Journal of Geovisualization and Spatial Analysis (2026) 10:51
https://doi.org/10.1007/s41651-026-00293-0

RESEARCH

Trustworthy Transfer Learning in GeoAI: Assessing Pretrained Models for Rooftop Solar PV Mapping Under Geographic Domain Shift

Tahsin Hossain[1] · **Tan Yigitcanlar**[1,2] · **Niklas Tilly**[1] · **Filip Biljecki**[3]

Received: 24 July 2026 / Accepted: 24 October 2026
© The Author(s) 2026

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 independent local test datasets within Brisbane. Local calibration substantially improved detection performance, increasing recall from 38.5% to 84.1% on the development dataset, with similar improvements observed on the independent 10-grid test dataset, indicating effective local adaptation within the sampled Brisbane environments. However, independent evaluation and qualitative error analysis revealed persistent false positives, missed detections, and geometric segmentation errors, highlighting residual uncertainty and the limitations of model performance within the evaluated geographic context. These findings suggest that transfer learning can substantially improve detection performance and provide useful evidence for assessing operational trustworthiness, although predictive improvements alone are insufficient to establish trustworthy deployment. By integrating complementary evidence relating to robustness, reliability, generalisation, transparency, and residual uncertainty, this study provides an evidence-based operational approach for assessing pretrained GeoAI models and informing responsible deployment decisions in geographically distinct environments.

Keywords GeoAI · Pretrained models · Transfer learning · Model trustworthiness · Rooftop solar PV mapping · Geographic domain shift · Solar panel detection

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 and Sun 2018; Choi 2023; Marasinghe et al. 2024; Hossain et al. 2025). GeoAI integrates artificial intelligence with Geographic Information Systems (GIS) and remote sensing to automate the analysis of large geospatial datasets. 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

✉ Tan Yigitcanlar
tan.yigitcanlar@qut.edu.au

Tahsin Hossain
tahsin.hossain@qut.edu.au

Niklas Tilly
niklas.tilly@hdr.qut.edu.au

Filip Biljecki
filip@nus.edu.sg

[1] Queensland University of Technology, Brisbane, Australia

[2] University of Johannesburg, Johannesburg, South Africa

[3] National University of Singapore, Singapore, Singapore

Published online: 29 October 2026

models from scratch, practitioners can now apply pretrained models to mapping and decision-making tasks across government, industry, and research (Hu et al. 2024; Nagy 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; Zhang et al. 2026a, b). 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 and 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. 2023a, b; Shaamala et al. 2025; Viswambharan and Singh 2025; Zhang et al. 2026a, b).

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 and Biljecki 2021; Boulmay 2022; Haynes et al. 2024; Kasmi et al. 2025). Geographic transfer may therefore produce false positives, missed detections, and spatial inconsistencies, introducing uncertainty into downstream analyses and decisions. Recognising these limitations, developers of operational GeoAI models, including Esri, recommend local calibration or retraining before deployment in new geographic environments (Viswambharan and Singh 2020; Singh and 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 and Biljecki 2021; Yasmeen et al. 2022; Lodhi et al. 2025). Accurate rooftop PV inventories support energy planning, network management, renewable energy policy, and climate resilience (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 and 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 2024). This growth has increased demand for spatially detailed PV inventories to support network planning, infrastructure investment, and urban sustainability (Zander 2020; Best and 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. Brisbane therefore provides a relevant urban case study for evaluating GeoAI-based rooftop PV mapping within a mature and rapidly evolving distributed solar environment.

This setting is also useful for evaluating geographic domain shift and the transferability of pretrained GeoAI models. 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). Esri's pretrained US Solar Panel Detection model illustrates this approach and has supported regional adaptation, including New Zealand (Esri 2023, 2025a). Applying this model to Brisbane provides an opportunity to examine local adaptation under geographic domain shift. It therefore raises an important practical question: can pretrained GeoAI models be trusted when deployed beyond the geographic environments for which they were originally developed?

Previous rooftop PV detection research has demonstrated geographic generalisation challenges and the benefits of incorporating target-region training data (Ranalli

et al. 2024). Similarly, transfer learning studies have primarily emphasised improving predictive performance through local model adaptation (Weiss et al. 2016; Boulmay 2022; Franklyn 2023; Nguyen et al. 2025). In parallel, trustworthy AI research highlights the importance of robustness, reliability, transparency, uncertainty, and accountability in operational deployment (OECD 2019; NIST 2023; Ghamisi et al. 2025). However, comparatively limited attention has been given to integrating these perspectives in the evaluation of pretrained GeoAI models, particularly by examining whether improvements following transfer learning are maintained on independent local datasets and what residual errors and uncertainties remain. This study addresses that gap by combining pre- and post-transfer learning performance comparisons, independent evaluation, and qualitative error analysis to assess multiple dimensions of operational trustworthiness beyond predictive accuracy alone.

To address this gap, this study examines the operational trustworthiness of a pretrained GeoAI model through rooftop PV mapping in Brisbane, Australia. It investigates whether local adaptation through transfer learning provides sufficient evidence to inform trustworthy deployment decisions under geographic domain shift. 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 an empirical assessment of pretrained GeoAI model trustworthiness following transfer learning under geographic domain shift, rather than the development of a new detection algorithm or trustworthiness framework. Specifically, the study makes three contributions: (a) evaluating the effectiveness of local model adaptation through independent performance assessment; (b) integrating evidence of robustness, reliability, generalisation, transparency, and residual uncertainty within an operational GeoAI workflow; and (c) demonstrating how these complementary assessments can inform responsible deployment decisions beyond conventional predictive accuracy.

The remainder of this paper is organised as follows. Section "Literature Background" reviews the relevant literature and establishes the operational trustworthiness perspective. Section "Research Design" presents the research design and methodology, Section "Analysis and Results" reports the results, Section "Findings and Discussions" discusses their broader implications, and Section "Conclusion" concludes the study.

Literature Background

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 and 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 and Biljecki 2021). Variations in building architecture, rooftop materials, vegetation, illumination conditions, sensor characteristics, spatial resolution, 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; Deng et al. 2026). 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.

Geographic generalisation limitations in rooftop PV detection have been documented in previous research. Ranalli et al. (2024) systematically evaluated neural network models using six aerial imagery datasets from the United States, France, and Germany. They found that diversifying training data improved average performance on some unseen datasets but did not achieve the performance of models trained with target-region data. Their experiments further demonstrated that incorporating relatively small amounts of local training data could substantially improve detection performance. Building on these findings, the present study shifts the emphasis from improving cross-regional generalisation to assessing whether local adaptation of a publicly available pretrained GeoAI model provides sufficient evidence for trustworthy operational deployment.

Beyond geographic generalisation, recent rooftop PV detection research has investigated strategies for improving model adaptation to new regions. Zech et al. (2024) evaluated deep active learning across six aerial imagery datasets, comparing uncertainty-based and diversity-based sampling approaches for PV segmentation, including fine-tuning and joint-learning scenarios. Their results showed that entropy-based sample selection substantially reduced annotation requirements, achieving performance comparable to full-data training with approximately 3% of the available images in their imbalanced-data experiments. These findings demonstrate the potential of targeted sample selection for efficient model adaptation. In contrast, the present study focuses on evaluating the operational trustworthiness of a locally fine-tuned pretrained GeoAI model rather than optimising annotation efficiency through active learning.

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 and Yang 2010; Weiss et al. 2016). Recent studies have demonstrated that pretrained representations combined with local or few-shot adaptation can substantially improve PV detection and segmentation performance, particularly where labelled target-domain data are limited (Boulmay 2022; Franklyn 2023; Nguyen et al. 2025; Deng et al. 2026). 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.

Taken together, these studies address related but distinct challenges in rooftop PV mapping. Ranalli et al. (2024) examined geographic generalisation and the contribution of target-region training data, whereas Zech et al. (2024) investigated how active learning could reduce annotation requirements during model adaptation. These contributions provide important evidence regarding model transferability and adaptation efficiency, but they do not directly address the broader question examined here: how multiple forms of evaluation evidence can be combined to assess the operational trustworthiness of a publicly available pretrained GeoAI model after local calibration. The present study therefore complements these approaches by examining independent local performance, observable failure modes, and residual segmentation uncertainty alongside conventional detection metrics.

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.

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 23,894 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 established trustworthy AI principles, this study adopts an operational trustworthiness perspective for evaluating pretrained GeoAI models under geographic domain shift, rather than proposing a new theoretical framework. As illustrated in Fig. 1, a pretrained model transferred to a geographically distinct environment may require local calibration through transfer learning. However, improved predictive performance alone does not establish trustworthy deployment. Instead, deployment decisions should be informed by complementary evidence relating to robustness, reliability, generalisation, transparency, and residual uncertainty. These dimensions provide the conceptual basis for interpreting the empirical findings of this study.

Table 1 operationalises the conceptual assessment framework presented in Fig. 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.

For the purposes of this study, these five dimensions are distinguished by the specific aspect of model behaviour they examine. In this study, robustness is considered in relation to geographic domain shift; however, the empirical comparison primarily evaluates the effectiveness of local adaptation through transfer learning rather than robustness in the conventional sense of maintaining stable performance across multiple domains or perturbations. Reliability concerns the detection performance demonstrated on an independent local test dataset excluded from model development. Generalisation refers more specifically to the extent to which performance improvements observed on the development dataset are maintained on previously unseen Brisbane imagery. Transparency concerns the interpretability of observable model behaviour through confidence-score inspection and analysis of successful detections, false positives, false negatives, and segmentation errors, rather than explanation of the model's internal decision processes. Residual uncertainty refers to the remaining detection and geometric limitations after local calibration, examined through error analysis and sensitivity to increasingly stringent IoU thresholds. Although these dimensions are related and may draw on overlapping empirical evidence, each addresses a distinct evaluation question. Together, they provide complementary evidence relevant to operational trustworthiness rather than a definitive or universal measure of deployment readiness.

Collectively, the conceptual framework (Fig. 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.

Research Design

Consistent with the operational trustworthiness assessment framework introduced in Section "Literature Background" (Fig. 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.

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

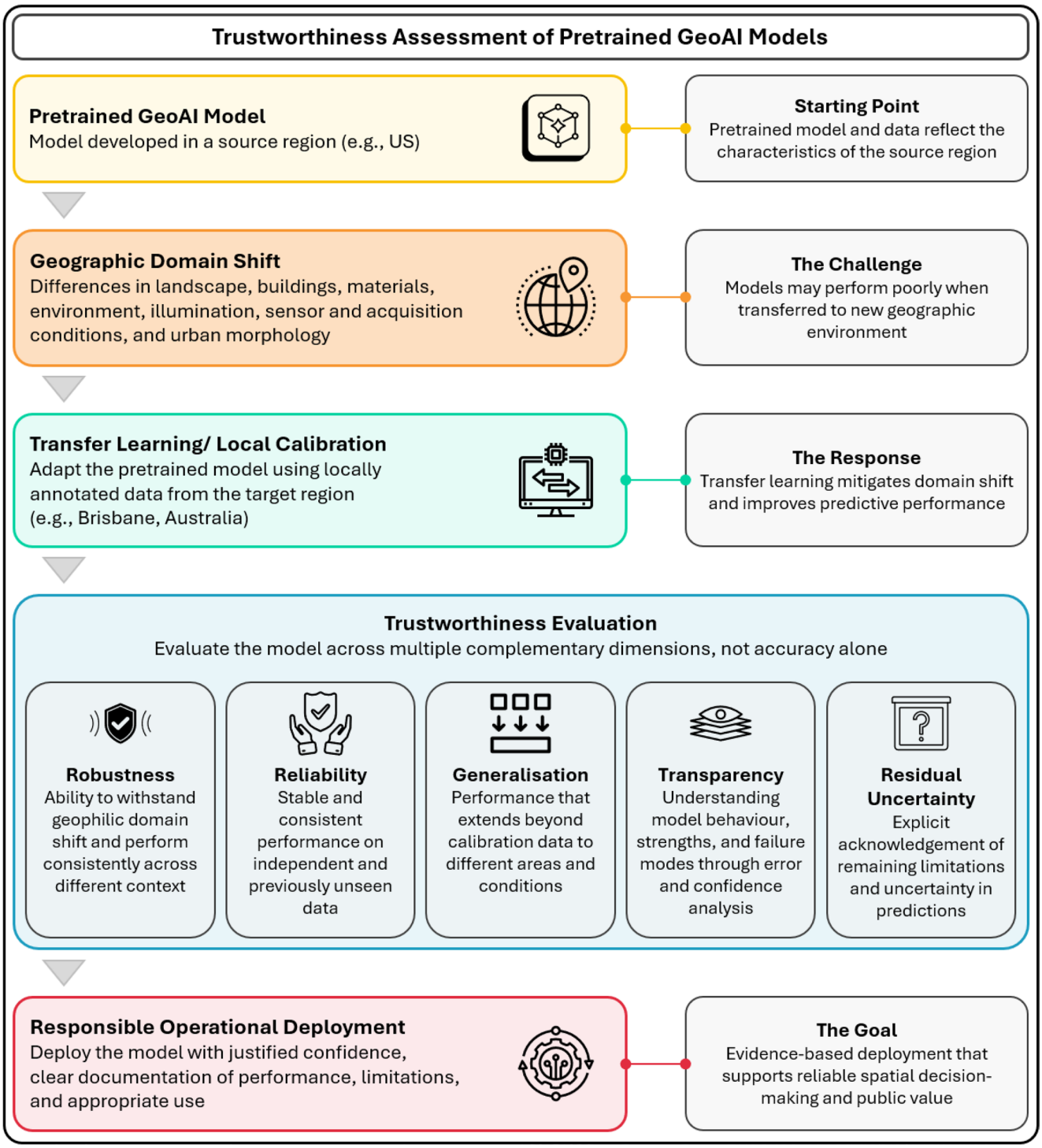


Fig. 1 Conceptual trustworthiness assessment framework adopted in this study for evaluating pretrained GeoAI models under geographic domain shift

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

Table 1 Operational trustworthiness dimensions evaluated in this study

Trustworthiness dimension	Operational interpretation	Evidence used in the study	References
Robustness	Sensitivity to geographic domain shift and improvement following local calibration	Pretrained versus locally calibrated model comparison in Brisbane (adaptation-related evidence)	(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 and 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. 2023a, b; Andorful et al. 2026; Huang et al. 2026)
Residual uncertainty	Explicit acknowledgement of remaining model limitations	Intersection over Union (IoU) sensitivity analysis and segmentation quality assessment	(Hu et al. 2024; W. Zhang et al. 2026a, b)
Responsible deployment	Evidence supporting operational adoption	Practical recommendations derived from empirical findings	(Wenda Li et al. 2023a, b; Marasinghe et al. 2024; Shi et al. 2025)

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.

Study Area and Dataset Preparation

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 and 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 was guided by variation in urban land use, building density, rooftop morphology, rooftop PV installation configurations, and environmental conditions likely to affect image-based detection. Specifically, the selected cells encompassed dense inner-city rooftopscapes, predominantly residential areas with pitched roofs, industrial and commercial buildings with extensive rooftop arrays, coastal and bayside environments, leafy and hilly suburbs with tree shading and variable roof orientations, and outer urban areas with relatively uniform building designs. The purpose of this selection was to capture contrasting local detection conditions rather than to obtain a statistically representative sample of all Brisbane grid cells.

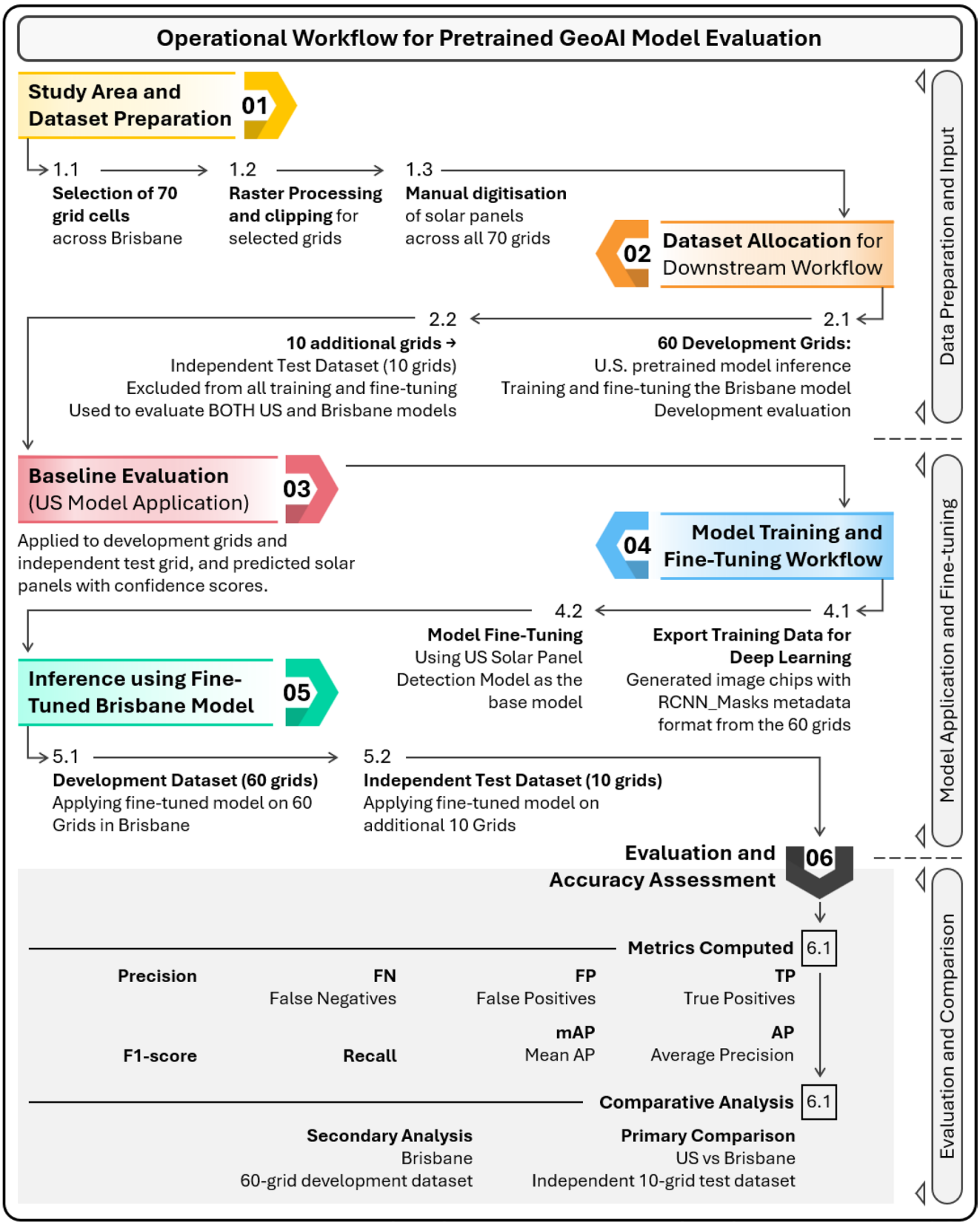


Fig. 2 Methodological workflow adopted to operationalise the trustworthiness assessment framework

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
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

The spatial distribution of the selected grid cells, together with representative imagery and examples of manually digitised rooftop PV installations, is presented in Fig. 3.

Although purposive stratified sampling enabled the inclusion of diverse urban typologies, the non-random selection of grid cells limits the statistical representativeness of the evaluation datasets. Consequently, the reported accuracy metrics reflect model performance within the selected urban environments and should not be interpreted as unbiased estimates of performance across the entire Brisbane LGA.

Raster Processing and Digitisation

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 66 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.

The manual annotation procedure was based on visual interpretation of the 10 cm RGB aerial imagery. Rooftop PV features were identified from their visible appearance and spatial arrangement, and polygons were delineated to represent the observable extent of each installation feature. Where adjacent modules formed a visually continuous array, they were represented as a grouped polygon rather than being systematically annotated as separate modules. Accordingly, the reference dataset represents visually identifiable rooftop PV installation features, which may comprise individual panels or grouped arrays, rather than a complete inventory of individual photovoltaic modules. The interpretation of installation boundaries may be less certain where panels are partially obscured, closely spaced, or affected by shadows and image characteristics. Following digitisation, the reference polygons were visually reviewed against the source aerial imagery to identify and correct apparent annotation errors and inconsistencies in the delineation of rooftop PV installation features.

Figure 3 illustrates examples of the resulting reference annotations across different rooftop PV configurations. This method aligns with annotation approaches adopted in

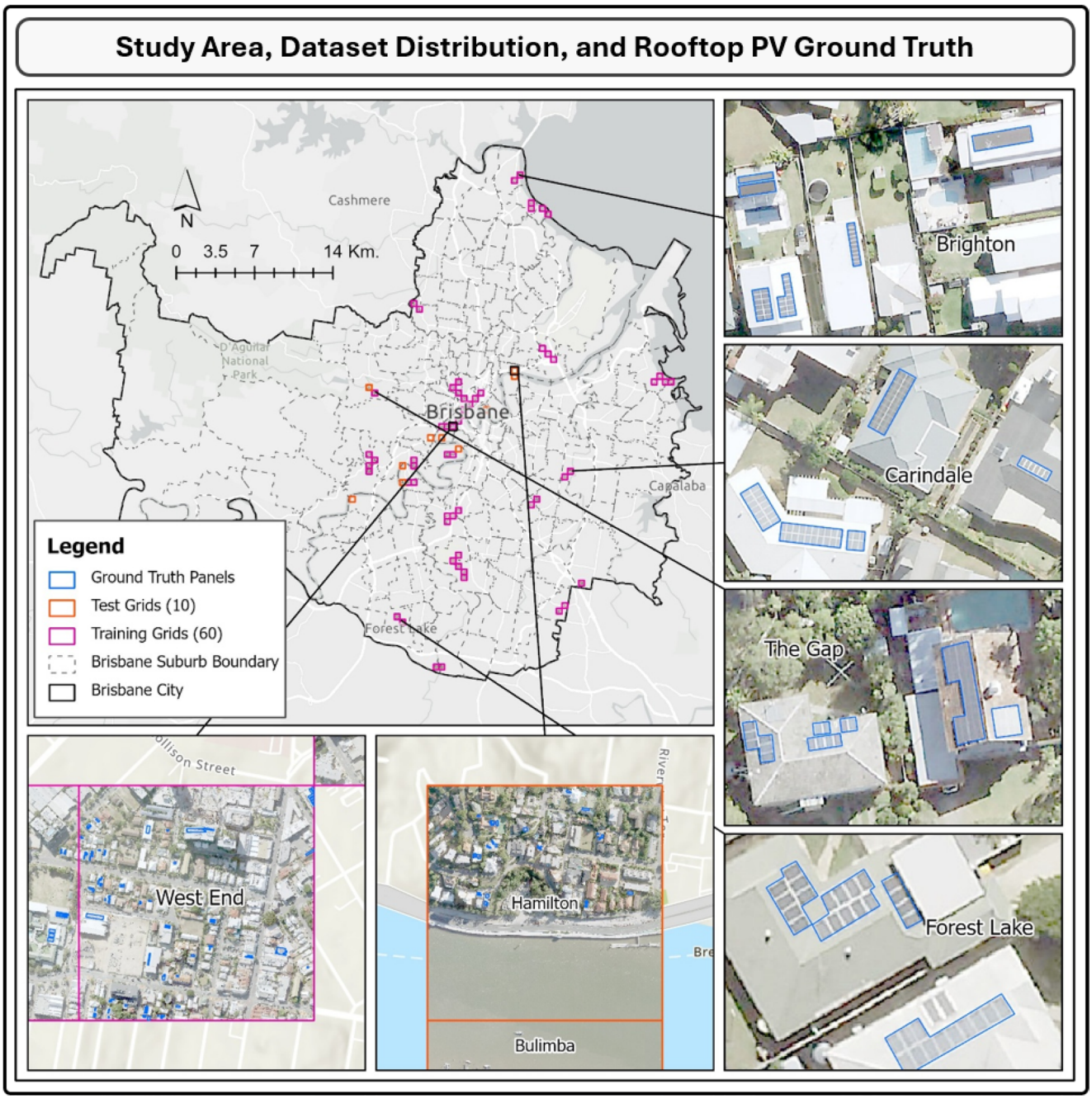


Fig. 3 Study area and examples of manually digitised rooftop PV installation features showing variation in polygon representation, ranging from single panels to grouped installations, derived from the high-resolution aerial imagery (Brisbane_2020_10cm_SISP_LGA), sourced from the Queensland SISP

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.

Dataset Allocation

The 70 selected grid cells were allocated into two datasets according to their role in the experimental design. The development dataset comprised 60 grid cells used to establish the baseline performance of the pretrained US model, perform local calibration through transfer learning, and compare the pretrained and locally calibrated models under identical evaluation conditions. The remaining 10 grid cells were reserved exclusively for independent testing. Both datasets were allocated purposively to capture variation in rooftop configurations, building densities, and urban characteristics across the selected Brisbane environments.

The independent test grids were excluded entirely from training-chip generation, model fine-tuning, and internal validation. Manually digitised rooftop PV installation features provided reference annotations for evaluating both models on these previously unseen local grid cells. Accordingly, development-dataset results were used to examine model adaptation, while the independent test dataset provided the primary evidence of performance beyond the development data. Independence refers specifically to the exclusion of test grids and their associated imagery and annotations from model development; it does not imply spatial isolation or evaluation in a separate geographic region. Furthermore, the purposive selection of grid cells does not provide a statistically representative assessment of the wider Brisbane LGA.

Baseline Inference Using the Pretrained US Model

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.

Transfer Learning and Local Calibration Workflow

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 Fig. 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 and Viswambharan 2021; Esri 2025b; Viswambharan and 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.

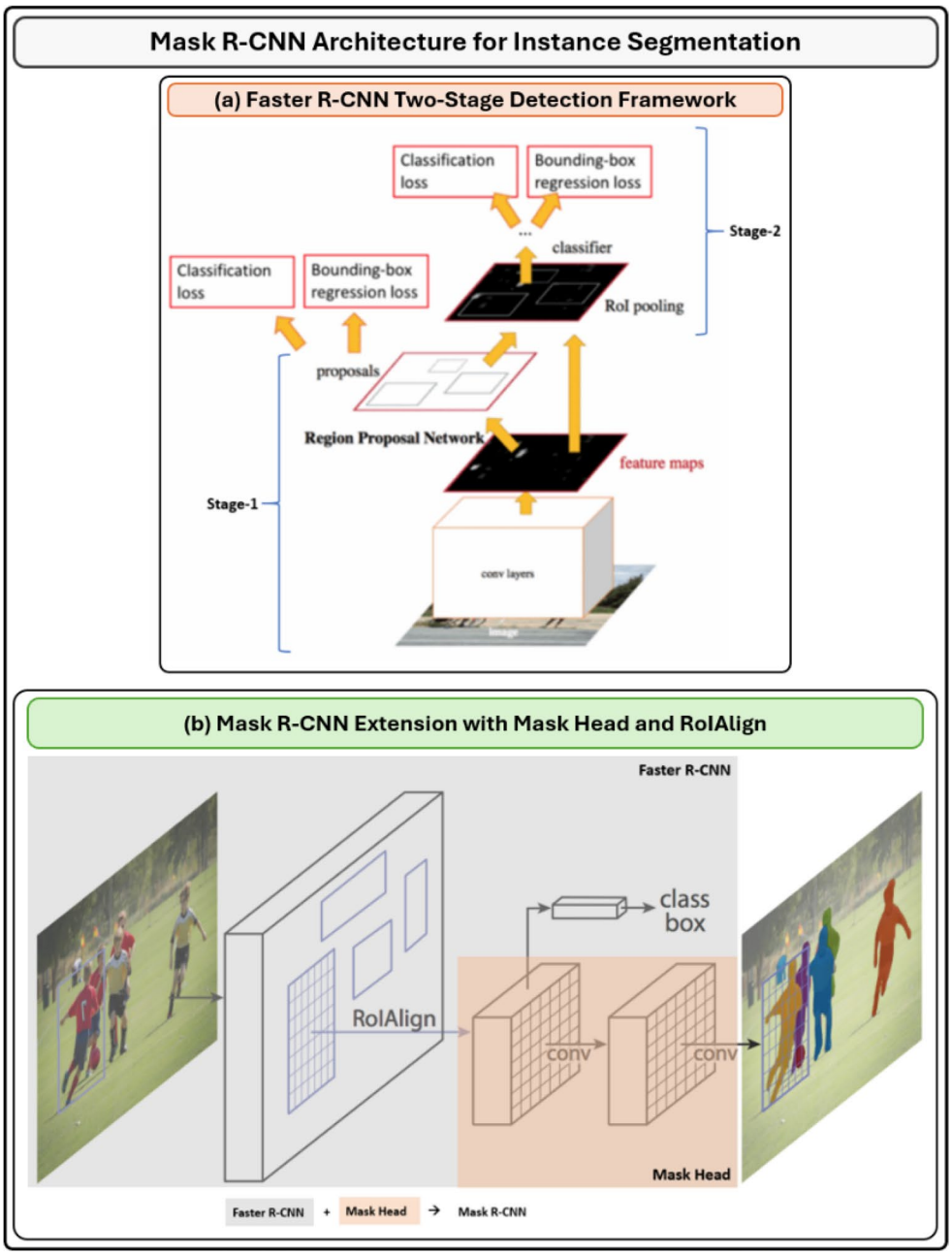


Fig. 4 Overview of the Mask R-CNN architecture used for instance segmentation. (**a**) The Faster R-CNN framework, which consists of a two-stage process: a region proposal network and classification with bounding-box regression. (**b**) The Mask R-CNN extension, which introduces a mask prediction branch (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)

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 starting model for local calibration. 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.

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 "Operational Trustworthiness Evaluation".

Operational Trustworthiness Evaluation

The evaluation combined quantitative object detection metrics with independent test performance, confidence-score analysis, and qualitative assessment of prediction errors and segmentation quality. These complementary analyses were used to operationalise the trustworthiness dimensions introduced in Section "Literature Background" (Table 1), while recognising the geographic and methodological scope of the Brisbane case study.

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.

Three IoU thresholds (0.50, 0.75, and 0.95) were selected to examine model performance under progressively stricter spatial matching requirements. An IoU threshold of 0.50 was used to assess general object detection and approximate spatial localisation, while 0.75 provided a more demanding assessment of segmentation agreement. The $\text{IoU} \geq 0.95$ threshold was included deliberately as a stringent sensitivity test of near-exact polygon alignment, rather than as a conventional minimum requirement for operational detection. Evaluating performance across these thresholds enabled the study to distinguish improvements in rooftop PV identification from the more demanding task of accurately delineating installation boundaries. This distinction is particularly relevant to the assessment of residual geometric uncertainty following transfer learning.

Object-level evaluation was interpreted using one-to-one matching between predicted and reference rooftop PV installation features at each specified IoU threshold. Under this evaluation principle, a predicted polygon can contribute to a true positive only when it is successfully matched to a reference polygon at the required IoU threshold. Where multiple predicted polygons overlap a single reference installation feature, at most one can contribute to a true positive, while additional unmatched predictions contribute to

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	$TP \Big/ (TP + FP)$	0 to 1
Recall	Proportion of detected reference rooftop PV features out of all reference rooftop PV features	$TP \Big/ (TP + FN)$	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

false positives. Conversely, where a single predicted polygon overlaps multiple reference features, at most one reference feature can be matched, with the remaining unmatched references contributing to false negatives. Predictions and reference features that do not satisfy the required matching criteria are counted as false positives and false negatives, respectively. Consequently, fragmented or merged predictions may affect both detection counts and spatial matching performance, particularly at stricter IoU thresholds.

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 "Literature Background". Specifically:

- Geographic adaptation, as evidence relevant to robustness, was evaluated by comparing the pretrained US model with the locally calibrated Brisbane model under identical evaluation conditions.
- 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.

These dimensions are operationally distinct but not statistically independent. Reliability is assessed through performance on held-out local imagery, whereas generalisation concerns the change in performance between development and independent datasets. Robustness is interpreted within the specific United States-to-Brisbane transfer scenario rather than as evidence of stability across multiple geographic domains. Similarly, transparency is assessed through observable prediction behaviour and error patterns, while residual uncertainty concerns the limitations revealed by those analyses. The latter does not constitute formal probabilistic uncertainty quantification.

Both models were evaluated on the development and independent test datasets using the same object detection evaluation protocol. Development-dataset results were used to examine model adaptation following local calibration, whereas the independent 10-grid test dataset provided the primary evidence of performance on previously unseen Brisbane imagery. Quantitative comparisons were complemented by qualitative inspection of successful detections, false positives, false negatives, and segmentation errors to characterise the remaining limitations of the locally calibrated model.

Analysis and Results

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}. Training was terminated after 32 epochs through

the early stopping mechanism, which monitors validation performance and limits further training when improvement ceases.

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. Training and validation losses followed broadly similar trajectories, with relatively small differences during the later epochs.

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 from 25 to 38 min, 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.

Overall, the training results showed a progressive reduction in training and validation losses, accompanied by an increase in Average Precision before performance plateaued. These trends indicate that the fine-tuning process converged within the observed training period. Model performance on the development and independent test datasets is presented in the following sections.

Model Adaptation under Geographic Domain Shift

The pretrained US model and the locally calibrated Brisbane model were first compared on the 60-grid development dataset to examine changes in detection performance following local calibration. Because these data contributed to model development, the comparison provides evidence of adaptation within the examined US-to-Brisbane transfer scenario rather than an independent assessment of geographic robustness.

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 high-confidence predictions, with mean confidence values exceeding 97% and a large proportion of detections assigned confidence scores above 98%. Local calibration produced only minor changes in these distributions despite substantial differences in detection performance reported below.

Table 4 compares detection outcomes for the two models on the development dataset. At IoU≥0.50, TP increased from 1,460 to 3,192 and FNs declined from 2,334 to 602 following local calibration. However, FPs also increased from 271 to 1,255. Similar patterns were observed at IoU≥0.75. At IoU≥0.95, true-positive counts were substantially lower for both models.

Figure 7 demonstrates that local calibration substantially improved detection performance at operationally meaningful

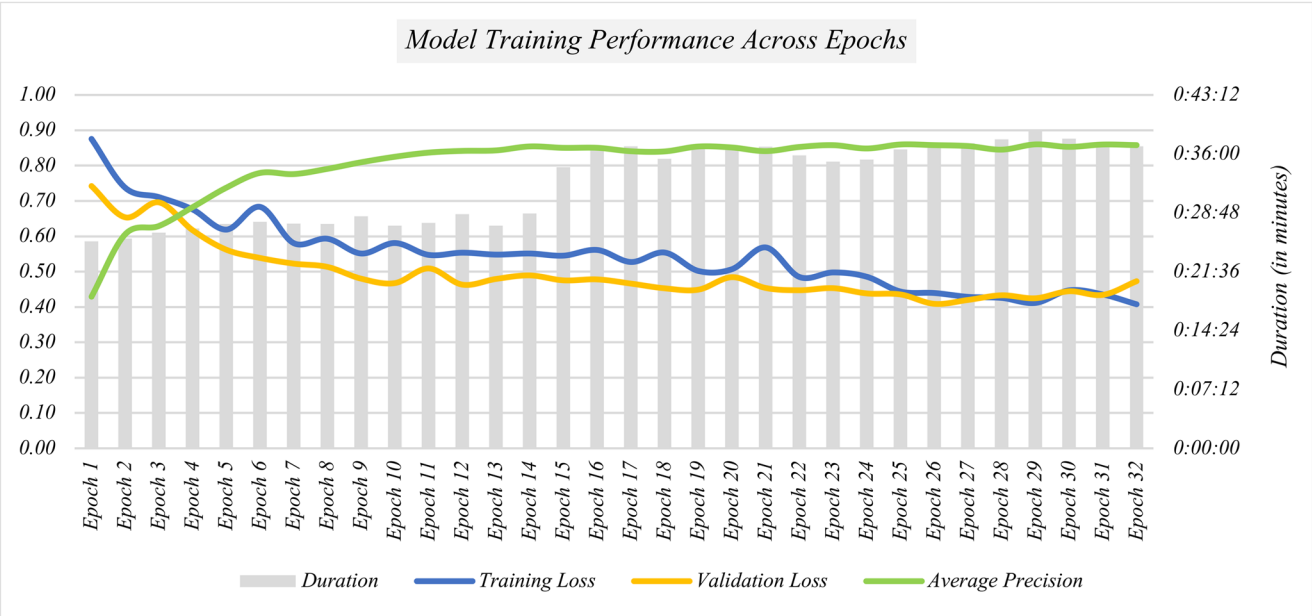


Fig. 5 Evolution of key training metrics across all epochs, including training loss, validation loss, average precision, and epoch duration

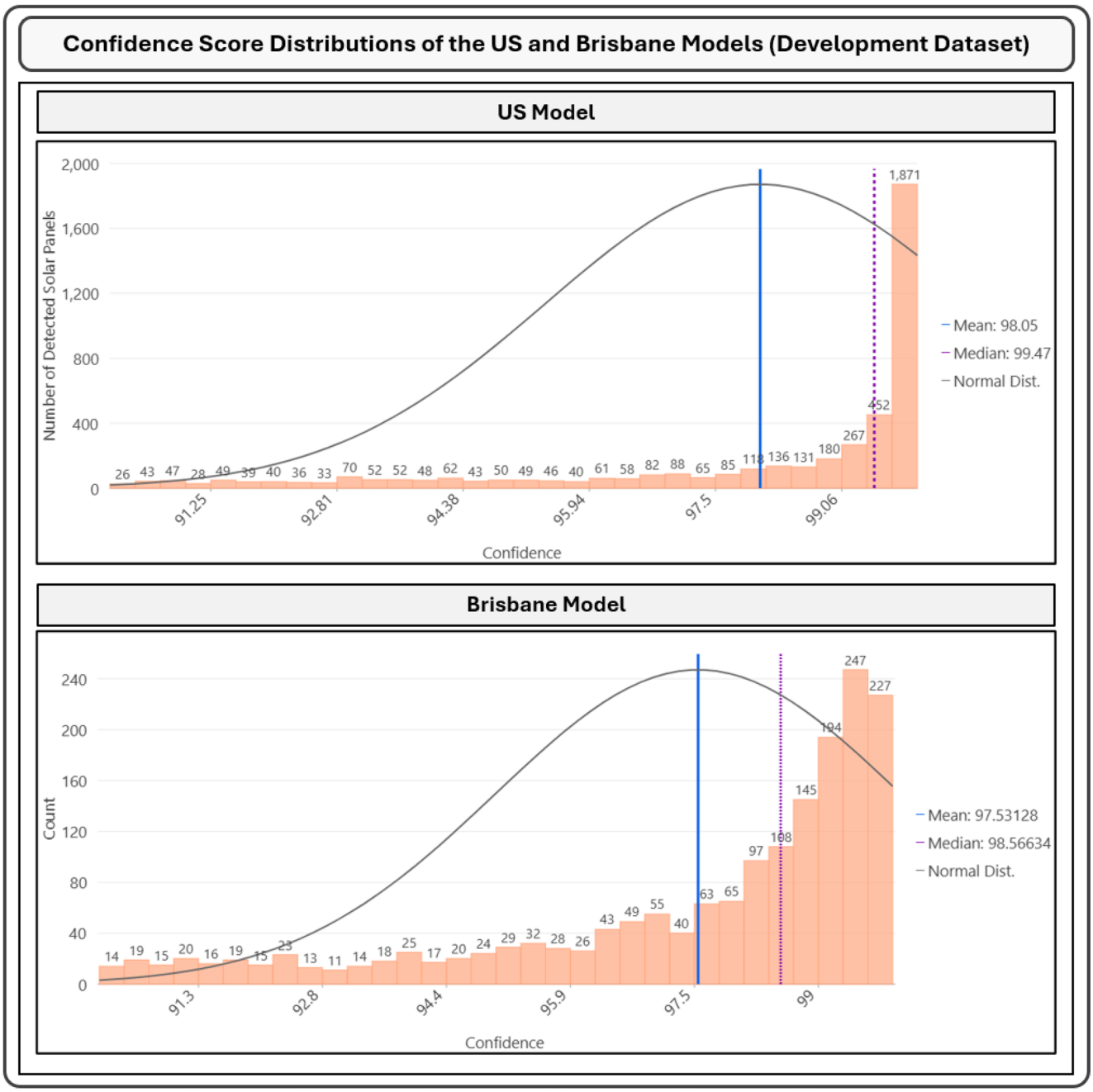
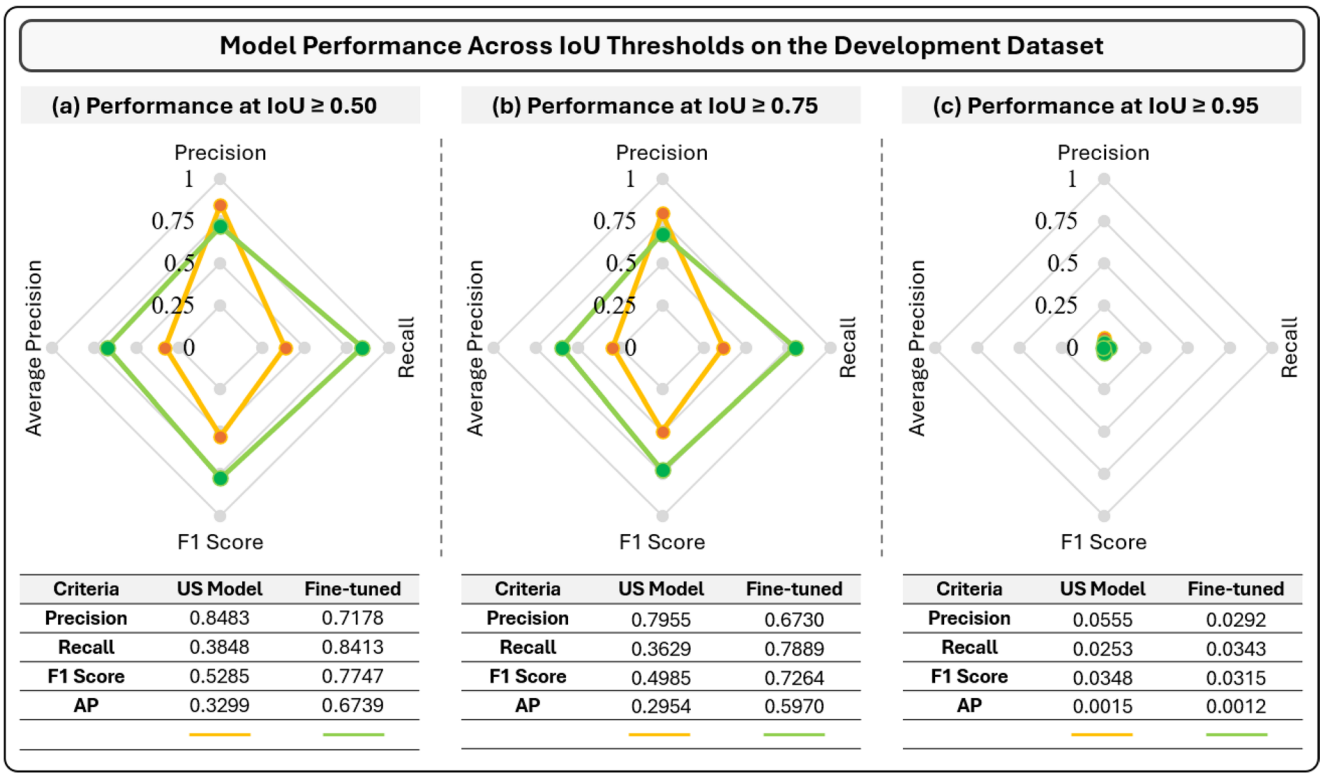


Fig. 6 Comparison of confidence score distributions produced by the pretrained US model and the locally calibrated Brisbane model on the development dataset

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



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

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

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

Fig. 7 Comparative performance of the pretrained US model and the locally calibrated Brisbane model across three IoU thresholds, illustrating performance improvements following local adaptation to the Brisbane Environment

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 demonstrate that transfer learning substantially improved the pretrained model's adaptation to the Brisbane environment, although they do not establish robustness across multiple geographic domains.

At IoU≥0.95, performance declined substantially for both models, indicating limited near-exact agreement between predicted and reference polygons. Because this stringent threshold is sensitive to small boundary differences, including those potentially associated with manually digitised reference features, these results are interpreted as a sensitivity assessment of geometric agreement rather than a conventional operational detection threshold. Because the development dataset contributed to model fine-tuning, these results describe model performance on data used during development. The following section reports the corresponding evaluation on the independent 10-grid test dataset, which was excluded from model development.

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 reliability and generalisation beyond the development grids within Brisbane, as it comprises previously unseen local imagery excluded from model development. 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 of both models on the independent 10-grid test dataset. Both models produced high-confidence predictions, with mean confidence values exceeding 97% and median values approaching 98–99%. The overall distributions changed only modestly following local calibration. These distributions are descriptive and do not constitute formal confidence calibration, which would require examining the relationship between individual prediction scores and their detection outcomes.

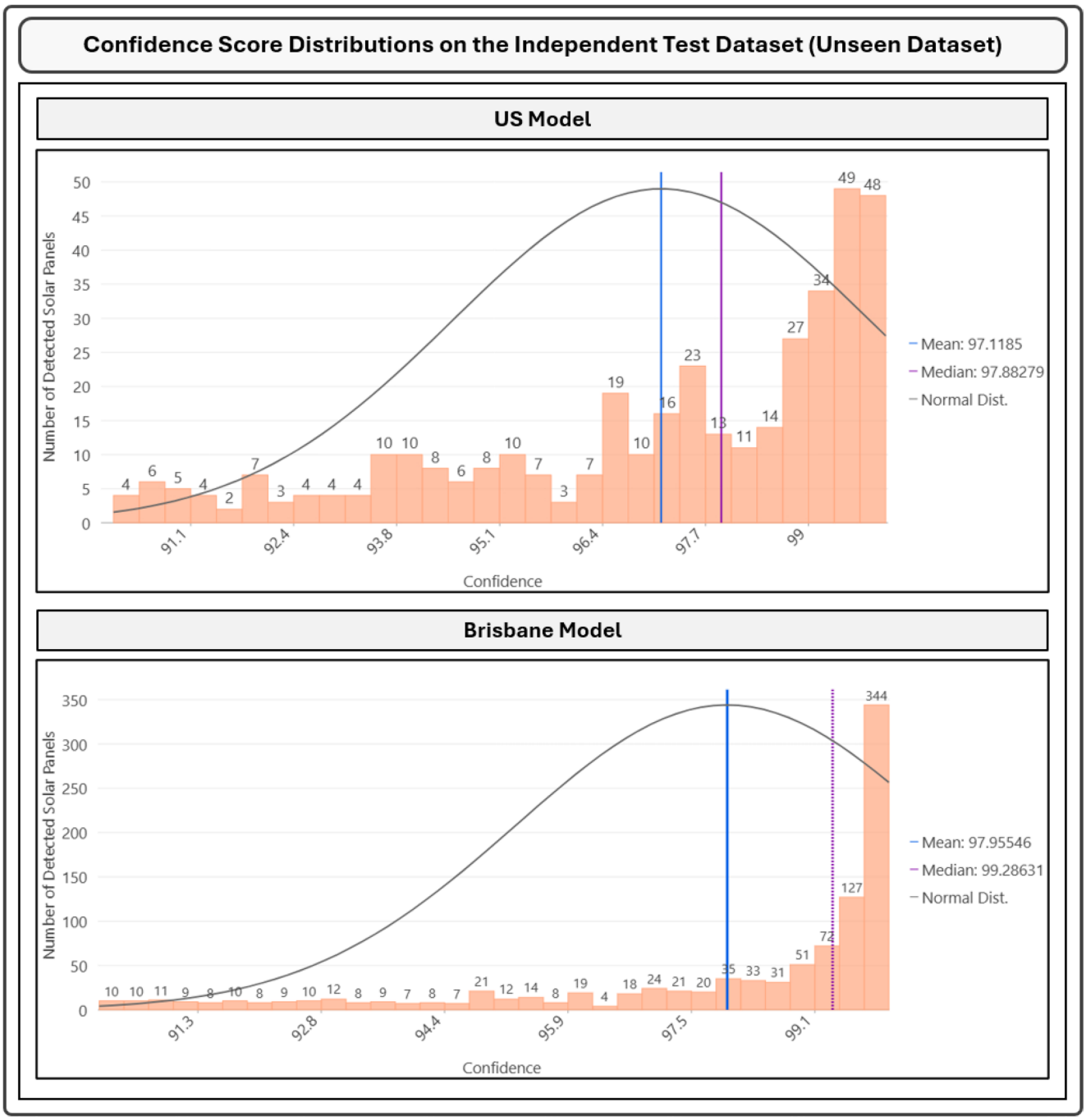


Fig. 8 Confidence score distributions for the pretrained US model and the locally calibrated Brisbane model on the independent test dataset

Table 5 compares the detection outcomes for both models on the independent test dataset. At IoU≥0.50, true positives increased from 300 to 689 and false negatives decreased from 543 to 154 following local calibration. However, false positives also increased from 76 to 301. Similar changes were observed at IoU≥0.75. At IoU≥0.95, both models recorded substantially fewer true positives, indicating limited near-exact agreement between predicted and reference polygons.

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

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

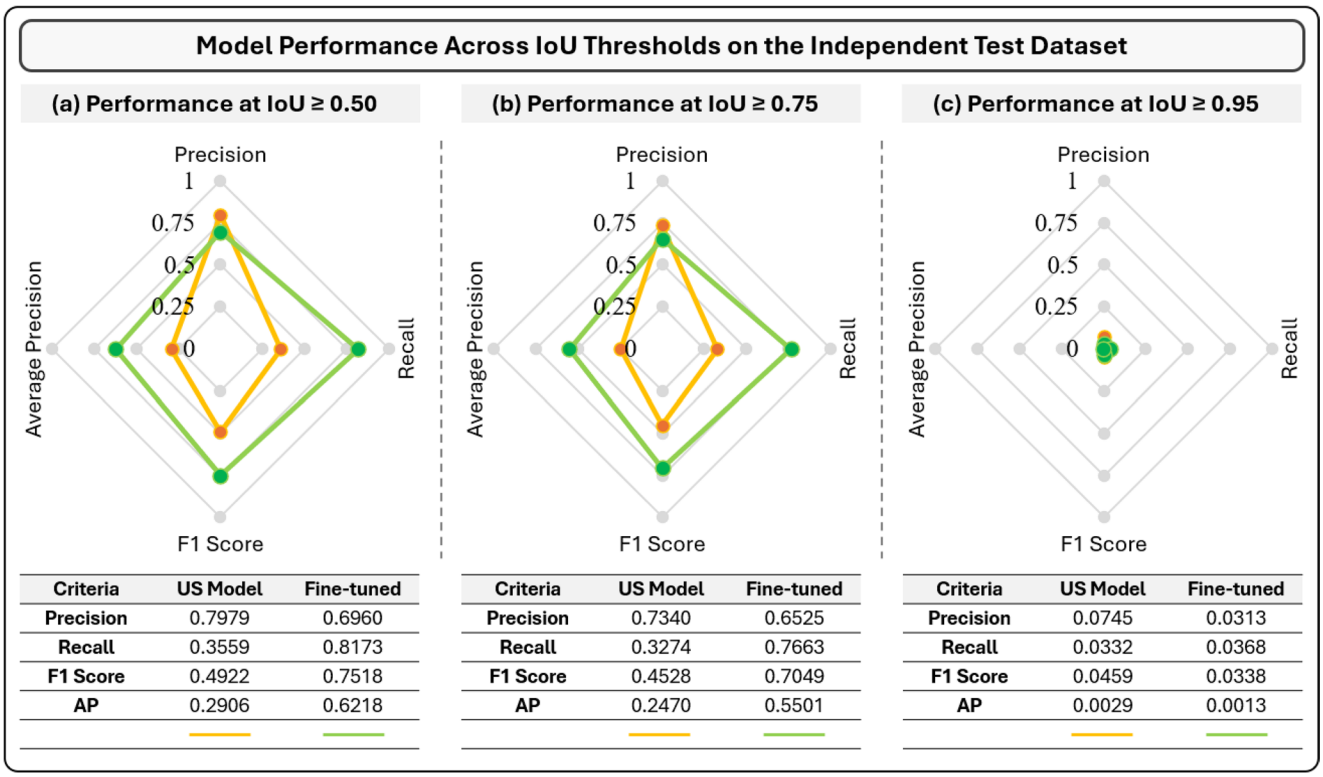


Fig. 9 Comparative detection performance of the pretrained US model and the locally calibrated Brisbane model across three IoU thresholds on the independent test dataset

were observed at IoU≥0.75, demonstrating that the benefits of local calibration extended beyond the calibration dataset and generalised to independent data.

Although local calibration substantially improved recall and F1-score, the increase in false-positive detections indicates that these improvements involved a trade-off between detection completeness and precision. Consequently, the observed performance gains should be interpreted in relation to application-specific requirements, particularly where false-positive detections may affect downstream spatial analysis.

Performance on the independent test dataset was slightly lower than on the development dataset at IoU≥0.50 and IoU≥0.75. Nevertheless, the locally calibrated Brisbane model retained higher recall and F1-scores than the pretrained US model on these previously unseen Brisbane grids. These results provide evidence that the observed improvements were maintained beyond the development dataset, within the geographic and sampling scope of the independent evaluation.

The independent test results showed improved detection performance following local calibration, alongside persistent false positives, missed detections, and reduced spatial matching performance at stricter IoU thresholds. The following section examines representative prediction outcomes and segmentation errors to characterise these remaining limitations.

Transparency and Residual Uncertainty

This section presents a qualitative assessment of representative prediction outcomes from the pretrained US model and the locally calibrated Brisbane model. The analysis examines recovered rooftop PV detections, false-positive patterns, segmentation inaccuracies, and prediction behaviour

on the independent Brisbane test grids. These observations complement the quantitative results by characterising the remaining detection and geometric errors.

Recovery of Missed Rooftop PV Installations through Transfer Learning

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 (Fig. 10) display detections from the US model, while F-a to F-f (Fig. 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.

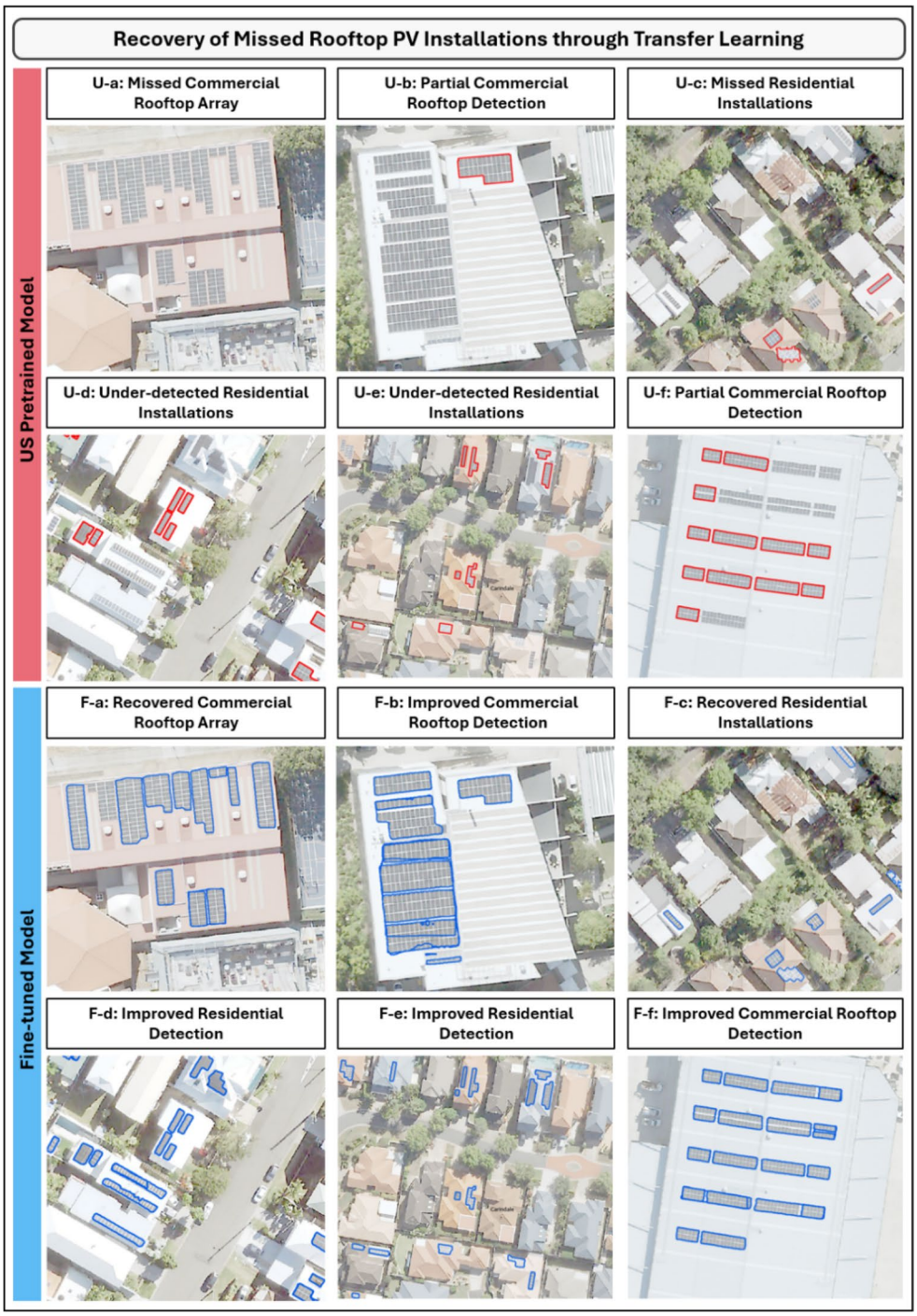


Fig. 10 Comparative examples of rooftop solar panel detections. Images U-a to U-f show detections produced by the pretrained US model, while F-a to F-f display results from the fine-tuned Brisbane model for the same locations

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 qualitative observations provide visual evidence supporting the quantitative improvements reported in Sections "Model Adaptation under Geographic Domain Shift" and "Reliability and Generalisation on an Independent Test Dataset", reinforcing that transfer learning substantially improves model adaptation under geographic domain shift.

The quantitative results further clarify the extent of the remaining omission errors. At IoU$\geq$0.50, false negatives decreased from 2,334 to 602 on the development dataset and from 543 to 154 on the independent test dataset following local calibration (Tables 4 and 5). Nevertheless, the 154 unmatched reference features on the independent test dataset indicate that some rooftop PV installations remained undetected or failed to meet the required spatial matching criterion. The examples in Fig. 10 suggest that smaller installations, irregular array configurations, and complex rooftop surroundings can present particular detection challenges. These observations identify potential failure conditions rather than establishing the relative frequency of each error type across the entire dataset.

Persistent False Positive Detection Patterns

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 Fig. 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 visual examples in Fig. 11 reveal several recurring sources of false-positive detections. The pretrained US model misidentified construction materials, scaffolding, reflective surfaces, playgrounds, sports courts, vehicles, road markings, shadows, and rooftop features such as solar hot water systems as rooftop PV installations. These errors suggest that visually similar non-PV features can be confused with rooftop PV arrays. The locally calibrated Brisbane model also produced false positives, including detections associated with complex shadows, pedestrian bridges, and urban edges. Importantly, the quantitative results show that local calibration increased the absolute number of false positives at IoU$\geq$0.50, from 271 to 1,255 on the development dataset (Table 4) and from 76 to 301 on the independent test dataset (Table 5). This increase occurred alongside substantial reductions in false negatives, indicating a trade-off between improved detection coverage and additional incorrect predictions. Thus, although local calibration improved overall detection performance, it did not reduce the total false-positive burden.

These findings demonstrate that false-positive errors remain an important limitation following local calibration. The examples in Fig. 11 illustrate plausible sources of confusion, while Tables 4 and 5 quantify the increase in unmatched predictions.

Residual Geometric Uncertainty in Instance Segmentation

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 Fig. 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 Fig. 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 Fig. 12) demonstrates improved detection coverage and spatial consistency overall; however, geometric distortions remain visible. For example, in Fig. 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



Fig. 11 Examples of FPs from the US pretrained model and the fine-tuned Brisbane model

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$\geq$0.95), where minor boundary misalignments significantly reduce intersection-over-union scores. 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.

The IoU sensitivity results provide additional quantitative evidence of geometric disagreement. On the independent test dataset, the locally calibrated model achieved 689 true positives at IoU$\geq$0.50, compared with 646 at IoU$\geq$0.75 and only 31 at IoU$\geq$0.95 (Table 5). This progressive decline indicates that many predictions meeting the less stringent matching criteria did not achieve near-exact overlap with their reference polygons. Figure 12 illustrates possible manifestations of this disagreement, including fragmented masks, incomplete array coverage, irregular contours, and boundary misalignment. However, these threshold-based counts do not distinguish prediction-boundary errors from uncertainty in manually delineated reference polygons, nor do they establish the frequency of individual geometric error categories.

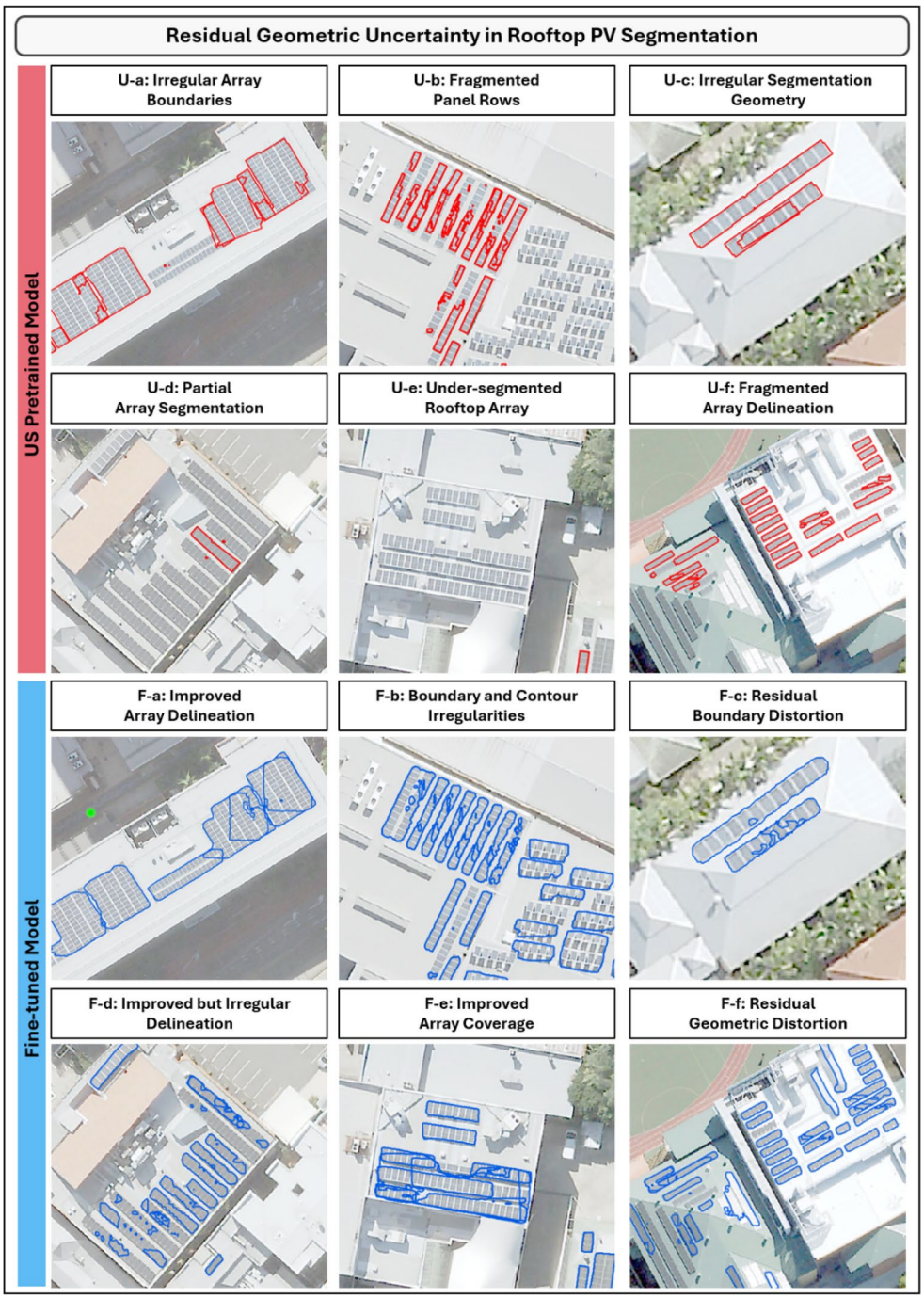


Fig. 12 Comparative examples of geometric distortions in solar panel segmentation. Panels U-a to U-f display outputs from the pretrained US model, while F-a to F-f present the corresponding results from the fine-tuned Brisbane model for identical rooftop locations

Model Behaviour on Previously Unseen Urban Environments

Figure 13 presents representative detection results from the independent 10-grid Brisbane test dataset. The upper row shows examples of rooftop PV installations with relatively well-aligned predicted polygons across different rooftop configurations. The lower row illustrates more challenging cases involving large or densely arranged arrays, where segmentation inaccuracies remain visible.

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.



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

These errors may be associated with scene complexity, dense panel arrangements, and ambiguity in delineating installation boundaries. 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.

The qualitative examples illustrate improved recovery of rooftop PV installations following local calibration, alongside persistent false positives and geometric segmentation inaccuracies. These observations are consistent with the quantitative results, which showed fewer missed reference features but more unmatched predictions after fine-tuning, together with substantial sensitivity to stringent IoU matching criteria.

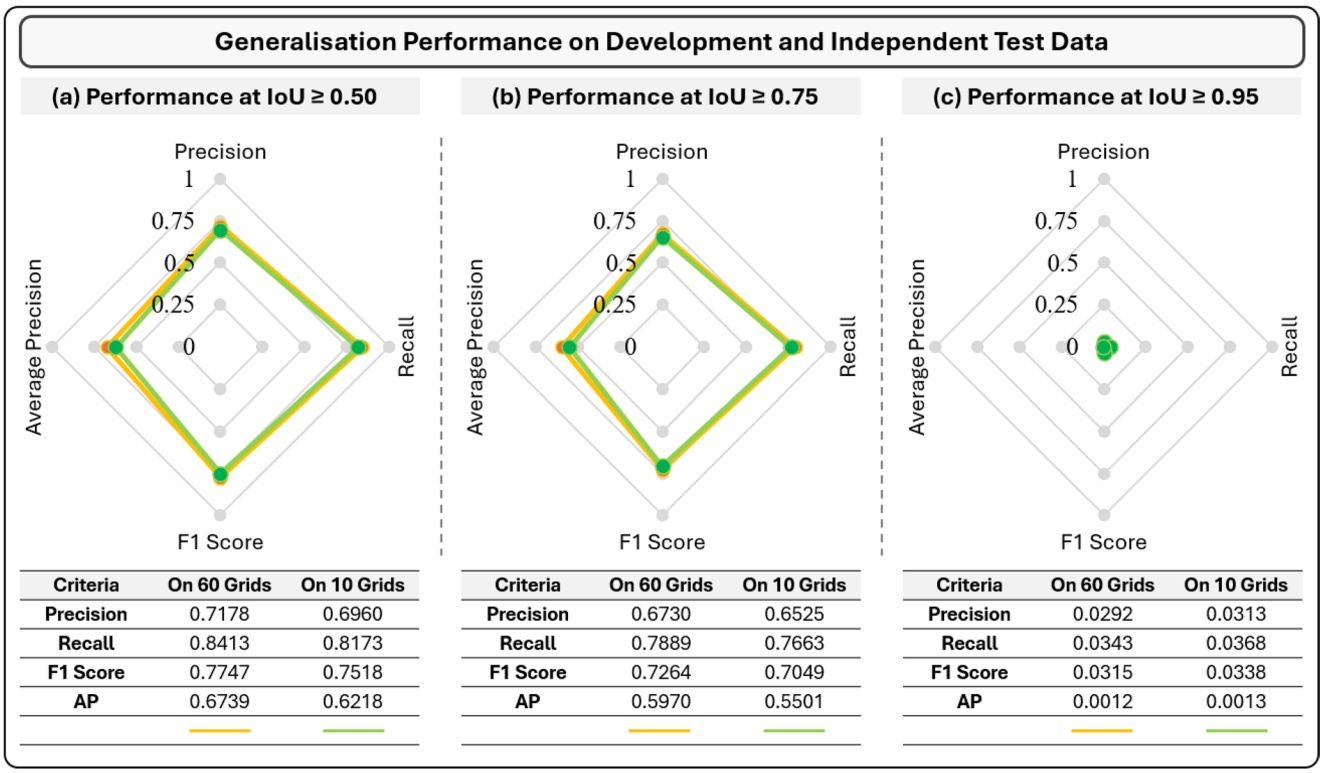
Findings and Discussions

Transfer Learning Improves Deployment but Does Not Establish Trustworthiness

The findings demonstrate that local calibration substantially improved the detection performance of the pretrained US rooftop PV model in the Brisbane case study. The most pronounced improvement was in recall, indicating that the locally calibrated model recovered a substantially greater proportion of reference PV installation features. However, this improvement was accompanied by increased false-positive detections and lower precision, revealing a trade-off between improved detection coverage and additional unmatched predictions. This distinction is operationally important because the suitability of a detection model depends not only on the number of installations identified but also on the consequences of incorrect predictions for its intended application.

These findings are consistent with previous research highlighting the importance of target-region data for rooftop PV detection. Ranalli et al. (2024) demonstrated that incorporating local training data can improve performance when models are transferred across geographic regions, while Zech et al. (2024) showed that active learning can reduce the annotation effort required for effective adaptation. The present study complements these contributions by examining the performance of an operationally available pretrained model before and after local calibration, with particular attention to independent local evaluation and remaining prediction errors. Unlike studies focused primarily on geographic generalisation or annotation efficiency, the emphasis here is on the evidence needed to inform deployment decisions after adaptation.

Figure 14 shows that the locally calibrated model maintained broadly comparable performance on the development and independent test datasets at IoU thresholds of 0.50 and 0.75, although some reduction was observed on the held-out grids. This provides evidence that the benefits of local



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

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

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

Fig. 14 Comparison of the locally calibrated Brisbane model's performance on the 60-grid development dataset and the independent 10-grid Brisbane test dataset across the evaluated detection metrics

calibration extended beyond the imagery used during model development. Nevertheless, because the independent evaluation was restricted to 10 purposively selected grids within Brisbane, the results support local generalisation within the sampled conditions rather than broader geographic transferability. The limitations of this evaluation are considered further in Section "Study Limitations and Future Research".

The qualitative error analysis provides possible explanations for the remaining limitations. Visually similar non-PV rooftop and urban features may contribute to false-positive detections, while small or irregular installations, complex roof structures, and partial occlusion may contribute to missed detections. Segmentation inaccuracies may also reflect difficulties in delineating closely arranged or visually ambiguous panel boundaries. These explanations are consistent with the representative examples examined in Section "Transparency and Residual Uncertainty" but were not independently tested as causal factors. Importantly, the substantial decline in spatial matching performance at IoU$\geq$0.95 also reflects the sensitivity of near-exact polygon comparisons to both prediction errors and uncertainty in manually digitised reference boundaries. Together, these findings show why improved detection performance must be interpreted in relation to the intended use of the model outputs.

Synthesising Operational Trustworthiness for Pretrained GeoAI Deployment

While Section "Transfer Learning Improves Deployment but Does Not Establish Trustworthiness" 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, the findings suggest that assessing operational trustworthiness requires multiple complementary sources of evidence that collectively characterise model behaviour under geographic domain shift. While these assessments provide a stronger basis for evaluating deployment suitability, they do not independently establish that the model is trustworthy across all geographic or operational contexts. This synthesis closes the conceptual perspective introduced in Section "Literature Background" 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.

The findings can be interpreted through the operational trustworthiness dimensions established in Section "Literature Background" (Table 1). Local calibration substantially improved detection performance under the examined geographic transfer scenario, while independent evaluation demonstrated that these improvements were maintained on previously unseen grid cells within Brisbane. The comparison between development and independent datasets provided evidence of local generalisation, although broader geographic robustness was not established. Qualitative error analysis further revealed persistent false positives, missed detections, and segmentation inaccuracies, while the IoU sensitivity analysis highlighted uncertainty in precise polygon delineation. Together, these findings demonstrate the value of integrating complementary evaluation evidence rather than relying exclusively on improvements in predictive accuracy.

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. Evidence of local adaptation, independent test performance, and generalisation within the sampled Brisbane environments provides confidence that the observed improvements extend beyond the development 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. Nevertheless, the evidence presented here should be interpreted within the scope of the Brisbane case study, as the limited independent evaluation and remaining model uncertainties do not constitute definitive confirmation of trustworthy deployment across broader geographic or operational contexts.

Figure 15 synthesises the implications of the empirical findings for evidence-based deployment decisions. Whereas Fig. 1 establishes the conceptual assessment perspective adopted at the beginning of the study, Fig. 15 illustrates how the results inform a progression from accuracy-centred model evaluation towards deployment decisions supported by complementary evidence. Specifically, the figure highlights that improvements achieved through transfer learning should be considered alongside independent performance evaluation, observed model failure modes, and residual uncertainty. It therefore serves as an interpretive synthesis of the study's findings rather than a repetition of the conceptual framework introduced in Section "Literature Background".

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.

Broader Implications for Operational GeoAI

The findings have practical implications for organisations adopting pretrained GeoAI models through commercial GIS platforms and other operational geospatial workflows. Before applying a model to a new geographic environment, practitioners should assess its initial performance using locally relevant reference data and consider local calibration where substantial performance degradation is observed. Following calibration, evaluation should include imagery excluded from model development, with attention to both omission and commission errors and to the spatial agreement required by the intended application. The Brisbane case study illustrates why improved recall should be considered alongside the additional false positives introduced during fine-tuning, rather than treating improved overall detection performance as sufficient evidence for deployment.

Operational requirements should also determine how model outputs are verified and used. For applications such as preliminary rooftop PV inventory development,

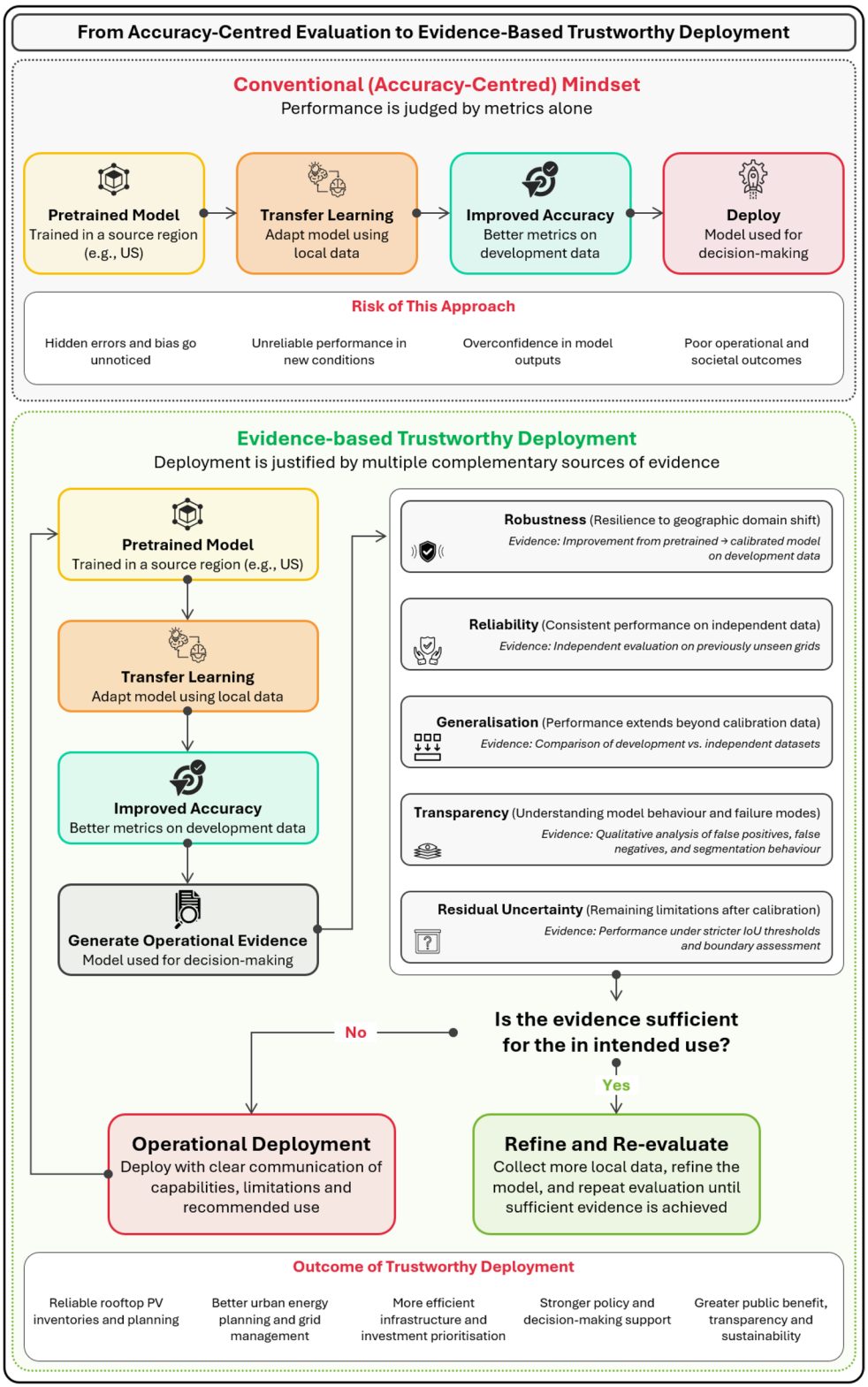


Fig. 15 From accuracy-centred evaluation to evidence-based trustworthy deployment of pretrained GeoAI models

the locally calibrated model may provide a useful starting point, subject to appropriate quality assurance. Applications requiring accurate installation boundaries, rooftop area estimates, or capacity modelling may require additional polygon verification and refinement because geometric discrepancies remained evident at stricter IoU thresholds. Similar considerations may apply when pretrained GeoAI models are adapted for building extraction, infrastructure

monitoring, and other mapping tasks. The broader implication is that model suitability should be assessed against application-specific error tolerances and supported by independent evaluation, transparent reporting of limitations, and proportionate human review.

Study Limitations and Future Research

Several limitations should be considered when interpreting the findings of this study. First, the independent test dataset comprised only 10 grid cells within Brisbane. Although these grids were excluded from model development and provided evidence of performance on previously unseen local imagery, their relatively small number limits the scope of the reliability and generalisation assessment. Furthermore, the purposive selection of the 70 study grids was intended to capture diverse urban conditions rather than provide a statistically representative sample of the Brisbane Local Government Area. Consequently, the reported performance metrics should not be interpreted as unbiased estimates of model accuracy across Brisbane.

Second, the evaluation was conducted entirely within Brisbane. Although the study examined the adaptation of a pretrained US model to an Australian urban environment, the independent test grids were located within the same city as the development grids. The findings therefore demonstrate local transferability to previously unseen sampled areas rather than broader geographic generalisation. Model performance in other Australian cities, rural environments, or regions with different imagery characteristics remains untested.

Third, the reference dataset was developed through manual visual interpretation of high-resolution aerial imagery. Although the digitised polygons were visually reviewed, uncertainty may remain in the identification and delineation of rooftop PV installations, particularly where features are obscured by shadows, closely spaced, or visually ambiguous. Furthermore, individual reference polygons may represent single panels or grouped installations. These annotation characteristics can influence object-level matching and the interpretation of segmentation accuracy.

Finally, performance assessment was sensitive to the selected Intersection over Union (IoU) thresholds. While $IoU \geq 0.50$ and $IoU \geq 0.75$ provided useful measures of detection and spatial agreement, $IoU \geq 0.95$ imposed a particularly stringent requirement for near-exact polygon alignment. At this threshold, relatively small differences between predicted and reference boundaries may substantially reduce measured performance, including differences arising from annotation uncertainty. Accordingly, the results at $IoU \geq 0.95$ should be interpreted as a sensitivity assessment of geometric agreement rather than a conventional operational acceptance criterion.

Although confidence score distributions were examined to provide insights into model prediction behaviour, formal confidence calibration was not performed. Consequently, the reported confidence scores should not be interpreted as direct estimates of prediction correctness. Future investigations could incorporate confidence calibration techniques to further assess the reliability of model predictions under geographic domain shift.

The analysis was based on high-resolution aerial imagery acquired in 2020. Consequently, the resulting reference annotations and model predictions reflect rooftop PV installations visible at the time of image acquisition rather than the current distribution of installations. Future research could incorporate more recent and multi-temporal imagery to assess model performance under changing rooftop PV configurations and urban development conditions.

An additional limitation is the absence of a model trained from scratch using the Brisbane dataset. Consequently, although the results demonstrate substantial improvements following local calibration of the pretrained model, they do not isolate the contribution of pretrained weights relative to training an equivalent model without pretraining. Future research could investigate this distinction through a controlled comparison using identical training and evaluation data.

Future research should evaluate pretrained and locally calibrated GeoAI models using larger, more geographically diverse, and preferably probability-based independent samples. Validation across multiple Australian cities and contrasting imagery conditions would provide stronger evidence of geographic generalisation. Further investigation of annotation consistency, confidence calibration, and geometric segmentation uncertainty would also strengthen the evidence available for assessing trustworthy operational deployment.

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.

The findings of this study indicate that transfer learning can be an effective strategy for adapting pretrained GeoAI

models to new geographic environments, although its effectiveness may vary depending on local imagery characteristics, training data availability, and operational conditions. In the Brisbane rooftop PV mapping case study, local calibration substantially improved detection performance, with these improvements largely maintained on an independent set of 10 previously unseen Brisbane grids. Nevertheless, the limited independent evaluation does not establish broader geographic generalisation. Furthermore, the independent evaluation and qualitative error analysis revealed persistent false positives, missed detections, and geometric segmentation errors, particularly under stricter spatial matching criteria. These findings highlight that, although transfer learning can improve detection performance, predictive improvements alone are insufficient to establish trustworthy deployment. Accordingly, the results provide empirical evidence to inform the assessment of operational trustworthiness within the sampled Brisbane environments, rather than definitive confirmation of deployment readiness across broader geographic contexts.

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 datasets excluded from model development should become standard practice for assessing operational reliability, with geographically diverse testing undertaken where broader generalisation is required. 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41651-026-00293-0.

Authors' Contributions TH: Data collection, processing, investigation, analysis, and writing – original draft. TY: Supervision, conceptualisation, writing - review & editing; NT, FB: Writing - review & editing. All authors have read and agreed to the final version of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability Data are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate This project does not involve human participants, animals, or plants and, therefore, does not require ethical approval or a formal ethics statement.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahmed S, El-Shazely A, Ahmed W (2025) Deep learning for building footprint extraction using UAV-based orthoimages. J Indian Soc Remote Sens 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. J King Saud Univ Comput Inform Sci 37(8):258

Andorful F, Herfort B, Melanda EA, Antonio ND, Zipf A, Camboim SP (2026) Transparency and trust in collaborative mapping: Concerns and dilemmas in AI-assisted road integration within OpenStreetMap. Annals Am Association 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 Econ 112:106124

Best R, Esplin R, Hammerle M, Nepal R, Reynolds Z (2023) Do solar panels increase housing rents in Australia? Hous Stud 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. Appl Soft Comput 149:9, 111013

Choi Y (2023) GeoAI: Integration of artificial intelligence, machine learning, and deep learning with GIS. Appl Sci 13(6):3895

Clark C, Pacifici F (2023) A solar panel dataset of very high resolution satellite imagery to support the Sustainable Development Goals. Sci 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 Sens 17(19):3389

Deng R, Guo Z, Zhang P, Li J, Xu X, Chen Q, Chen Y, Yan J (2026) PANEL: A photovoltaic-specific vision-language model for zero-shot and few-shot PV interpretation tasks in remote sensing. ISPRS J Photogrammetry Remote Sens 238:337–355

Doan T, Wilson D, Rashford S, Bosley E (2021) Ambient temperatures, heatwaves and out-of-hospital cardiac arrest in Brisbane, Australia. Occup Environ Med 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 Int J Geo-Information 14(12):491

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/

Esri (2023) Solar panel detection - New Zealand. Retrieved 1 February 2026 from https://qutaus.maps.arcgis.com/home/item.html?id=23d46b7e7f7d41abae01885b64834af8

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 Cult Soc 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. Sol 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. Adv Eng Inform 64:103104

Ghamisi P, Yu W, Marinoni A, Gevaert CM, Persello C, Selvakumaran S, Girotto M, Horton BP, Rufin P, Hostert P, Pacifici F, Atkinson PM (2025) Responsible artificial intelligence for earth observation: Achievable and realistic paths to serve the collective good. IEEE Geoscience Remote Sens 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. Eng Appl Artif Intell 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 Appl 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 Clim 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 Am Association Geographers 115(1):148–166

Hu Y, Goodchild M, Zhu AX, 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. Ann 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, pp 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-17

83081743-1.0.1.1-WNWSabGh4Jrqb_RkfffdSQriDwKUrhZnRiZ7hnoFpeQ

Jagatheesaperumal SK, Bibri SE, Huang J, Rajapandian J, Parthiban B (2024) Artificial intelligence of things for smart cities: advanced solutions for enhancing transportation safety. Comput Urban Sci 4(1):10

Jiao D, Sun J (2018) Real-time visualization of geo-sensor data based on the protocol-coupling symbol construction method. ISPRS Int J 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. Environ Data Sci, 4, e22

Kaya O, Klepacka A, Florkowski W (2019) Achieving renewable energy, climate, and air quality policy goals: Rural residential investment in solar panel. J Environ Manage 248:109309

Li W, Lee H, Wang S, Hsu C-Y, Arundel ST (2023a) 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 (2023b) The making of responsible innovation and technology: An overview and framework. Smart Cities 6(4):1996–2034

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 Soc 115:105875

Lodhi MK, Tan Y, Li Y, Khan MN, Naeem S (2025) Deep learning ensemble and multi-criteria GIS for high-fidelity rooftop solar potential mapping. J Geovisualization Spat Anal 9(2):38

Long Y, Xu X, Huo Z (2025) Urban rooftop photovoltaic potential model: A study on assessment methods and model framework. Energy Build 345:116138

Lu F, Cheng S, Wang P (2025) GeoAI enabled urban computing: Status and challenges. Ann 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. Int J Appl Earth Obs Geoinf 125:103563

Marasinghe R, Yigitcanlar T, Mayere S, Washington T, Limb M (2024) Towards responsible urban geospatial AI: Insights from the white and grey literatures. J Geovisualization Spat Anal 8(2):24

Nagy LA, Szabó S, Burai P, Bertalan L (2024) Improving urban mapping accuracy: Investigating the role of data acquisition methods and SfM processing modes in UAS-based survey through explainable AI metrics. J Geovisualization Spat Anal 8(1):15

Narayanan R, Parthkumar P, Pippia R (2021) Solar energy utilisation in Australian homes: A case study. Case Stud Therm Eng 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 SJ, Yang Q (2010) A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 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. J Soft Comput Civil Eng 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. Sol 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

Ranalli J, Zech M, Tetens H-P (2024) Diversifying training data does not improve generalizability of neural network models for PV identification. J Renew Sustain Energy, 16(6)

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 Geographic Inform Sci 52(4):331–349

Simin L, James M, H (2024) Climate change adaptation strategies in coastal cities of Australia: A comparative study of Sydney, Melbourne, and Brisbane. Stud Social Sci Humanit 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: Spatio-temporal mapping of distributed solar power for New York City. WIRE Energy Environ, 9(5), e374

Tayal D (2017) Achieving high renewable energy penetration in Western Australia using data digitisation and machine learning. Renew Sustain Energy Rev 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 models 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 UW (2024) A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX 13:102946

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J 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. Landsc Urban Plann 214:104167

Wu P, Zhang Z, Peng X, Wang R (2024) Deep learning solutions for smart city challenges in urban development. Sci Rep 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

Zech M, Tetens H-P, Ranalli J (2024) Toward global rooftop PV detection with Deep Active Learning. Adv Appl Energy 16:100191

Zhang W, Wang L, Cheng S, Lu F (2026a) Ensemble learning for spatiotemporal data: Methods, applications, challenges and opportunities. Int J Geogr Inf Sci, 1–49

Zhang X, Yang F, He Z, Li W, Yang M, Computers (2026b) Environ Urban Syst, 123, 102370

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.