously unseen geographic conditions. Unlike the development dataset, performance on this dataset reflects operational behaviour rather than model adaptation during transfer learning.

Figure 8 compares the confidence score distributions produced by the pretrained US model and the locally calibrated Brisbane model on the independent test dataset. Similar to the development dataset, both models generated highly confident predictions, with mean confidence values exceeding 97% and median confidence values approaching 98–99%. Local calibration produced only modest changes in the overall confidence distribution despite substantially improving detection performance. This finding further demonstrates that prediction confidence alone provides little evidence of operational reliability or model generalisation. Highly confident predictions may still contain systematic omission and commission errors when models are deployed in previously unseen geographic environments. Independent evaluation against reference data therefore remains essential for determining trustworthy deployment.

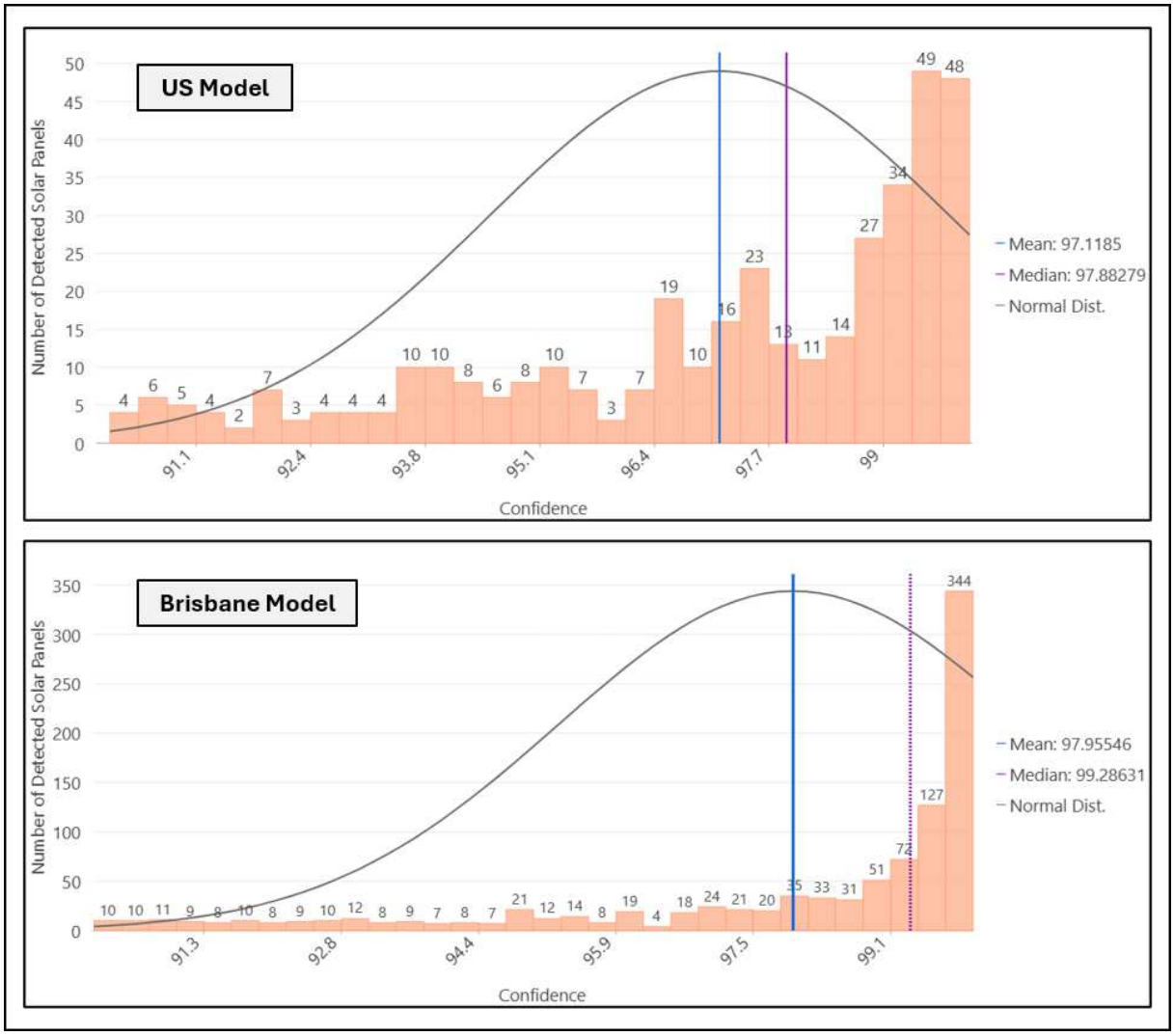


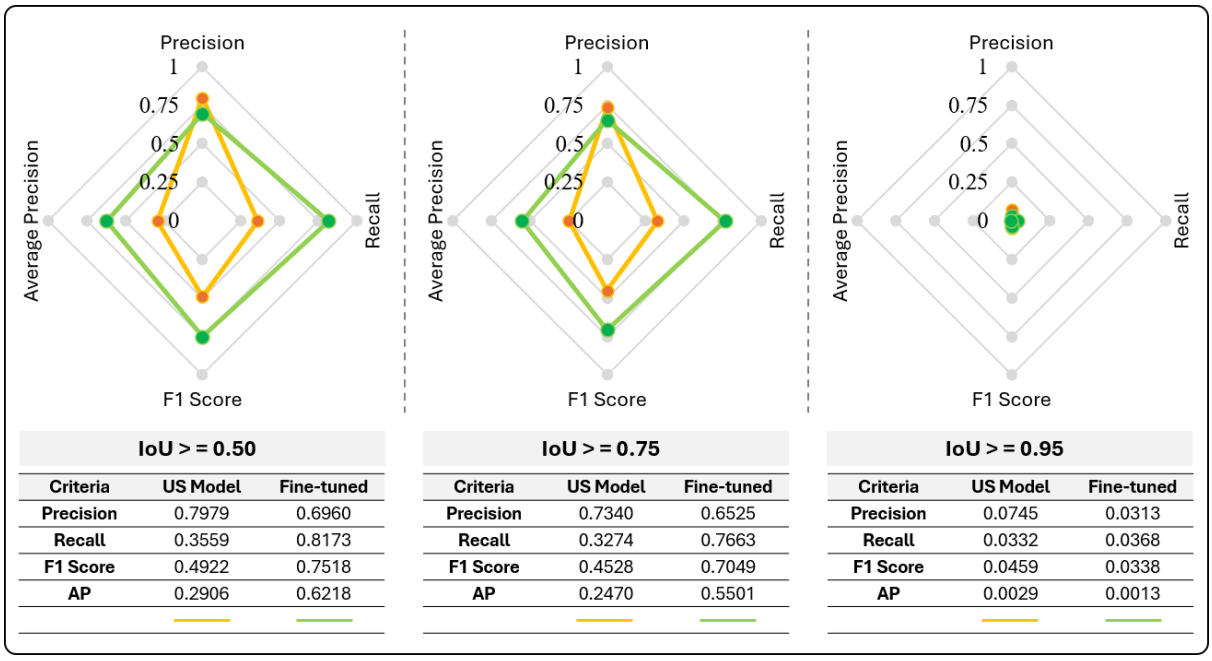
Figure 8: Confidence score distributions for the pretrained US model and the locally calibrated Brisbane model on the independent test dataset

Table 5 compares the detection outcomes of the pretrained US model and the locally calibrated Brisbane model on the independent dataset. At IoU ≥ 0.50, transfer learning increased the number of TPs from 300 to 689 while reducing FNs from 543 to 154. Similar improvements were observed at IoU ≥ 0.75. These results indicate that the improvements achieved through transfer learning were not confined to the development dataset but generalised to previously unseen geographic areas. This provides stronger evidence of operational reliability than evaluation on the calibration dataset alone. However, both models experienced a substantial decline in performance at IoU ≥ 0.95, highlighting the continuing challenge of achieving accurate pixel-level segmentation under strict geometric matching criteria.

Table 5: Comparison of detection outcomes for the pretrained US model and the locally calibrated Brisbane model on the independent dataset

IoU	Model	True Positive	False Positive	False Negative
IoU > = 0.50	US Model	300	76	543
	Brisbane Model	689	301	154
IoU > = 0.75	US Model	276	100	567
	Brisbane Model	646	344	197
IoU > = 0.95	US Model	28	348	815
	Brisbane Model	31	959	812

Figure 9 provides a more comprehensive comparison of model performance across multiple evaluation metrics. At IoU ≥ 0.50, recall increased from 35.59% to 81.73%, while the F1-score improved from 0.4922 to 0.7518 and Average Precision more than doubled following transfer learning. Although precision decreased moderately, the locally calibrated Brisbane model achieved a substantially better balance between omission and commission errors, indicating improved operational reliability. Similar improvements were observed at IoU ≥ 0.75, demonstrating that the benefits of local calibration extended beyond the calibration dataset and generalised to independent data.



IoU > = 0.50

Criteria	US Model	Fine-tuned
Precision	0.7979	0.6960
Recall	0.3559	0.8173
F1 Score	0.4922	0.7518
AP	0.2906	0.6218

IoU > = 0.75

Criteria	US Model	Fine-tuned
Precision	0.7340	0.6525
Recall	0.3274	0.7663
F1 Score	0.4528	0.7049
AP	0.2470	0.5501

IoU > = 0.95

Criteria	US Model	Fine-tuned
Precision	0.0745	0.0313
Recall	0.0332	0.0368
F1 Score	0.0459	0.0338
AP	0.0029	0.0013

Figure 9: Comparative detection performance of the pretrained US model and the locally calibrated Brisbane model across three IoU thresholds on the independent test dataset

Compared with the development dataset, performance declined slightly on the independent dataset, an expected outcome reflecting genuine model generalisation rather than memorisation of the calibration data. Nevertheless, the locally calibrated Brisbane model consistently outperformed the pretrained US model across all operationally relevant metrics, demonstrating that the benefits of transfer learning extended beyond the calibration dataset.

From the perspective of trustworthy GeoAI deployment, these findings provide evidence that local calibration improves not only robustness but also operational reliability and generalisation. However, the persistence of FPs, missed detections, and declining performance under stricter IoU thresholds demonstrates that meaningful residual uncertainty remains following transfer learning. Consequently, independent evaluation should be regarded as an essential prerequisite for trustworthy deployment rather than merely a final step in model validation. Although the independent evaluation demonstrates that transfer learning substantially improves operational reliability and model generalisation, quantitative performance metrics alone cannot fully explain why errors persist. The following section therefore examines model failure modes to improve transparency and better understand the residual uncertainties that remain after local calibration.

4.4. Transparency and Residual Uncertainty

The quantitative analyses presented in Sections 4.2 and 4.3 established evidence that transfer learning substantially improved the robustness, reliability, and generalisation of the pretrained GeoAI model. The final two dimensions of the operational trustworthiness framework concern transparency and residual uncertainty. Accordingly, this section complements the quantitative findings through qualitative error analysis, examining representative examples of model behaviour to explain why prediction errors occur and to identify the residual uncertainties that remain following local calibration.

To improve transparency, qualitative assessment focused on four complementary aspects of model behaviour: (i) recovery of previously missed rooftop PV installations following transfer learning, (ii) persistent FP detection patterns, (iii) residual geometric uncertainty in instance segmentation, and (iv) model behaviour on previously unseen urban environments.

4.4.1. *Recovery of Missed Rooftop PV Installations through Transfer Learning*



Figure 10: Comparative examples of rooftop solar panel detections. Images U-a to U-f show detections produced by the pretrained US model, while F-a to F-f display results from the fine-tuned Brisbane model for the same locations

Transparency requires understanding not only whether model performance improves, but also how transfer learning changes model behaviour. Figure 10 presents representative examples where the pretrained US model failed to detect rooftop PV installations that were subsequently recovered following local calibration. Panels U-a to U-f (Figure 10) display detections from the US model, while F-a to F-f (Figure 10) show outputs from the fine-tuned Brisbane model at the same locations.

Across large commercial rooftops (U-a, U-b, U-f), the US model either failed to detect substantial portions of solar arrays or produced incomplete and fragmented outputs. These errors suggest limited generalisation to differences in rooftop material, array configuration, and illumination conditions present in the Australian context. In medium-density residential areas (U-c, U-d, U-e), under-detection was particularly pronounced, with smaller or irregularly arranged installations frequently omitted. Such FNs indicate that the pretrained model's learned feature representations did not fully capture the variability of local rooftop morphology.

Following fine-tuning with locally annotated samples, the Brisbane model (F-a to F-f) demonstrates substantially improved detection coverage and spatial consistency. In each instance, the US model either entirely missed the solar panels (FNs) or detected only partial segments with incorrect geometry. In contrast, the fine-tuned Brisbane model showed a markedly improved capacity to identify these panels, producing detections with more accurate alignment and geometry. These results underscore the value of local calibration via transfer learning, particularly for urban form and rooftop heterogeneity that is not well represented in global or foreign

datasets. These qualitative observations provide visual evidence supporting the quantitative improvements reported in Sections 4.2 and 4.3, reinforcing that transfer learning substantially improves model adaptation under geographic domain shift.

4.4.2. Persistent False Positive Detection Patterns



Figure 11: Examples of FPs from the US pretrained model and the fine-tuned Brisbane model

Despite the substantial reduction in missed detections, transfer learning did not eliminate all prediction errors. One of the most important remaining sources of uncertainty was the occurrence of FP detections. As shown in Figure 11, examples from both the US pretrained model (top two rows) and the fine-tuned Brisbane model (bottom two rows) reveal distinct misclassification patterns rooted in each model's limitations.

The US model frequently misidentified elements from construction zones, such as scaffolding and reflective debris, due to their spectral and geometric resemblance to PV panels. It also confused playgrounds and sports courts, especially those with rectilinear markings or coloured surfaces, for solar arrays, often influenced by shadows or viewing angles. Additional errors included misclassifying dark vehicles, road paint, and long shadows, as well as rooftop features like solar hot water systems, revealing a lack of spatial understanding of local rooftop configurations. The fine-tuned model trained on Brisbane data showed notable improvements, with fewer and less severe FPs. However, occasional errors still occurred, particularly in areas with complex shadows, pedestrian bridges, or urban edges, where certain features resembled panel footprints.

These observations highlight the importance of local training data for reducing FP detections while also demonstrating that certain urban features remain visually ambiguous, even after local calibration. From a trustworthy deployment perspective, these remaining FPs demonstrate that local calibration reduces, but does not eliminate, the need for post-processing and human verification in operational workflows.

4.4.3. *Residual Geometric Uncertainty in Instance Segmentation*

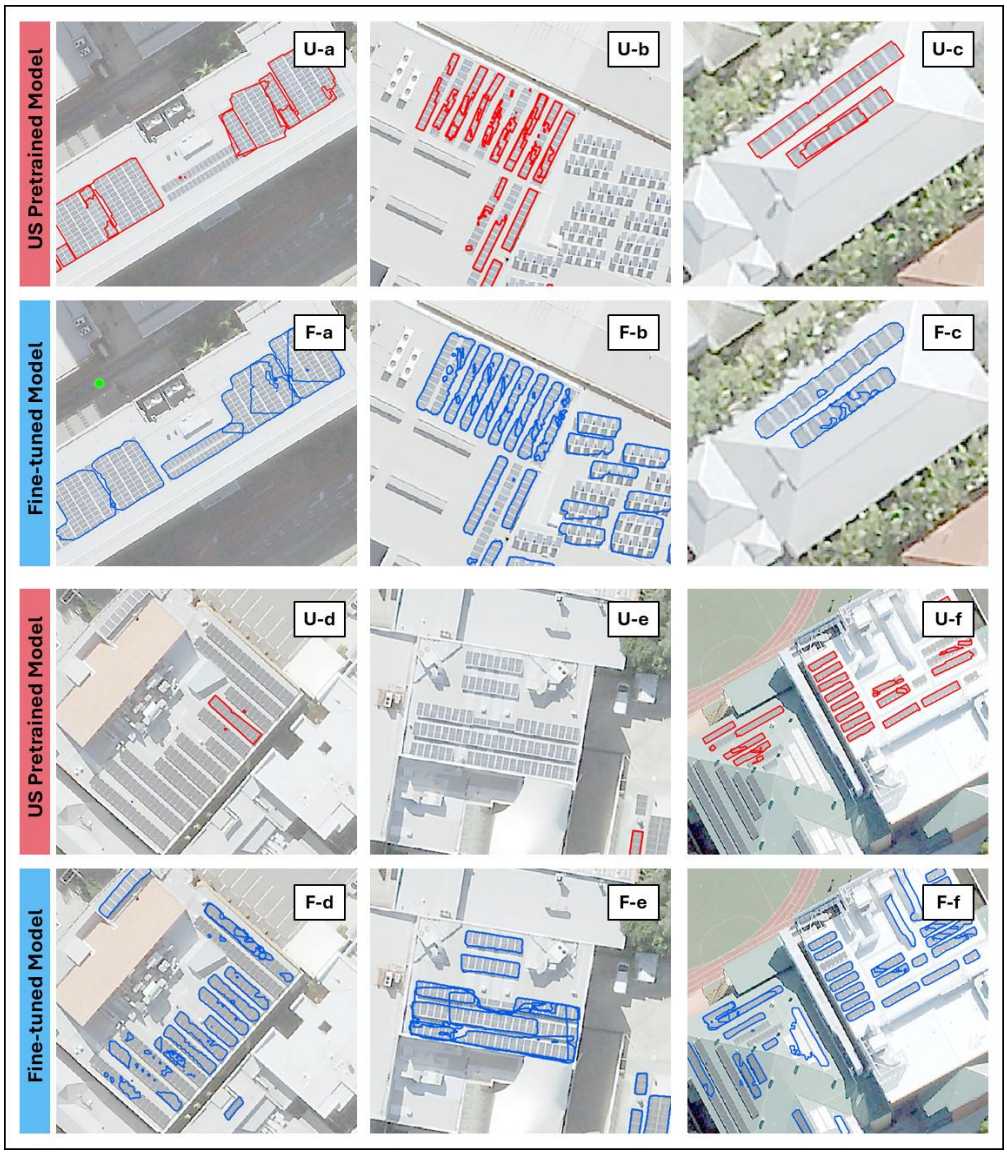


Figure 12: Comparative examples of geometric distortions in solar panel segmentation. Panels U-a to U-f display outputs from the pretrained US model, while F-a to F-f present the corresponding results from the fine-tuned Brisbane model for identical rooftop locations

Trustworthy deployment depends not only on detecting rooftop PV installations but also on accurately delineating their spatial extent. Qualitative inspection revealed that geometric uncertainty remained even when detections were classified as TPs. These patterns are illustrated in Figure 12, where panels U-a to U-f present segmentation outputs from the pretrained US model, and F-a to F-f show the corresponding results from the fine-tuned Brisbane model for the same rooftop locations.

Across several commercial rooftop settings in Figure 12 (U-a, U-b, U-d, U-f), the pretrained model produced detections that broadly overlapped with the solar arrays but exhibited irregular boundaries, jagged contours, and partial segmentation of continuous panel rows. In other cases (e.g., U-b and U-e), only portions of larger installations are captured, indicating under-segmentation despite correct overall localisation. These results suggest that while the pretrained model recognises panel presence, it struggles to delineate coherent array structures under varying rooftop configurations.

The fine-tuned Brisbane model (F-a to F-f in Figure 12) demonstrates improved detection coverage and spatial consistency overall; however, geometric distortions remain visible. For example, in Figure 12 (F-b and F-d), the model successfully delineates more complete panel clusters than the pretrained counterpart, yet introduces boundary noise and contour irregularities along elongated arrays. These distortions suggest that,

although local calibration improved object detection, accurately delineating complex rooftop PV arrays remains challenging due to the variability of rooftop geometry and panel arrangement.

Importantly, these geometric inconsistencies do not necessarily translate into false detections but can substantially influence performance under stricter evaluation criteria. This is reflected in the sharp decline in metrics at higher IoU thresholds (e.g., IoU ≥ 0.95), where minor boundary misalignments significantly reduce intersection-over-union scores. From a practical perspective, such distortions may limit the reliability of outputs for applications requiring precise rooftop area estimation, capacity modelling, or automated asset quantification. These findings demonstrate that conventional quantitative metrics alone cannot fully characterise segmentation quality. Qualitative assessment therefore provides an important complementary source of evidence when evaluating trustworthy GeoAI deployment.

4.4.4. Model Behaviour on Previously Unseen Urban Environments

The final aspect of transparency concerns understanding model behaviour when transferred to entirely unseen urban environments. Figure 13 therefore presents representative detection examples from the independent test dataset. These examples illustrate how the model performs when applied to rooftop configurations that were not encountered during model development. The top row of Figure 13 illustrates well-aligned detections of solar panels, varying shapes and sizes, often under different lighting conditions, rooftop orientations, and background clutter. These examples suggest that the locally calibrated model learned feature representations that generalise to previously unseen urban environments.



Figure 13: Representative detection results produced by the fine-tuned Brisbane model on the independent 10-grid test dataset

Nevertheless, the bottom row highlights more challenging scenarios involving buildings with large and dense solar arrays. While the model still detects panels in these scenes, the predicted polygons exhibit notable geometric distortions. These include over-segmentation, fragmented or inconsistent boundaries, and slight misalignments with actual panel edges. Unlike in simpler rooftops, here the model struggles to maintain structural coherence across densely packed panel clusters or irregular rooftop layouts.

These errors appear to arise not from geographic unfamiliarity alone, but from scene complexity and panel density. Factors such as occlusions from rooftop infrastructure, tightly grouped panel configurations, and the presence of fine-grained architectural detail may introduce ambiguity in edge delineation. These observations indicate that transfer learning substantially improves the model's ability to recognise rooftop PV installations across diverse urban environments. However, the persistence of geometric uncertainty demonstrates that model adaptation does not completely eliminate operational uncertainty in previously unseen urban settings.

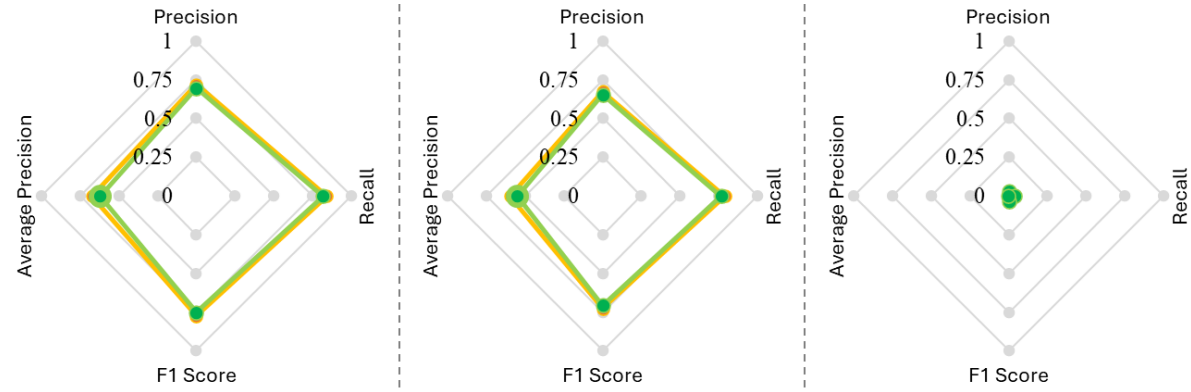
Collectively, the qualitative analyses presented in this section complement the quantitative evaluations reported in Sections 4.2 and 4.3 by providing greater transparency into model behaviour under operational conditions. Rather than simply demonstrating improved accuracy, the visual analyses explain how transfer learning reduces missed detections, improves segmentation consistency, and enhances model adaptation to a new geographic environment while simultaneously revealing the residual uncertainties that remain. These findings reinforce the central argument of this study: transfer learning substantially improves several dimensions

of trustworthy GeoAI deployment but, on its own, does not provide sufficient evidence for trustworthy deployment. Responsible deployment therefore requires independent evaluation, transparent reporting of model behaviour, and explicit acknowledgement of residual uncertainties alongside conventional quantitative performance assessment.

5. Findings and Discussions

5.1. Transfer Learning Improves Deployment but Does Not Establish Trustworthiness

The principal finding of this study is that transfer learning substantially improved the operational performance of the pretrained GeoAI model when deployed in a geographically distinct urban environment. Local calibration consistently increased robustness to geographic domain shift, improved reliability on previously unseen data, and enhanced model generalisation compared with the original pretrained US model. These improvements demonstrate the practical value of transfer learning for adapting pretrained GeoAI models to local conditions and support its use as an effective strategy for mitigating geographic domain shift within operational geospatial workflows. However, the findings also demonstrate that improved predictive performance alone should not be interpreted as sufficient evidence for trustworthy deployment.



IoU > = 0.50

Criteria	On 60 Grids	On 10 Grids
Precision	0.7178	0.6960
Recall	0.8413	0.8173
F1 Score	0.7747	0.7518
AP	0.6739	0.6218

IoU > = 0.75

Criteria	On 60 Grids	On 10 Grids
Precision	0.6730	0.6525
Recall	0.7889	0.7663
F1 Score	0.7264	0.7049
AP	0.5970	0.5501

IoU > = 0.95

Criteria	On 60 Grids	On 10 Grids
Precision	0.0292	0.0313
Recall	0.0343	0.0368
F1 Score	0.0315	0.0338
AP	0.0012	0.0013

Figure 14: Comparison of the locally calibrated Brisbane model on the development dataset (60 grids) and an independent operational evaluation dataset (10 grids). The relatively small reduction in performance indicates that improvements achieved through transfer learning generalise beyond the calibration data thereby highlighting the importance of independent evaluation before operational deployment

Figure 14 provides additional evidence supporting this interpretation by comparing the performance of the locally calibrated Brisbane model on the development dataset and the independent evaluation dataset. Although performance declined modestly on previously unseen data, the differences across Precision, Recall, F1-score, and Average Precision were comparatively small at the operationally relevant IoU thresholds (0.50 and 0.75). This suggests that the improvements achieved through transfer learning extended beyond the calibration dataset and reflected genuine model generalisation rather than overfitting to the calibration dataset. From the perspective of trustworthy GeoAI, this distinction is important because operational deployment depends on consistent performance under previously unseen geographic conditions rather than performance on the data used for model development.

Nevertheless, successful generalisation should not be equated with trustworthy deployment. Despite the substantial improvements observed following local calibration, the independent evaluation and qualitative error analysis revealed persistent FPs, missed detections, and geometric segmentation errors. Furthermore, model performance declined considerably under stricter IoU thresholds, indicating that accurately delineating rooftop PV boundaries remained challenging even after transfer learning. These findings demonstrate that transfer learning substantially improves the operational deployment of pretrained GeoAI models but does not eliminate all sources of uncertainty that may influence operational decision-making.

These findings reinforce an important distinction between model adaptation and trustworthy deployment. Although transfer learning effectively mitigated geographic domain shift and substantially improved model performance, trustworthy deployment requires evidence beyond successful model adaptation alone. The persistence of residual errors and uncertainty demonstrates that deployment decisions should be informed by multiple complementary sources of operational evidence rather than predictive accuracy alone. The following section synthesises these complementary sources of evidence and discusses their implications for the responsible deployment of pretrained GeoAI models.

5.2. Synthesising Operational Trustworthiness for Pretrained GeoAI Deployment

While Section 5.1 demonstrated that transfer learning substantially improves the operational performance of pretrained GeoAI models, the results also indicate that no single performance metric is sufficient to determine whether a model is ready for operational deployment. Instead, trustworthy deployment emerged from multiple complementary sources of operational evidence that collectively characterised model behaviour under geographic domain shift. This synthesis closes the conceptual perspective introduced in Section 2 by demonstrating how established trustworthy AI principles can be operationalised for evaluating pretrained GeoAI models in real-world deployment scenarios. Rather than relying solely on conventional accuracy assessment, the present study considered robustness, reliability, generalisation, transparency, and residual uncertainty as complementary dimensions that together provide a more comprehensive understanding of deployment readiness.

Table 6 summarises how each operational trustworthiness dimension was evaluated and the principal evidence generated in this study. Robustness was demonstrated through the substantial improvement achieved after local calibration under geographic domain shift. Reliability was assessed using an independent evaluation dataset that confirmed consistent performance beyond the calibration data. Generalisation was examined by comparing performance between the development and independent datasets, where only modest reductions in performance were observed. Transparency was supported through qualitative analysis of model behaviour, including systematic examination of FPs, FNs, and geometric segmentation errors. Finally, residual uncertainty was assessed through evaluation across multiple IoU thresholds, revealing that although transfer learning substantially improved object detection, precise pixel-level delineation remained challenging under stricter geometric matching criteria.

To synthesise the findings, Table 6 revisits the operational trustworthiness dimensions introduced in Section 2 and summarises the empirical evidence generated for each dimension. Rather than representing a new trustworthiness framework, the table illustrates how established trustworthy AI principles were operationalised and evaluated within the Brisbane case study.

Table 6: Synthesis of findings across the operational trustworthiness dimensions evaluated in this study

Dimension	Operational evidence	Key finding
Robustness	Development dataset comparison	Transfer learning substantially reduced the effects of geographic domain shift.
Reliability	Independent evaluation	Performance improvements remained stable on previously unseen data.
Generalisation	Development vs independent datasets	Limited performance degradation indicated successful generalisation beyond the calibration data.
Transparency	Qualitative error analysis	Error analysis revealed strengths and limitations not captured by summary accuracy metrics.
Residual uncertainty	IoU sensitivity and geometric assessment	Boundary delineation remained a persistent source of uncertainty despite improved detection performance.

Collectively, these dimensions demonstrate that trustworthy deployment cannot be inferred from improved predictive accuracy alone. Rather, deployment decisions should be supported by multiple complementary forms of evidence that characterise how a model behaves under operational conditions. Robustness, reliability, and generalisation provide confidence that model performance extends beyond the calibration data, while transparency and explicit acknowledgement of residual uncertainty help users understand the limitations of model outputs and the circumstances under which errors may occur. From this perspective, trustworthiness is not a property that a model either possesses or lacks; instead, it reflects the extent to which sufficient evidence exists to justify confidence in model behaviour for a particular operational application.

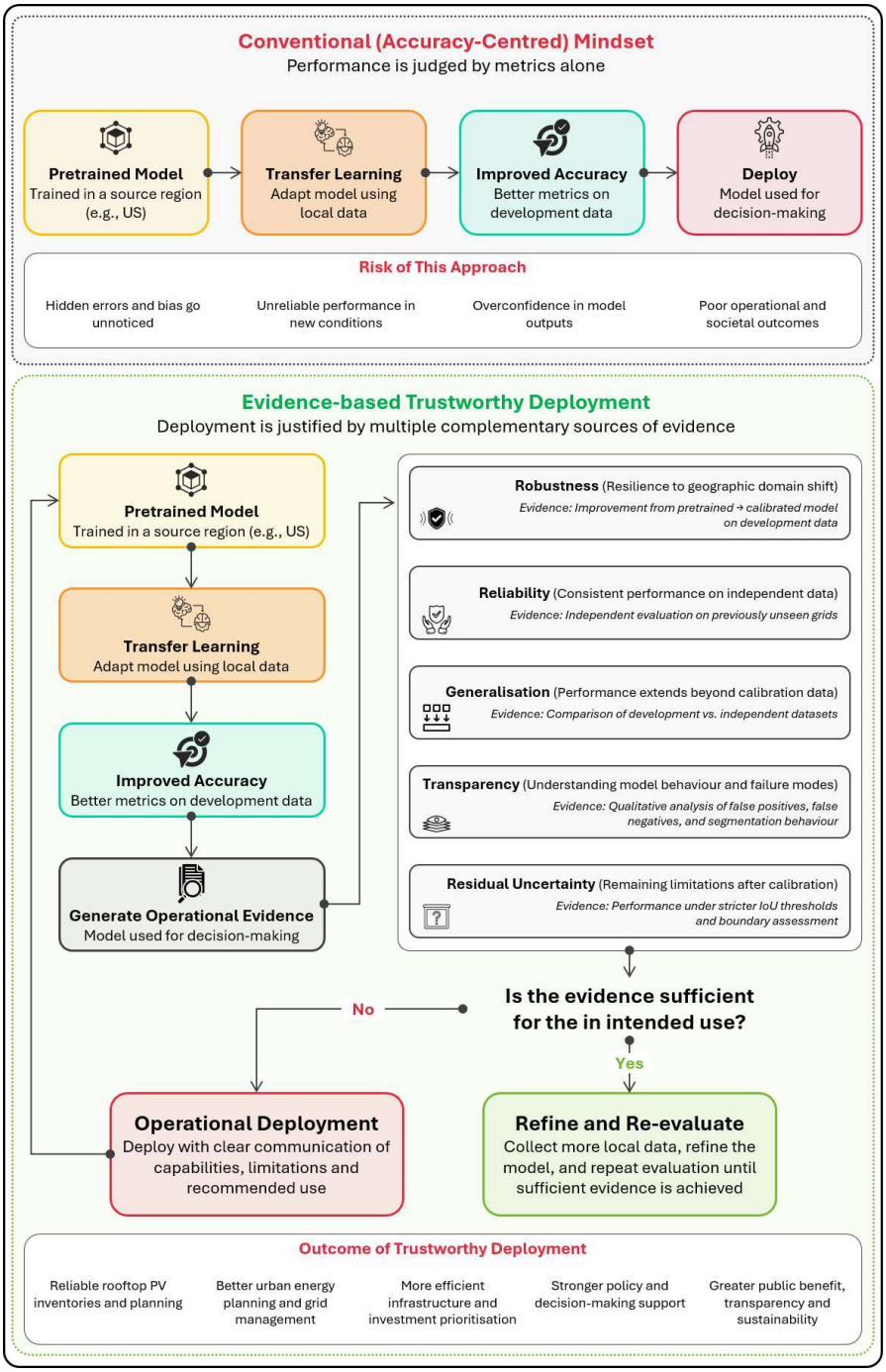


Figure 15: From accuracy-centred evaluation to evidence-based trustworthy deployment of pretrained GeoAI models

Figure 15 synthesises the principal conceptual contribution of this study. Rather than proposing a new trustworthy AI framework, the figure illustrates an evidence-based perspective for evaluating pretrained GeoAI models prior to operational deployment. Building upon established trustworthy AI principles, the conceptual model emphasises that transfer learning should be viewed as one component within a broader evidence generation process rather than the final stage of model validation. Although transfer learning substantially mitigates geographic domain shift, deployment decisions should additionally consider whether sufficient evidence has been obtained regarding robustness, reliability, generalisation, transparency, and residual uncertainty before model outputs are relied upon in operational decision-making.

This perspective has implications beyond rooftop solar PV mapping. Increasingly, pretrained GeoAI models are being applied to applications such as building extraction, land-cover classification, infrastructure monitoring, disaster assessment, and environmental change detection. As geospatial foundation models become embedded within commercial GIS platforms and cloud-based Earth observation services, practitioners will increasingly

face the challenge of determining whether pretrained models can be trusted when transferred across geographic regions. The evidence-based perspective presented here provides one practical approach for supporting these decisions by demonstrating how established trustworthy AI principles can be translated into an operational evaluation process. Although additional dimensions may be required for particular applications, the central principle remains broadly applicable: deployment decisions should be informed by multiple complementary sources of evidence rather than predictive accuracy alone.

The operational trustworthiness dimensions should not be interpreted as a checklist in which all dimensions carry equal importance. Rather, they represent complementary forms of evidence that collectively support deployment decisions. The relative importance of each dimension will depend on the application, consequences of model error, and operational context. Accordingly, this study should be viewed as demonstrating one practical approach for translating established trustworthy AI principles into operational GeoAI evaluation rather than proposing a universal trustworthiness framework.

5.3. Broader Implications for Operational GeoAI

The findings of this study have implications that extend beyond rooftop PV mapping. As pretrained GeoAI models become increasingly embedded within commercial GIS platforms, cloud-based Earth observation services, and emerging geospatial foundation models, practitioners are progressively shifting from developing models to adapting existing ones for local applications. While this transition has substantially lowered the technical barriers to artificial intelligence adoption, it has also increased the importance of understanding how pretrained models behave when transferred across geographically distinct environments. Consequently, the challenge for operational GeoAI is no longer simply improving predictive accuracy, but establishing sufficient evidence that model outputs can be trusted for the decisions they are intended to support.

The evidence generated in this study suggests that transfer learning should be considered a standard component of cross-regional GeoAI deployment rather than an optional model refinement step. However, the findings also demonstrate that local calibration alone is insufficient to justify operational deployment. Independent evaluation, transparent reporting of model behaviour, and explicit acknowledgement of residual uncertainty remain necessary for understanding the strengths and limitations of model predictions. Adopting such practices can improve confidence in GeoAI outputs while reducing the risk of deploying models that perform well under development conditions but exhibit systematic errors when applied operationally.

Although this study focused on rooftop PV detection, the same deployment challenges are likely to arise across a wide range of GeoAI applications, including building footprint extraction, land-cover classification, infrastructure monitoring, disaster assessment, environmental change detection, and ecological mapping. As geospatial foundation models continue to mature and pretrained models become increasingly reusable across applications and geographic regions, operational evaluation frameworks that integrate predictive performance with robustness, reliability, transparency, and uncertainty assessment are likely to become increasingly important. Future research should therefore focus on developing standardised protocols for evaluating pretrained GeoAI models across diverse geographic contexts and investigating additional evidence sources, such as uncertainty quantification, model explainability, and continuous post-deployment monitoring, to further support responsible operational AI.

Ultimately, this study suggests that the future of operational GeoAI will depend not only on increasingly powerful pretrained models but also on increasingly rigorous methods for evaluating their deployment. From this perspective, the central question is no longer "Can a pretrained GeoAI model be transferred to a new geographic region?" but rather "Has sufficient evidence been generated to justify trusting the model under its intended operational conditions?" Answering this question will be fundamental to ensuring that GeoAI systems support reliable, transparent, and responsible spatial decision-making.

6. Conclusion

The increasing availability of pretrained GeoAI models is fundamentally changing how artificial intelligence is adopted within geospatial practice. By lowering the technical barriers to deep learning, commercial GIS platforms and emerging geospatial foundation models are enabling practitioners to apply sophisticated AI models to a growing range of operational applications. As these models are increasingly transferred across geographic regions, however, the challenge is no longer limited to achieving high predictive accuracy. Rather, the central challenge is determining whether sufficient evidence exists to justify trusting model outputs under the environmental, geographic, and operational conditions in which they will be used.

This study demonstrates that transfer learning is an effective strategy for adapting pretrained GeoAI models to geographically distinct environments. Using rooftop PV mapping in Brisbane, Australia, as a representative case study, local calibration substantially improved robustness to geographic domain shift, enhanced reliability on previously unseen data, and supported effective model generalisation beyond the calibration dataset. At the same time, the independent evaluation and qualitative error analysis revealed persistent FPs, missed detections, and geometric segmentation errors, particularly under stricter spatial matching criteria. These findings demonstrate that while transfer learning substantially improves operational performance, improved predictive accuracy alone should not be interpreted as sufficient evidence for trustworthy deployment.

The principal contribution of this research is therefore not the development of a new deep learning architecture or the proposal of a new trustworthy AI framework. Instead, the study demonstrates how established trustworthy AI principles can be operationalised as complementary sources of evidence for evaluating the deployment of pretrained GeoAI models. By integrating evidence relating to robustness, reliability, generalisation, transparency, and residual uncertainty, the study presents an evidence-based operational perspective that extends beyond conventional accuracy assessment. This perspective encourages practitioners to evaluate not only how accurately a model performs, but also whether sufficient evidence exists to justify confidence in its behaviour before it is deployed to support planning, policy, or infrastructure decisions.

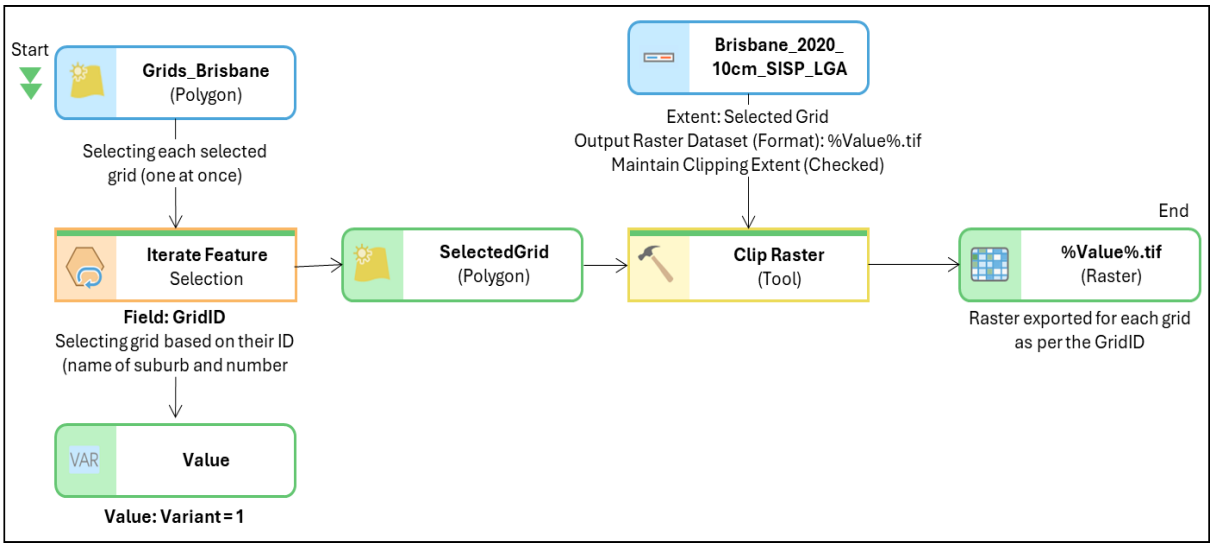
The findings also suggest several practical recommendations for operational GeoAI deployment. First, transfer learning should be regarded as the beginning rather than the end of the model validation process whenever pretrained models are transferred across geographic regions. Second, independent evaluation using geographically separate datasets should become standard practice for assessing operational reliability and generalisation. Finally, predictive performance should be complemented by transparent assessment of model behaviour, failure modes, and residual uncertainty to enable more informed and responsible deployment decisions. Collectively, these practices provide a more comprehensive basis for evaluating deployment readiness than predictive accuracy alone.

Although this study focused on rooftop solar PV detection, the broader implications extend to many GeoAI applications where pretrained models are increasingly reused across geographic contexts, including building extraction, land-cover mapping, infrastructure monitoring, disaster assessment, and environmental change detection. Future research should evaluate the evidence-based operational perspective presented in this study across diverse application domains, geographic regions, and emerging geospatial foundation models. Additional work should also investigate complementary sources of operational evidence, including uncertainty quantification, explainability, fairness assessment, and continuous post-deployment monitoring, to support more comprehensive evaluation of trustworthy GeoAI systems.

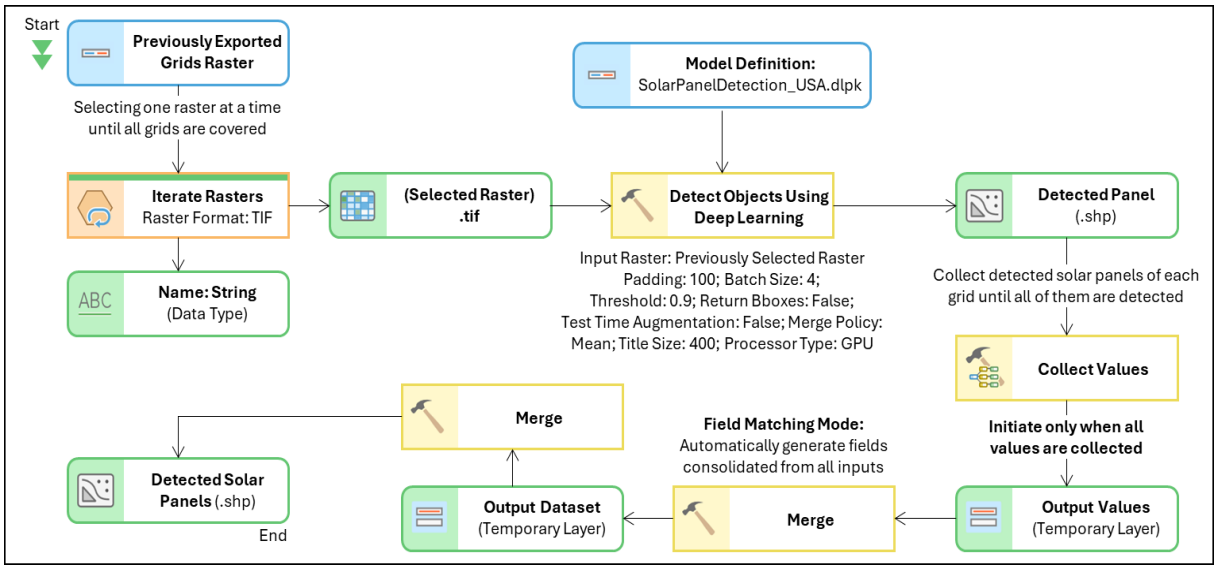
Ultimately, the future of operational GeoAI will depend not only on increasingly powerful pretrained models but also on increasingly rigorous approaches for evaluating their deployment. As artificial intelligence becomes embedded within routine geospatial decision-making, the central question is no longer whether pretrained models can be transferred across geographic regions, but whether sufficient operational evidence has been generated to justify trusting their outputs. Developing evidence-based deployment practices will therefore be fundamental to ensuring that GeoAI supports reliable, transparent, and responsible spatial decision-making.

Appendices

Appendix A: ModelBuilder workflow for clipping raster tiles per grid cell in ArcGIS



Appendix B: Automated ModelBuilder workflow used to apply the US pretrained model to the Brisbane raster tiles for baseline inference



References

Ahmed, S., El-Shazely, A., & Ahmed, W. (2025). Deep learning for building footprint extraction using UAV-based orthoimages. Journal of the Indian Society of Remote Sensing, 53(4), 1243-1262.

Alhawsawi, A., Khan, S., & Rehman, F. (2025). Deep learning framework for crowd congestion detection in smart cities via encoding motion irregularities using recurrence plots. Journal of King Saud University Computer and Information Sciences, 37(8), 258.

Andorful, F., Herfort, B., Melanda, E. A., Antonio, N. D., Zipf, A., & Camboim, S. P. (2026). Transparency and trust in collaborative mapping: Concerns and dilemmas in AI-assisted road integration within OpenStreetMap. Annals of the American Association of Geographers, 116(3), 675-696.

Arribas, L., Lechón, Y., Perula, A., Domínguez, J., Ferres, M., Navarro, J., Zarzalejo, L., García Barquero, C., & Cruz, I. (2021). Review of data and data sources for the assessment of the potential of utility-scale hybrid wind–solar PV power plants deployment, under a microgrid scope. Energies, 14(21), 7434.

Best, R., & Chareunsy, A. (2022). The impact of income on household solar panel uptake: Exploring diverse results using Australian data. Energy Economics, 112, 106124.

Best, R., Esplin, R., Hammerle, M., Nepal, R., & Reynolds, Z. (2023). Do solar panels increase housing rents in Australia? Housing Studies, 38(10), 1918-1935.
Boulmay, B. (2022). Introducing pretrained deep learning models for energy. Retrieved 30 January 2026 from https://www.esri.com/en-us/industries/blog/articles/introducing-pretrained-deep-learning-models-for-energy#:~:text=1,Africa%20and%20Australia%20as%20well
Chen, X., Li, Z., Peter, D., & Slowik, A. (2023). Circular economy oriented future building information processing: PSO for CNN approach. Applied Soft Computing, 149, 9, 111013.
Choi, Y. (2023). GeoAI: Integration of artificial intelligence, machine learning, and deep learning with GIS. Applied Sciences, 13(6), 3895.
Clark, C., & Pacifici, F. (2023). A solar panel dataset of very high resolution satellite imagery to support the Sustainable Development Goals. Scientific Data, 10(1), 636.
Coglan, J., Gharineiat, Z., & Tarsha Kurdi, F. (2025). Automatic rooftop solar panel recognition from UAV LiDAR data using deep learning and geometric feature analysis. Remote Sensing, 17(19), 3389.
Doan, T., Wilson, D., Rashford, S., & Bosley, E. (2021). Ambient temperatures, heatwaves and out-of-hospital cardiac arrest in Brisbane, Australia. Occupational and Environmental Medicine, 78(5), 349.
Energy AU. (2024). Australia hits rooftop solar milestone. Department of Climate Change, Energy, the Environment and Water, Australian Government. Retrieved 30 January 2026 from https://www.energy.gov.au/news/australia-hits-rooftop-solar-milestone#:~:text=Rooftop%20solar%20in%20Australia%20has,million%20installations%20across%20the%20country
Epasto, S. (2025). Spatializing trust: A GeoAI-based model for mapping digital trust ecosystems in Mediterranean smart regions. ISPRS International Journal of Geo-Information, 14(12), 491.
Esri. (2023). Solar panel detection - New Zealand. Retrieved 1 February 2026 from https://qutaus.maps.arcgis.com/home/item.html?id=23d46b7e7f7d41abae01885b64834af8
Esri. (2025a). Introduction to the model (ArcGIS pretrained models). Retrieved 31 January 2026 from https://doc.arcgis.com/en/pretrained-models/latest/imagery/introduction-to-solar-panel-detection-usa.htm#ESRI_SECTION1_422692F6E1BB4019A8F7F0C69B93F2F6
Esri. (2025b). Solar panel detection - USA. Retrieved 1 February 2026 from https://qutaus.maps.arcgis.com/home/item.html?id=c2508d72f2614104bfcfd5ccf1429284
Esri. (2026). How compute accuracy for object detection works (ArcGIS Pro 3.6). Retrieved 20 January 2026 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/image-analyst/how-compute-accuracy-for-object-detection-works.htm
Esri Developer. (2026). How mask R-CNN works? (ArcGIS API for Python). Esri. Retrieved 2 February 2026 from https://developers.arcgis.com/python/latest/guide/how-maskrcnn-works/
European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union.
Farahani, L., Izadpanahi, P., & Tucker, R. (2022). The death and life of Australian suburbs: Relationships between social activity and the physical qualities of Australian suburban neighbourhood centres. City, Culture and Society, 28, 100426.
Franklyn, J. (2023). Detecting solar panels in imagery using deep learning. Retrieved 1 February from https://storymaps.arcgis.com/stories/b9de83bf40644200a946ec5575efb56d
García, G., Aparcedo, A., Nayak, G., Ahmed, T., Shah, M., & Li, M. (2024). Generalized deep learning model for photovoltaic module segmentation from satellite and aerial imagery. Solar Energy, 274, 112539.
Ghahremani, A., Adams, S., Norton, M., Khoo, S., & Kouzani, A. (2025). Advancements in AI-Driven detection and localisation of solar panel defects. Advanced Engineering Informatics, 64, 103104.
Ghamisi, P., Yu, W., Marinoni, A., Gevaert, C. M., Persello, C., Selvakumaran, S., Girotto, M., Horton, B. P., Rufin, P., Hostert, P., Pacifici, F., & Atkinson, P. M. (2025). Responsible artificial intelligence for earth observation: Achievable and realistic paths to serve the collective good. IEEE Geoscience and Remote Sensing Magazine, 13(3), 72-96.
Ghimire, S., Bhandari, B., Casillas-Pérez, D., Deo, R., & Salcedo-Sanz, S. (2022). Hybrid deep CNN-SVR algorithm for solar radiation prediction problems in Queensland, Australia. Engineering Applications of Artificial Intelligence, 112, 104860.
Haynes, S., Johnston, P., & Elyan, E. (2024). Generalisation challenges in deep learning models for medical imagery: Insights from external validation of COVID-19 classifiers. Multimedia Tools and Applications, 83(31), 76753-76772.
Hossain, T., Yigitcanlar, T., Nguyen, K., & Xu, Y. (2025). Platform urbanism for resident safety: A real-time predictive microclimate risk monitoring and alert system. Urban Climate, 61, 102445.

Hou, C., Zhang, F., Kang, Y., Gao, S., Li, Y., Duarte, F., & Li, S. (2025). Transferred bias uncovers the balance between the development of physical and socioeconomic environments of cities. Annals of the American Association of Geographers, 115(1), 148-166.
Hu, Y., Goodchild, M., Zhu, A. X., Yuan, M., Aydin, O., Bhaduri, B., Gao, S., Li, W., Lunga, D., & Newsam, S. (2024). A five-year milestone: Reflections on advances and limitations in GeoAI research. Annals of GIS, 30(1), 1-14.
Huang, X., Tu, Z., Ye, X., & Goodchild, M. (2026). The role of open-source LLMs in shaping the future of GeoAI. Annals of GIS, 1-10.
Hui, J. (2018). mAP (mean average precision) for object detection. Medium. Retrieved 22 January 2026 from https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173
ISO. (2023). ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. I. O. f. Standardization. https://www.iso.org/standard/77304.html?__cf_chl_f_tk=RxRgwjgrgPMQVuwDKEIw9ETPS4birvBMJPcslhseqhM-1783081743-1.0.1.1-WNWSabGh4Jrqb_RkfffdSQriDwKUrhZnRiZ7hnoFpeQ
Jagatheesaperumal, S. K., Bibri, S. E., Huang, J., Rajapandian, J., & Parthiban, B. (2024). Artificial intelligence of things for smart cities: advanced solutions for enhancing transportation safety. Computational Urban Science, 4(1), 10.
Jiao, D., & Sun, J. (2018). Real-time visualization of geo-sensor data based on the protocol-coupling symbol construction method. ISPRS International Journal of Geo-Information, 7(12), 460.
Kasmi, G., Dubus, L., Saint-Drenan, Y., & Blanc, P. (2025). Space-scale exploration of the poor reliability of deep learning models: The case of the remote sensing of rooftop photovoltaic systems. Environmental Data Science, 4, e22.
Kaya, O., Klepacka, A., & Florkowski, W. (2019). Achieving renewable energy, climate, and air quality policy goals: Rural residential investment in solar panel. Journal of Environmental Management, 248, 109309.
Li, F., Yigitcanlar, T., Nepal, M., Nguyen, K., Dur, F., & Li, W. (2024). Assessing heat vulnerability and multidimensional inequity: Lessons from indexing the performance of Australian capital cities. Sustainable Cities and Society, 115, 105875.
Li, W., Lee, H., Wang, S., Hsu, C.-Y., & Arundel, S. T. (2023). Assessment of a new GeoAI foundation model for flood inundation mapping Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Hamburg, Germany.
Li, W., Yigitcanlar, T., Browne, W., & Nili, A. (2023). The making of responsible innovation and technology: An overview and framework. Smart Cities, 6(4), 1996-2034.
Long, Y., Xu, X., & Huo, Z. (2025). Urban rooftop photovoltaic potential model: A study on assessment methods and model framework. Energy and Buildings, 345, 116138.
Lu, F., Cheng, S., & Wang, P. (2025). GeoAI enabled urban computing: Status and challenges. Annals of GIS, 31(4), 537-555.
Manso-Callejo, M., Cira, C., Arranz-Justel, J., Sinde-González, I., & Sălăgean, T. (2023). Assessment of the large-scale extraction of photovoltaic (PV) panels with a workflow based on artificial neural networks and algorithmic postprocessing of vectorization results. International Journal of Applied Earth Observation and Geoinformation, 125, 103563.
Narayanan, R., Parthkumar, P., & Pippia, R. (2021). Solar energy utilisation in Australian homes: A case study. Case Studies in Thermal Engineering, 28, 101603.
Nguyen, H., Dung, T., Hung, T., & Tuan, H. (2025). Deep learning-based segmentation of rooftop photovoltaic panels in Ho Chi Minh City, Vietnam. 2025 17th International Conference on Knowledge and System Engineering, IEE Xplore.
NIST. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U. S. D. o. C. National Institute of Standards and Technology (NIST).
OECD. (2019). OECD AI Principles overview. Organisation for Economic Co-operation and Development. Retrieved 3 July 2026 from https://oecd.ai/en/ai-principles
Padilla, R., Netto, S., & Silva, E. (2020). A survey on performance metrics for object-detection algorithms. 2020 International Conference on Systems, Signals and Image Processing, IEEE Xplore.
Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.
Parsaeifar, R., Valinejadshoubi, M., Le Guen, A., & Valdivieso, F. (2025). AI-based solar panel detection and monitoring using high-resolution drone imagery. Journal of Soft Computing in Civil Engineering, 9(3), 41-59.
Peronato, G., Rastogi, P., Rey, E., & Andersen, M. (2018). A toolkit for multi-scale mapping of the solar energy-generation potential of buildings in urban environments under uncertainty. Solar Energy, 173, 861-874.

Queensland Government. (2025). Queensland imagery: Latest state program (all users). Retrieved 15 February 2026 from https://www.data.qld.gov.au/dataset/queensland-imagery-latest-state-program-public-basemap-service
Sadeq, M., Muhammad-Sukki, F., Ullah, M., & Sellami, N. (2025). AI-driven BIPV potential prediction: Integrating deep learning and PVGIS for accelerating building decarbonisation in the UK. Springer.
Shaamala, A., Yigitcanlar, T., Nili, A., & Nyandega, D. (2025). Machine learning applications for urban geospatial analysis: A review of urban and environmental studies. Cities, 165, 106139.
Shi, M., Janowicz, K., Verstegen, J., Currier, K., Wiedemann, N., Mai, G., Majic, I., Liu, Z., & Zhu, R. (2025). Geography for AI sustainability and sustainability for GeoAI. Cartography and Geographic Information Science, 52(4), 331-349.
Simin, L., & Michael James, H. (2024). Climate change adaptation strategies in coastal cities of Australia: A comparative study of Sydney, Melbourne, and Brisbane. Studies in Social Science & Humanities, 3(5), 64-71.
Singh, R., & Viswambharan, V. (2021). Pretrained deep learning models update (July 2021). Retrieved 31 January 2026 from https://www.esri.com/arcgis-blog/products/arcgis-living-atlas/announcements/pre-trained-deep-learning-models-update-july-2021
Szukalski, B. (2024). Overview of pretrained AI models in ArcGIS Living Atlas. Retrieved 25 September 2025 from https://community.esri.com/t5/arcgis-living-atlas-blog/overview-of-pretrained-ai-models-in-arcgis-living/ba-p/1367078
Taminiau, J., & Byrne, J. (2020). City-scale urban sustainability: Spatiotemporal mapping of distributed solar power for New York City. WIREs Energy and Environment, 9(5), e374.
Tayal, D. (2017). Achieving high renewable energy penetration in Western Australia using data digitisation and machine learning. Renewable and Sustainable Energy Reviews, 80, 1537-1543.
Viswambharan, V., & Singh, R. (2020). Introducing pretrained geospatial deep learning models. Retrieved 29 January 2026 from https://www.esri.com/arcgis-blog/products/arcgis/imagery/introducing-ready-to-use-deep-learning-models
Viswambharan, V., & Singh, R. (2025). Pretrained nodels in ArcGIS: Comparing task-specific and generalized vision-language models. Retrieved 26 September 2025 from https://www.esri.com/arcgis-blog/products/arcgis/geoai/task-and-generalized-pretrained-models
Waqas, M., & Humphries, U. W. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX, 13, 102946.
Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. Journal of Big Data, 3(1), 9.
Wu, A.., & Biljecki, F. (2021). Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. Landscape and Urban Planning, 214, 104167.
Wu, P., Zhang, Z., Peng, X., & Wang, R. (2024). Deep learning solutions for smart city challenges in urban development. Scientific Reports, 14(1), 5176.
Yasmeen, R., Yao, X., Ul Haq Padda, I., Shah, W., & Jie, W. (2022). Exploring the role of solar energy and foreign direct investment for clean environment: Evidence from top 10 solar energy consuming countries. Renewable Energy, 185, 147-158.
Yu, J., Wang, Z., Majumdar, A., & Rajagopal, R. (2018). DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. Joule, 2(12), 2605-2617.
Zander, K. (2020). Unrealised opportunities for residential solar panels in Australia. Energy Policy, 142, 111508.
Zhang, W., Wang, L., Cheng, S., & Lu, F. (2026). Ensemble learning for spatiotemporal data: Methods, applications, challenges and opportunities. International Journal of Geographical Information Science, 1-49.
Zhang, X., Yang, F., He, Z., Li, W., & Yang, M. (2026). City identity recognition: how representation bias influences model predictability and replicability? Computers, Environment and Urban Systems, 123, 102370.