Remote Sensing of Environment 340 (2026) 115413

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Making footprints move: Temporal disaggregation of building footprint data using Sentinel-2 imagery and Bayesian deep learning

Manuel Huber [a,*], Christian Geiß [a,b], Jessy Ribaira [a], Michael Schmitt [d], Hannes Taubenböck [a,c]

[a] Deutsche Zentrum für Luft- und Raumfahrt (DLR) Deutsches Fernerkundungsdatenzentrum (DFD) Georisiken und zivile Sicherheit, Oberpfaffenhofen, Munich, 82234, Bavaria, Germany
[b] Department of Geography, University of Bonn, Bonn, 53115, Germany
[c] Institute for Geography and Geology, Julius-Maximilians-Universität Würzburg, Würzburg, 97074, Germany
[d] Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, 85577, Germany

HIGHLIGHTS

- Novel method to assign temporal labels to high-resolution building footprints.
- Merging building footprints with Sentinel-2 data outperforms using each source alone.
- The new method offers a flexible solution for finer temporal urban monitoring.
- The framework assigns temporal metadata and uncertainty to building footprints.

ARTICLE INFO

Edited by Dr. Marie Weiss

Keywords:
Sentinel-2
Building footprint
Uncertainty
Deep learning

ABSTRACT

High-resolution building footprints from Overture, Google, Meta, and OpenStreetMap are essential for urban and environmental studies. However, these datasets often lack temporal metadata, limiting their utility for applications, that require spatio-temporal information, such as risk assessment, population estimation, and urbanization monitoring. This study presents a novel method for temporally disaggregating static building footprints by leveraging Sentinel-2 satellite imagery and a Bayesian U-Net segmentation model. The approach allows the assignment of time labels to individual building footprints and probabilistic uncertainty estimation. Beyond temporal labeling, disaggregation greatly improves label quality, boosting R^2 from -0.10 to 0.80 for building count and from 0.74 to 0.89 for built-up area accuracy. Overall, the method robustly generalizes, enabling flexible temporal disaggregation of high-resolution building footprints with uncertainty estimates that indicate prediction trustworthiness.

1. Introduction

Identifying building footprints is essential for monitoring urbanization and assessing its environmental impact. Building footprint information provides the basis for analyzing urban structure and impervious surface distribution and can serve as a proxy for population estimation (Huang et al., 2019; Palacios-Lopez et al., 2019; Sapena et al., 2022). Furthermore, such data constitute a key input for understanding urban land use, accessibility, urban heat island effects, air quality, and the distribution of green spaces, among others (Zhou et al., 2018; Taubenböck, 2025). Especially in rapidly growing regions, timely and spatially consistent building footprint information is crucial for analyzing demographic and socioeconomic development. Despite their importance, globally consistent and temporally resolved building footprint datasets remain limited. Existing high-resolution products are often static, regionally incomplete, or heterogeneous in quality, limiting their suitability for time-dependent analyses (Zhou et al., 2022; Prexl and Schmitt, 2023). Yet many applications require not only building locations but also information on when structures emerged. Temporal building data enable change detection and the reconstruction of urban growth processes (Huang et al., 2019; Palacios-Lopez et al., 2019;

* Corresponding author.
Email address: manuel.huber@dlr.de (M. Huber).

https://doi.org/10.1016/j.rse.2026.115413
Received 17 November 2025; Received in revised form 24 March 2026; Accepted 6 April 2026
Available online 10 April 2026

Sapena et al., 2022). Beyond explicit change analysis, temporally attributed footprints are highly relevant for mono-temporal applications. Urban climate modeling, exposure and risk assessment, population mapping, infrastructure planning, land-use analysis and the respective supervised training data generation typically rely on building inventories referenced to a specific date rather than on bi-temporal change layers (Taubenböck, 2025; Zhu et al., 2019; Ahmad et al., 2025; Zhou et al., 2023).

Therefore, identifying building footprints consistently across space and time is essential for monitoring urbanization dynamics and assessing their environmental impacts. However, traditional approaches such as field surveying and manual mapping are often time-consuming, labor-intensive, and costly (Liu et al., 2019), limiting their scalability for large-area or frequently updated assessments. This has led to a significant push toward automation, positioning building footprint segmentation as a key area of focus in both research and development (Liu et al., 2019). The development of building footprint extraction methods has evolved from statistical and threshold-based segmentation toward more advanced techniques such as deep learning. In the realm of deep learning, supervised approaches, in which models are trained on labeled data, were rapidly adopted due to their ability to learn complex, nonlinear patterns. In this context, satellite and aerial imagery have become essential resources for the automated extraction of building footprints, with convolutional deep neural networks (CNNs) playing a central role in this advancement (Liu et al., 2019; Prexl and Schmitt, 2023; Hu et al., 2021; Alshehhi et al., 2017; Neupane et al., 2025).

In general, CNNs have proven their effectiveness in a wide range of image analysis tasks. Among these, the U-Net architecture introduced by Ronneberger et al. (2015) has seen widespread adoption, both in its original form and through numerous adaptations (Hu et al., 2021; Lu et al., 2018; Prexl and Schmitt, 2023; Alshehhi et al., 2017). For example, Lu et al. (2018) proposed a deeper convolutional feature network to improve the delineation of building edges. Prexl and Schmitt (2023) utilized the standard U-Net architecture and achieved strong segmentation results across North America. Shi et al. (2019) enhanced footprint extraction by integrating deep CNNs with graph CNNs, while Hu et al. (2021) introduced a residual recurrent convolutional network incorporating an attention mechanism to further improve segmentation accuracy. More recently, Haghighi Gashti et al. (2024) demonstrated the potential of combining aerial imagery with OpenStreetMap building footprint labels within a U-Net framework to produce high-quality segmentation results alongside uncertainty estimates. A remaining challenge in many of these studies is the reliance on domain (geographical)-specific training, which limits the generalization of the models to regions where labeled data is sparse or unavailable. Hafner et al. (2022) addressed this by investigating the generalization capabilities of a standard CNN in various global locations, comparing and penalizing incorrect labels using predictions from two subnets: one using Sentinel-1 SAR and the other Sentinel-2 multispectral imagery. Still, it is important to acknowledge that alternative architectures, such as Vision Transformers (ViTs) (Liu et al., 2023a; Thisanke et al., 2023) and Mamba-based state space models (Dao and Gu, 2024; Bao et al., 2025; Zhu et al., 2024a), have recently gained strong attention in remote sensing and computer vision tasks. Such models have demonstrated competitive performance across a wide range of applications, including semantic segmentation. However, a recent comparative study indicates that the average improvement over state-of-the-art U-Net variants is often modest, with reported gains of approximately three percentage points in their validation metrics (Gibril et al., 2024). While such improvements are relevant, they need to be interpreted in the context of model complexity, data requirements, and practical deployment constraints. Moreover, probabilistic Bayesian formulations are not yet as well established for Vision Transformer architectures as they are for convolutional networks. In this study, the explicit modeling of aleatoric and epistemic uncertainty is a central component of the methodological framework, and the Bayesian U-Net provides a principled and well-understood way to obtain such uncertainty estimates.

In addition to current methodologies, several built-up area maps are openly available, e.g., GISA (Huang et al., 2022) (30 m; variant 10 m), World Settlement Footprint (WSF) (Marconcini et al., 2020) (10 m), Global Dynamic Land Cover (FROM-GLC) (Gong Peng et al., 2013) (30 m) and Global Human Settlement Layer (GHSL) (European Commission, Joint Research Centre, 2023) (10 m). However, their spatial resolution cannot consistently delineate individual buildings, creating a “resolution gap.” At the other end of the spatial resolution-spectrum are very-high-resolution polygon datasets. The creation of such high resolution building datasets has become a key area of focus for major technology companies, such as Microsoft’s Building Footprints project (Microsoft, 2025), Google’s Open Buildings dataset (Meka and Research, 2024), and the Overture Maps Foundation (OMF) initiative (Maps, 2025). These datasets are typically generated using very-high-resolution aerial or satellite imagery, which, while enabling precise segmentation, is expensive and time-consuming to process (Hu et al., 2021). Another critical limitation of these datasets is the absence or inconsistency of temporal metadata, such as clear timestamps indicating when a building was constructed or became visible. For instance, the OMF aggregates building data from heterogeneous sources—including Google, Microsoft, Meta, OpenStreetMap (OSM), and various government datasets—each with differing standards and limitations regarding metadata (Maps, 2025). In the case of OSM, the available timestamp typically reflects the most recent user edit, rather than the actual construction date of the building. Similarly, building footprints provided by Microsoft or Meta do not capture dates in the metadata.

Many studies emphasize the critical need for temporally disaggregated building footprint data, as existing datasets often lack explicit temporal annotations (Prexl and Schmitt, 2023; Zhu et al., 2024b; Guo et al., 2021; Gevaert et al., 2024; Uhl and Leyk, 2022; Liu et al., 2023b; Ahn et al., 2024; Zhu et al., 2019). In this context, Google’s recently released multi-temporal dataset covering the Global South marks a significant advancement in addressing the temporal dimension of building footprint data (Sirko et al., 2023; Meka and Research, 2024). The dataset covers multi-year raster files representing softmax segmentation outputs from 2016 to 2024, however it is subject to limitations. It relies heavily on costly, proprietary very-high-resolution satellite imagery and incorporates Sentinel-2 image stacks, which limit its suitability for rapid-response applications and make it vulnerable to cloud cover. Another multi temporal built-up dataset was recently published for China, integrating super-resolution methods and semantic segmentation with Sentinel-2 imagery (Liu et al., 2023b). The model was trained on controlled, well-labeled building-footprint data and Sentinel-2 observations. This work demonstrates the value of multi-temporal building-footprint information, but still depends on curated high-resolution building footprint labels to produce a high-quality, stand-alone product—conditions that are not always feasible. Moreover, the inherent coarser resolution of Sentinel-2, even after super-resolution processing, limits the separation of small buildings and thus does not consistently enable a reliable detection of individual structures (Liu et al., 2023b).

In addition to the relevance of temporally disaggregating static building footprint datasets, it is equally important to assess the uncertainty of the segmentation results, especially when deploying models in real-world applications. Probabilistic uncertainty estimation plays a critical role in evaluating the true reliability of model predictions. Most non-probabilistic models provide semantic uncertainty, often inferred from the softmax outputs of deep neural networks, which primarily reflect the model’s confidence in its predictions. In contrast, probabilistic uncertainty estimation allows capturing both epistemic uncertainty (stemming from the model’s parameters or lack of knowledge) and aleatoric uncertainty (arising from inherent noise or ambiguity in the input data) (Sale et al., 2024; Baumann et al., 2023; Gal and Ghahramani, 2016; Hertel et al., 2025). Uncertainty quantification is particularly

important in remote sensing, as datasets are often highly heterogeneous across regions and time. Consequently, both models and their uncertainty estimates must be robust to variations in data quality (He and Jiang, 2023; Ji and Tang, 2025). To develop such probabilistic models, this study employs Bayesian deep learning methods, by incorporating prior distributions and performing Bayesian inference (Shridhar et al., 2019; LaBonte et al., 2019; Hertel et al., 2025). Another approach to derive uncertainty is Monte Carlo dropout (MCD), which gained widespread popularity as a practical approximation of Bayesian neural networks by applying dropout layers during both training and inference (Gal and Ghahramani, 2016; Haghighi Gashti et al., 2024). However, MCD imposes limitations on the flexibility in defining prior distributions, particularly in terms of their mean and variance (Gal and Ghahramani, 2016; Dechesne et al., 2021; LaBonte et al., 2019). Hence, this study employs a fully Bayesian neural network to preserve maximum modeling flexibility.

Despite significant advancements in automated building footprint segmentation using CNNs—particularly U-Net-based architectures—several key challenges remain when scaling these methods globally. A major limitation is that existing building footprint datasets often lack consistent temporal metadata, which not only restricts the analysis of urban development dynamics but also complicates the robust training of models based on incomplete or temporally misaligned reference data. Recent studies address this issue through noisy and weak supervision strategies in remote sensing, including pseudo-label refinement (Mirpulatov et al., 2023), noisy pretraining schemes (Liu et al., 2024), and weakly supervised building extraction frameworks (Chen et al., 2022; Usmani et al., 2023). While these approaches enhance robustness during training, they primarily focus on improving segmentation accuracy under imperfect labels. Our approach builds upon this line of research by generating pseudo-labels and incorporating them as adaptive loss weights during fine-tuning, while additionally integrating aleatoric and epistemic uncertainty in the temporal disaggregation task.

This work addresses the aforementioned challenges by developing a fully open-source, end-to-end pipeline that utilizes Sentinel-2 imagery, integration of a Bayesian U-Net and open-source building footprint data. In addition to methodological advancements, this study extends the approach to a global scale by testing the developed model and temporal disaggregation method across geographically diverse locations. Furthermore, a systematic evaluation is integrated to assess the impact of different training strategies by comparing models trained on local domains with those trained at a global scale. This evaluation considers model and prediction uncertainty in addition to overall segmentation performance. This extensive evaluation is designed to assess both the scalability and generalization capabilities of the method, with reference to prior work.

The structure of the paper is as follows: Section 2 describes the datasets and pre-processing steps. Section 3 elaborates on the proposed method, model architecture, uncertainty estimation and experimental setup. The results and discussion are presented in Section 4. The limitations and future work are presented in Section 5. Section 6 contains the conclusion of this work.

2. Data

This section reviews utilized building datasets and details Sentinel-2 acquisition and pre-processing steps for training, validation and testing.

2.1. Building data

This study leverages four different building datasets for training validation, evaluation, and benchmarking. For training and internal model validation, the OMF building footprint dataset is used. It consists of global 2D polygonal representations of structures and, in some cases, includes associated metadata such as estimated height or usage type when available. All data from OMF are provided under an open license, allowing developers, researchers, and organizations to use and build upon the dataset freely. Therefore, the OMF dataset provides an excellent foundation for this study.

For the evaluation, the SpaceNet 7 Multi-Temporal Urban Development Challenge dataset is employed. This dataset provides independent, multi-temporal, high-resolution building data, including monthly annotations of building footprints across globally distributed locations (SpaceNet, 2025; Van Etten et al., 2021). Fig. 1 shows the centroids of the SpaceNet 7 locations (Mas et al., 2013), of which 60 locations include footprint annotations, while 20 locations do not.

For benchmarking the final product, the Google 2.5D and China Building Rooftop Area (CBRA) datasets are utilized to compare model performance at the overlapping test locations. The Google's 2.5D Open Buildings product (Meka and Research, 2024) provides annual building segmentations for the Global South. The data is provided in raster format with a pixel size of 4 m on an annual basis (Sirko et al., 2023). The CBRA is a multi-annual raster dataset (2016–2021) with 2.5 m resolution, covering the whole of China.

2.2. Sentinel-2 data processing

For imagery, Sentinel-2 data is used, which has been freely available since 2017 and provides consistent, multispectral satellite images. This study adopts an annual time step to generate temporally labeled building footprints. Consequently, Sentinel-2 imagery is processed into yearly composites.

Sentinel-2 Harmonized MSI imagery (European Space Agency (ESA) and Google Earth Engine, 2023) is used for both training and inference. Google Earth Engine (GEE) is used to create a processing pipeline to generate yearly, cloud-free composites for each study location. Cloud masking is performed using the Sentinel-2 QA60 band, which identifies

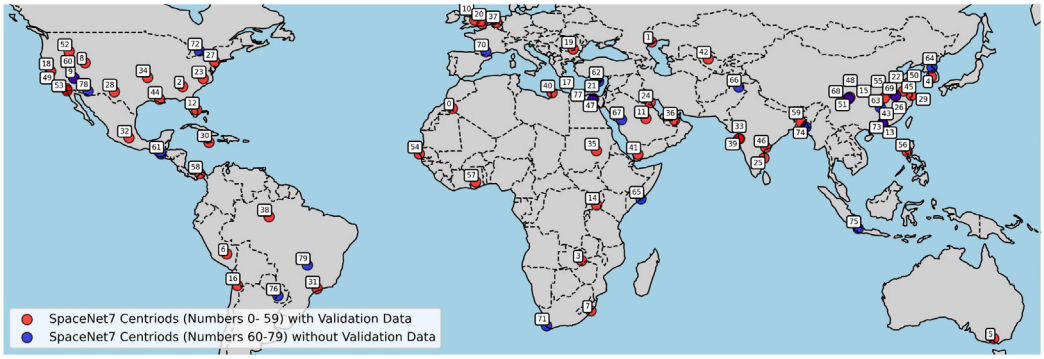

Fig. 1. This figure illustrates a global map, highlighting all SpaceNet 7 testing locations (red, ranging between 0 and 59) and additional locations without testing data (blue ranging between 60 and 79). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

and removes pixels affected by clouds and cirrus (20% cloud cover limit) to retain only clear-sky observations.

The resulting image collections are reduced to a median composite and clipped to the respective Area Of Interest (AOI). The AOI is then subdivided into grid tiles of 256×256 pixels. To ensure a balanced dataset for each tile, the percentage of built-up area is calculated by identifying pixels associated with building footprints, which are all bands greater than or equal to 11, in the GHSL dataset (European Commission, Joint Research Centre, 2023). The GHSL is used in this study due to its direct accessibility via GEE, independence from training tasks and its established application in previous building segmentation research (Hafner et al., 2022). The tiles are then categorized into three urban density classes: 1) High (urban density > 30%), 2) Medium (10% – −30%), Low (5% – −10%). Tiles with building coverage below 5% are discarded to eliminate the sparsely populated tail of the dataset. For each urban density class, up to 1000 representative tiles are selected, prioritizing those nearest to the AOI centroid. If insufficient tiles are available, the AOI is iteratively buffered outward, starting with a 100km radius around the AOI, and the sampling is repeated until the required distribution is met or a maximum of three iterations is reached with each iteration increasing the radius by 50km. This method ensures a spatially diverse and balanced training dataset, which is essential for developing generalizable machine learning models across heterogeneous urban environments.

In the final pre-processing step, four spectral bands—Near-Infrared (NIR), Red, Green, and Blue—are selected based on their highest spatial resolution of 10 meters (GSD). The raw reflectance values of these bands are normalized by dividing by 10,000 and subsequently resampled to a finer spatial resolution of 2.5 meters using bicubic interpolation with an interpolation order of 3. Bicubic interpolation considers a 4 × 4 neighborhood of pixels and computes new pixel values using cubic polynomials in two dimensions. This processing pipeline is inspired by and adapted from previous studies (Prexl and Schmitt, 2023; Hafner et al., 2022). While super-resolution models present an alternative approach to enhance spatial detail, bicubic interpolation has been shown to achieve comparable results to single-image and even multi-channel, multi-image super-resolution techniques in certain remote sensing applications (Razzak et al., 2023). While super-resolution improves the reconstruction of finer spatial textures, such as building edges, this study does not focus on the optimizing delineation, but on exploiting the synergy between high-resolution footprint data and Sentinel-2 segmentation outputs. In this approach the geometric delineation of buildings is primarily controlled by the input footprint dataset, whereas the Sentinel-2-based prediction provides confidence about the presence and temporal attribution of built-up structures.

3. Methodology

Fig. 2 illustrates the overall methodology and its key steps for model training, spatio-temporal domain adaptation, temporal disaggregation and evaluation.

The overall workflow comprises two complementary modeling dimensions: a spatial training strategy (global versus local) and a refinement strategy (base versus master models). First, probabilistic segmentation models are trained either locally (locations-specific) or globally (using data from all available location), resulting in Local Base and Global Base models. In a second step, these base models are spatio-temporally fine-tuned using refined labels and uncertainty-aware training data, yielding the corresponding Local Master and Global Master models (defined per year and location). The final model outputs, namely softmax probabilities and associated uncertainty layers, are then used to temporally disaggregate the original OMF building footprint data. This approach synergistically combines the high spatial resolution of OMF products with the temporal information derived from Sentinel-2-based segmentation. By systematically comparing base and master models at both the global and local scales, the framework further enables an analysis of how predictive performance and uncertainty behavior evolve under different training and generalization strategies.

3.1. Training, validation and testing setup

As mentioned earlier, the SpaceNet 7 dataset is used exclusively to test the model output, which means that its temporal building footprints are not incorporated into training or validation. Still, the SpaceNet 7 dataset serves a dual role: on the one hand, it provides an independent benchmark for testing the trained models; on the other hand, it defines the baseline geographic distribution for selecting training data (OMF and Sentinel-2 imagery) when developing the global base model. Specifically, the global model is trained on data from 70 locations while holding out 10 locations (with temporal annotations) as an external test set. In contrast, the local models are trained exclusively on data available around each individual location (OMF and Sentinel-2), while excluding the corresponding testing tiles. This design simulates a realistic testing scenario in which models must generalize to unseen tiles within a known region.

Table 1 provides an overview of the data ranges per model type (global and local), as well as the amount of data used for spatio-temporal domain adaptation. This number represents the total data input, of which 20 percent is reserved as an external validation dataset - this applies to training the base model and the domain adaptation process. The exact range of tile counts and urban density values is summarized in Table A.8, and varies between each location due to the sampling approach described in the previous section.

In summary, the models are trained using processed Sentinel-2 imagery together with static OMF building footprint labels, which are resampled to match the spatial resolution of Sentinel-2. The model input consists of 256×256 tiles, where each Sentinel-2 tile (2.5 m pixel spacing) is paired with a corresponding binary building footprint mask at the same resolution. Ultimately, these image–label pairs are used to train the individual models, resulting in four distinct variants: Local Base, Local Master, Global Base, and Global Master.

3.2. Bayesian deep neural network

A standard U-Net model produces deterministic outputs (i.e., a single predicted value for each pixel) and, hence, is not able to derive aleatoric and epistemic uncertainty estimates. However, to capture uncertainty in predictions, Bayesian models are more suitable. They allow probabilistic reasoning in their learning process (Gawlikowski et al., 2023). To achieve this, a BCNN is developed on top of the standard U-Net architecture. In a typical neural network, model parameters (such as weights and biases) are learned as fixed values through backpropagation. In contrast, a Bayesian approach introduces probabilistic weights, where the model parameters are treated as random variables with associated probability distributions. To train the Bayesian U-Net, the network is initialized with a prior distribution over the model parameters. This prior represents the initial belief about the parameters before seeing the data. As the model is trained, it updates this prior to a posterior distribution, reflecting the learned knowledge after observing the training data (Hüllermeier and Waegeman, 2021). Additionally, a comparative experiment was conducted between a deterministic and Bayesian U-Net. On average the Bayesian variant achieved a slightly lower mean Intersection-over-Union (0.29) compared to the deterministic model (0.32). However, the Bayesian model provides additional uncertainty outputs that are crucial for downstream quality control, data reliability analysis, and model calibration. Considering the relatively small differences in performance, the Bayesian U-Net was adopted as the main model in this study due to its enhanced interpretability and robustness under varying training strategies.

Fig. 2(b) illustrates the BCNN utilized in this work. It incorporates probabilistic weights by using a prior distribution, allowing the model to capture uncertainty in its predictions. The prior distribution in this study is a Normal distribution with weak/uninformative prior, aligning with

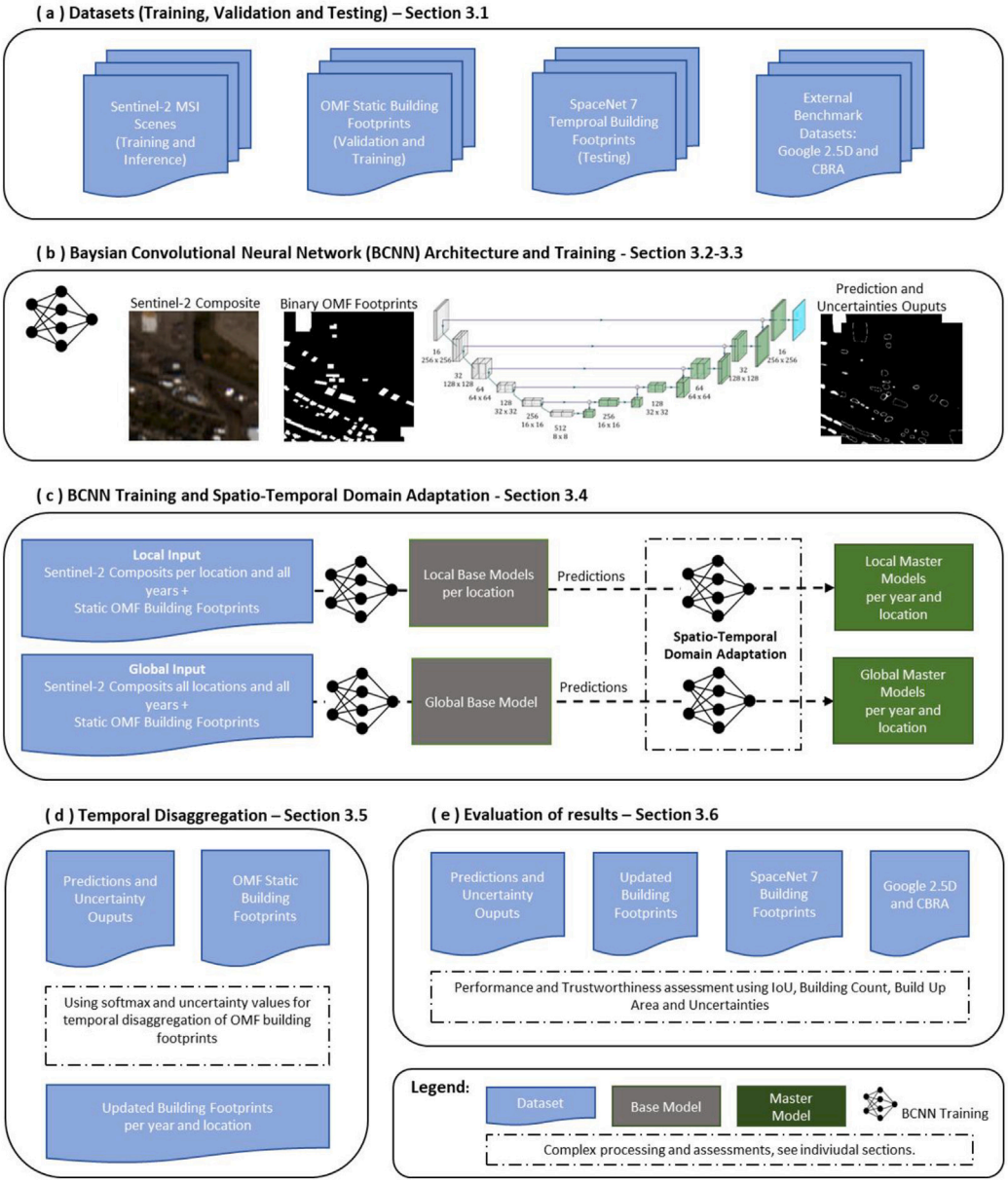

Fig. 2. This figure illustrates the methodology applied in this study, outlining (a) the overall datasets for training, validation and testing, (b) the applied BCNN architecture with input and output examples, (c) the training and spatio-temporal domain adaptation process of the individual BCNN's, (d) the temporal disaggregation step and (e) evaluation of the results.

Table 1
Amount of training data for local and global base and master models. Each number represents a 256x256 tile. * the exact number depends on the location, see Table A.8.

	Amount of training data	
	Base Training	Domain Adaptation
Global Model	7,351,449	183 - 60,974 *
Local Model	1386 - 408,636 *	
Temporal Baseline	2018 - 2024	per year (2018 or 2019)

the approach discussed by Sarma and Kay (2020), who emphasize the importance of incorporating prior knowledge to guide parameter estimation effectively. Further initialization parameters are: 4 input channels (R, G, B, NIR) with two output classes, a kernel size of three for the convolutional layers, with ReLU activation applied after each convolution. Padding is set to "same" to ensure the input and output dimensions remain consistent. The Adam optimizer with an initial learning rate of 0.0001 is used. In addition, a learning rate reduction strategy is applied to avoid plateauing with a decreasing rate of 0.5. The training is done using an NVIDIA A100-SXM4-80GB GPU.

3.3. Uncertainty estimation

The key aspect of the uncertainty estimation is the Bayesian framework using Monte Carlo (MC) approximation during inference. Instead of relying on a single set of predictions, the model can generate multiple predictions using different samples from the posterior distribution of the parameters. From this, three forms of uncertainty are derived: *aleatoric*, *epistemic*, and *total predictive uncertainty*. Still, it should be noted that the following distinction between aleatoric and epistemic uncertainty is not absolute; however, for the purpose of this study, they will be referred to as such. The key aspect of the uncertainty estimation is the

Bayesian framework using Monte Carlo (MC) approximation during inference. Instead of relying on a single set of predictions, the model can generate multiple predictions using different samples from the posterior distribution of the parameters. From this, three forms of uncertainty are derived: *aleatoric*, *epistemic*, and *total predictive uncertainty*.

Aleatoric Uncertainty captures the inherent noise and ambiguity present in the input data itself. It is estimated by computing the average entropy of the individual Monte Carlo predictions. Formally, it is defined as:

$$\mathcal{U}_{\text{aleatoric}}(i) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{H}(\mathbf{p}_i^{(t)}) \tag{1}$$

where $\mathbf{p}_i^{(t)}$ denotes the predicted class probability vector for pixel i from the t-th stochastic forward pass, and entropy $\mathcal{H}$ is defined as:

$$\mathcal{H}(\mathbf{p}) = -\sum_{c} p_c \log(p_c + \varepsilon) \tag{2}$$

with c denoting the class index and $\varepsilon = 10^{-15}$ added for numerical stability.

Epistemic Uncertainty uncertainty reflects the model's uncertainty due to limited knowledge (e.g., from insufficient training data) and can be estimated using the mutual information between the predictions and the model parameters. Based on Monte Carlo sampling, it is computed as the difference between the entropy of the mean prediction and the mean of the entropies:

$$\mathcal{U}_{\text{epistemic}}(i) = \mathcal{H}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_i^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{H}(\mathbf{p}_i^{(t)}) \tag{3}$$

Here, $\mathbf{p}_i^{(t)}$ denotes the predicted class probability vector for pixel i from the t-th Monte Carlo forward pass, and $\mathcal{H}(\cdot)$ denotes the entropy function.

The **total predictive uncertainty** is the sum of aleatoric and epistemic uncertainty:

$$\mathcal{U}_{\text{total}}(i) = \mathcal{U}_{\text{aleatoric}}(i) + \mathcal{U}_{\text{epistemic}}(i) \tag{4}$$

3.4. Spatio-temporal domain adaptation

Fig. 2(c) illustrates the workflow for creating the Local and Global Base models, which are trained in a supervised manner using the previously described Sentinel-2 imagery and raw OMF building footprint annotations. To obtain the Master models, via spatio-temporal domain adaptation, the Base models are further refined by (1) generating base model predictions and (2) reusing these predictions as averaged MC softmax probabilities as pseudo-labels to fine-tune the corresponding Base models. The fine-tuning process is guided by a modified loss function, where the softmax probabilities from the respective Base model are used as loss weights. The weighted binary cross-entropy loss used for fine-tuning the Master models is defined as follows, which is inspired by Ronneberger et al. (2015):

$$\mathcal{L}_{\text{WBCE}} = \frac{1}{\sum_i w_i}\sum_i w_i \cdot \text{BCE}(y_i, \hat{y}_i) \tag{5}$$

In this equation, $\mathcal{L}_{\text{WBCE}}$ represents the weighted binary cross-entropy loss. The term $y_i \in \{0, 1\}$ denotes the pseudo-label for each pixel i, while $\hat{y}_i \in (0, 1)$ corresponds to the predicted probability for the same pixel, typically obtained by applying the sigmoid function to the model output. The weight w_i is derived from the softmax output of the corresponding Base Model and reflects the Base model's confidence that a given pixel belongs to the building class. The binary cross-entropy term, $\text{BCE}(y_i, \hat{y}_i)$, is given by:

$$\text{BCE}(y_i, \hat{y}_i) = -\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{6}$$

By weighting each pixel's loss contribution according to w_i and normalizing over the total sum of weights, the model ensures, that more probable building pixels have a stronger influence during the fine-tuning process. This approach enhances the model's sensitivity to temporally and spatially relevant features in the imagery. As a result, this approach leads to the development of year-specific Master Models for both the Global and Local domains, allowing the model to better account for temporal variations in building presence and morphology.

3.5. Temporal disaggregation procedure

Once all model variants are trained, predictions and uncertainty estimates are generated for the test locations. This results in raster outputs for each test location, containing the *softmax* probabilities from the BCNN as well as *aleatoric* and *epistemic* uncertainty maps. Based on these outputs, the temporal disaggregation procedure proceeds through several steps: *Binary mask creation*, *Vectorization*, *Overlay with OMF footprints*, *Zonal statistics* and *Temporal assignment*.

The binary mask creation is formulated as a semantic segmentation task. The softmax output probabilities are converted into a binary prediction map using a fixed threshold of 0.5, which is commonly adopted in related building extraction studies (Liu et al., 2023b; Liu and Tang, 2023). The resulting binary raster is subsequently vectorized using the rasterio.features.shapes function from the Rasterio Python library, which extracts polygon geometries from connected pixel regions of the binary mask. To spatially align the prediction layer with the static OMF building footprints, the vectorized prediction polygons are overlaid with the OMF footprint layer using a geometric intersection operation. Matching between predicted and static polygons is determined via spatial overlap. If a predicted polygon intersects an existing OMF building polygon, the OMF polygon is retained, whereas if a predicted polygon does not intersect any OMF polygon, it is treated as a candidate for a newly detected building footprint. This spatial association is rule-based and relies on geometric intersection rather than a learnable parameter. No additional IoU-based matching threshold is applied at this stage, since the final decision about retaining or discarding polygons is handled downstream by the supervised XGBoost classifier. After the overlay step, the non-thresholded prediction raster (i.e., the softmax and uncertainty layers) is clipped to both the retained static polygons and the newly added polygons. Zonal statistics are then computed per polygon, yielding the mean softmax probability as well as aleatoric and epistemic uncertainty estimates for each footprint.

Since no single physically meaningful decision boundary exists to directly combine softmax probabilities with aleatoric and epistemic uncertainty estimates, a supervised classification step is introduced for the final polygon selection. For this purpose, each building footprint is represented by polygon-level features, including the mean softmax probability and the averaged uncertainty estimates derived from the Bayesian network outputs. Using the testing building polygons as reference, binary labels are assigned to each candidate footprint. A polygon is labeled as true if it overlaps with a testing polygon by at least 10%, and false otherwise. Different overlap thresholds were tested, but lower thresholds yielded more stable and robust results, as they better account for minor geometric misalignments between predicted and reference footprints. For this purpose, a global classification threshold is learned using an XGBoost classifier trained jointly across all test locations. XGBoost is a gradient-boosted decision tree algorithm that has demonstrated strong performance, computational efficiency, and robustness across a wide range of classification tasks, including land cover mapping and remote sensing applications (Georganos et al., 2018). In addition to its predictive accuracy, XGBoost enables transparent feature importance analysis, allowing a quantitative assessment of the contribution of uncertainty measures relative to softmax confidence and other auxiliary features. The classification performance is assessed using the Area Under the Receiver Operating Characteristic Curve (AUC) and the F1 score, which jointly capture ranking performance and classification balance under class imbalance. These metrics are first computed at the location level and then aggregated by averaging across all locations to ensure spatially

balanced evaluation. To assess the robustness and generalizability of the classification step, additional experiments are conducted, including data bootstrapping, limited hyperparameter sensitivity analysis, and a systematic comparison between location-specific XGBoost models and a single global model trained across all regions. The final trained XGBoost model is then applied to generate the definitive labeling of the combined building footprint dataset, resulting in three building footprint categories:

- **Existing footprint**: A static building polygon is confirmed when the mean raster prediction value within it exceeds the global threshold.
- **False footprint**: A static polygon exists, but the model prediction does not meet the threshold or a static polygon does not contain any prediction values (empty).
- **New footprint**: The model predicts a building with high confidence in an area not covered by any existing static footprint.

3.6. Evaluation of results

The evaluation of this work is organized into four core questions: (1) How do model performance and model uncertainty vary between different training strategies? (2) What are the generalization capabilities of the global model?, (3) How does the temporal disaggregation improve the final segmentation with respect to IoU, building count and built-up area?, and, (4) How does the model compare to existing state-of-the-art products?

All models (corresponding to different training strategies) are assessed with respect to both, segmentation performance and predictive uncertainty. By contrasting local and global models, new insights are provided into uncertainty behavior and the reliability of predictions across diverse urban structures. The generalization capacity of the global model is tested on ten designated hold-out locations, representing unseen geographic regions, while local models are evaluated within their respective domains.

To benchmark performance against the state of the art, predictions are further compared to the Google 2.5D multi-temporal dataset and the CBRA dataset, covering locations in the Global South and China, respectively. Testing is conducted both at the pixel level and at the level of aggregated building counts and built-up area, allowing us to assess spatial accuracy as well as completeness in structure detection.

IoU is used to quantitatively evaluate the segmentation performance of the models. The IoU is a well established metric to assess how well the predicted segmentation masks align with the ground truth annotations and is used in many previous studies (Prexl and Schmitt, 2023; Hafner et al., 2022; Haghighi Gashti et al., 2024; Yu et al., 2023). It is defined as a measure of overlap between the predicted foreground region and the ground truth, normalized by their union. A higher IoU indicates a greater degree of overlap between prediction and ground truth. Although IoU primarily measures spatial overlap, it is appropriate here because the objective is not classical change detection but the generation of a temporally attributed footprint layer for each timestamp. Performance is therefore evaluated on the complete footprint map per year rather than only on newly added or removed buildings. The raw OMF dataset serves as a naïve baseline without temporal disaggregation, and improvements are assessed by comparing IoU, building count, and built-up area between the original and updated layers.

The model uncertainty is estimated using a cost-based approach. This cost function is used to quantify the per-pixel uncertainty in the model's predictions and is based on the work of Yang et al. (2016). Let $\hat{y}_i \in [0, 1]$ denote the predicted probability (average softmax output of 32 model runs) of pixel i belonging to the foreground class, and let $y_i \in \{0, 1\}$ represent the corresponding ground truth label. The cost for each pixel is calculated using the binary cross-entropy formulation:

$$\text{Cost}_i = -y_i \cdot \log(\hat{y}_i) - (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{7}$$

This cost captures the model's confidence in its predictions. Higher costs indicate greater uncertainty or error in the predicted probability for a given pixel. From this, the average cost is then given by:

$$\text{Average Cost} = \frac{1}{N} \sum_{i=1}^{N} \text{Cost}_i \tag{8}$$

where N is the total number of pixels. This cost-based metric not only reflects prediction confidence at a fine-grained level, but also enables a comparative analysis of uncertainty across different model types.

4. Results and discussion

4.1. How do model performance and model uncertainty vary between different training strategies?

To assess the quality of segmentation results, the IoU is estimated across all models and both testing years —2018 (green) and 2019 (red). Fig. 3 presents kernel density estimation (KDE) plots of IoU distributions per model and year across all testing locations.

Fig. 3, shows that the mean IoU values (vertical lines) are generally higher for 2019 than for 2018, except for the local base model (c), where the IoU remains nearly constant. A contributing factor to this difference could be the improved temporal alignment between the training labels and the input imagery. Given the overall trend of urban growth, if the OMF building footprints correspond to later timestamps, the model may develop a bias toward more recent structures. This effect is also reflected in the wide IoU distribution between the raw OMF input and the ground-truth testing data. In general, the base predictions (a) and (c) exhibit lower mean IoU values and a wider spread across all locations. In contrast, both the local (d) and global (b) master models show a concentration of IoU values in the higher range, indicating that the spatio-temporal domain adaptation substantially improves model performance.

Table 2 presents aggregated metrics over all testing locations, including IoU, average cost, and both types of uncertainty (epistemic and aleatoric), separately reported for building and non-building pixels. Overall, the base models exhibit slightly lower uncertainty and cost values for both building and non-building classes. However, these improvements in uncertainty and cost come at the expense of lower accuracy, as reflected by lower IoU scores. In contrast, the master models, particularly the global master model, achieve the highest IoU, indicating superior segmentation performance. Comparing the global master model with the local master model IoU, the global master model benefits from training on a substantially larger and more diverse dataset prior to site-specific adaptation. This broader data distribution enables the global master models to learn more generalized spatial and spectral building representations, which are subsequently refined through fine-tuning. In contrast, the local master model is initialized from a more homogeneous and site-specific baseline, which may increase the risk of overfitting and amplify local label noise during fine-tuning. Despite slightly elevated uncertainty metrics, the global master also maintains low average cost, suggesting confident and well-calibrated predictions overall. Another important observation is that the spatio-temporal adaptation applied in the master models appears to introduce slightly higher uncertainty, especially compared to their base counterparts. This increase may be interpreted as a trade-off between accuracy and uncertainty. One likely explanation is that the adaptation process involves fine-tuning with potentially imperfect or temporally misaligned training labels, which can introduce variability and noise into the learning process. After examining aggregated model cost, performance, and uncertainties individually, it is also important to explore how these core metrics relate to each other and to external variables that could influence them, such as structural variabilities as represented by various urban densities and the number of training tiles. This is particularly relevant when comparing local models with the global master model, since both differ strongly in terms of available training data.

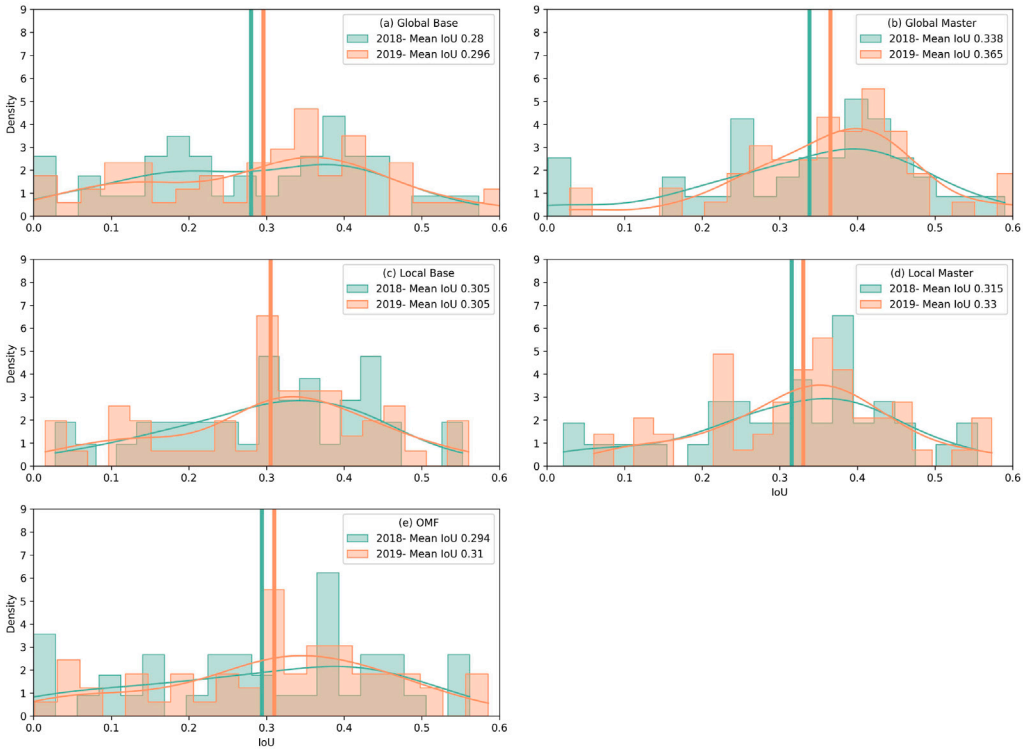

Fig. 3. This figure shows histogram plots of Intersection over Union (IoU) scores for all testing locations across both time steps —2018 (green) and 2019 (red)— segmented by model type. Each histogram illustrates the distribution of IoU values, while the vertical dashed lines indicate the mean IoU for each year and model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 2
Aggregated IoU, average cost, and uncertainty per model type. (Unc. = Uncertainty, GM = Global Master Model, GB = Global Base Model, LM = Local Master Model, LB = Local Base Model). Bold values indicate the best-performing metric per column. Arrows denote whether lower (↓) or higher (↑) values are preferable for interpretation, with the bold value representing the best-performance.

Metric	GB	GM	LB	LM
IoU	0.290	**0.351**↑	0.305	0.324
Aleatoric Unc. Building	**0.398**↓	0.528	0.442	0.541
Aleatoric Unc. None-Building	**0.023**↓	0.032	0.031	0.046
Epistemic Unc. Building	**0.039**↓	0.046	0.049	0.056
Epistemic Unc. None-Building	**0.005**↓	0.007	0.013	0.018
Total Unc. Building	**0.437**↓	0.573	0.491	0.597
Total Unc. None-Building	**0.028**↓	0.040	0.044	0.065
Average Cost	**0.059**↓	**0.059**↓	0.062	0.070

Fig. 4 shows the correlation matrices for each model - (a) Local Base, (b) Local Master, (c) Global Base, and (d) Global Master — illustrating how performance (IoU), model uncertainty, and average cost are related to the availability of training data, label quality and urban density. Label quality refers to the IoU between the raw OMF input and the testing data. Strong positive correlations are shown in red, negative correlations in blue, and weak/no correlations in white. Firstly, in Fig. 4, urban density can be interpreted as a proxy for scene complexity, reflecting the transition from rural to highly urbanized environments. Across all model types, the correlation matrix shows that urban density is positively correlated with non-building uncertainty and average cost. At the same time, urban density also correlates positively with IoU. One explanation could be, that in the case of very high urban density, much more of the non-building pixels are bordering building pixels, which is exactly the zone at which the model expresses the highest uncertainty. In addition, only a very weak correlation between aleatoric uncertainty and the number of training tiles is seen. This is consistent with Kendall and Gal (2017), who describe heteroscedastic uncertainty, a form of aleatoric uncertainty, that depends on the input data itself and cannot be reduced simply by adding more data. On the other hand, a strong negative correlation can be observed between epistemic uncertainty and the number of training tiles, for both building and non-building pixels. This aligns with findings of previous studies (Kendall and Gal, 2017; Ji and Tang, 2025), which confirm that epistemic uncertainty can be reduced by including more training data. Another interesting observation is the correlation between IoU and the number of training tiles. The number of training tiles shows a three times weaker correlation with the IoU of the global master model, compared to the IoU from the local base and local master models. This suggests that while additional data effectively reduces uncertainty, its effect on pure performance (IoU) is less direct and may depend on other factors such as spatial diversity, label quality or representativeness of the training samples. Label quality is the dominant explanatory variable across all model types in regard to IoU. Its regression coefficient is positive and highly significant in every case, indicating that improvements in label completeness directly translate into higher IoU scores. This effect is strongest for the global base model, where label quality alone explains a large fraction of the observed performance variance. The fine-tuned master models exhibit reduced dependence on label quality compared to their base counterparts. This indicates that fine-tuning mitigates error propagation from imperfect labels, making these models more robust in regions with noisy or incomplete footprint data.

Fig. 5 illustrates selected relationships between training tile count, building density, and the derived performance and uncertainty metrics, highlighting in more detail some of the significant correlations already discussed. Each scatter plot aggregates values per location and year, with quartile-interpolated lines included for each model type. To account for the wide variation in training samples, the training tile axis is shown on a logarithmic scale.

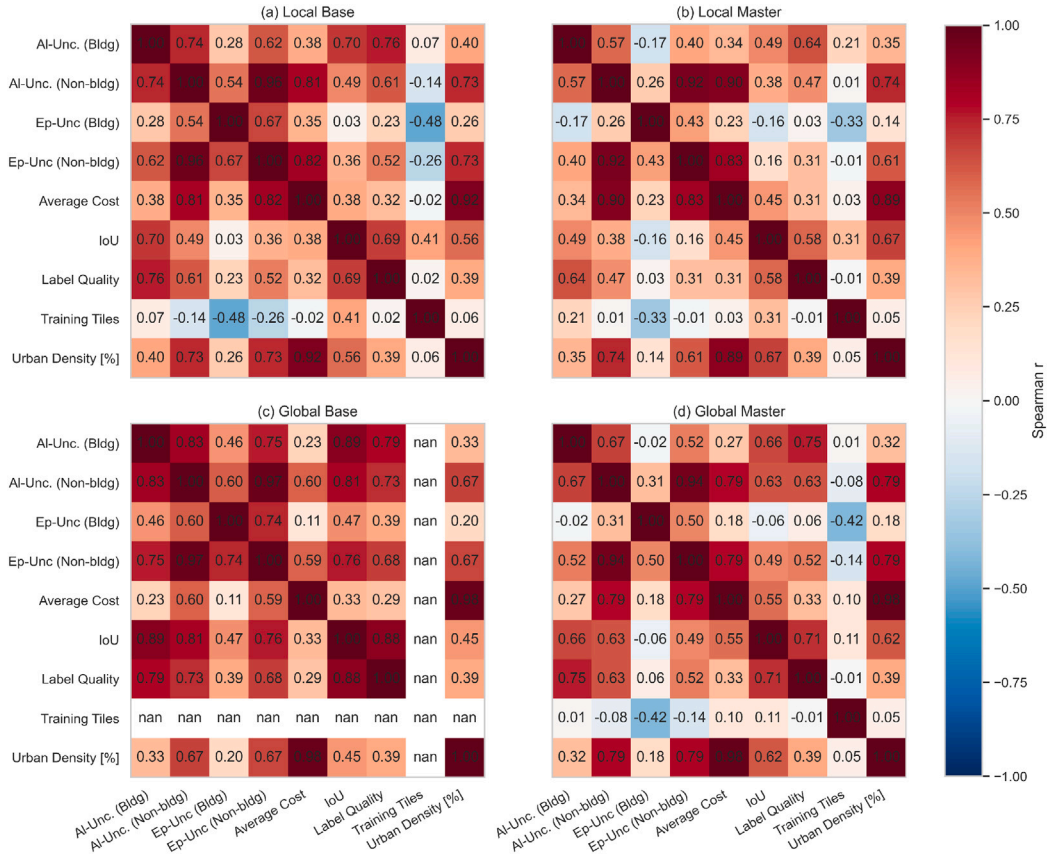

Fig. 4. This figure, shows the correlation matrix for each model — (a) local Base, (b) local Master, (c) global Base, and (d) global master — illustrating relationships between derived uncertainties, average cost, IoU, label quality (which is the IoU of the input footprints), number of training tiles, and urban density - using the Spearman correlation coefficient.

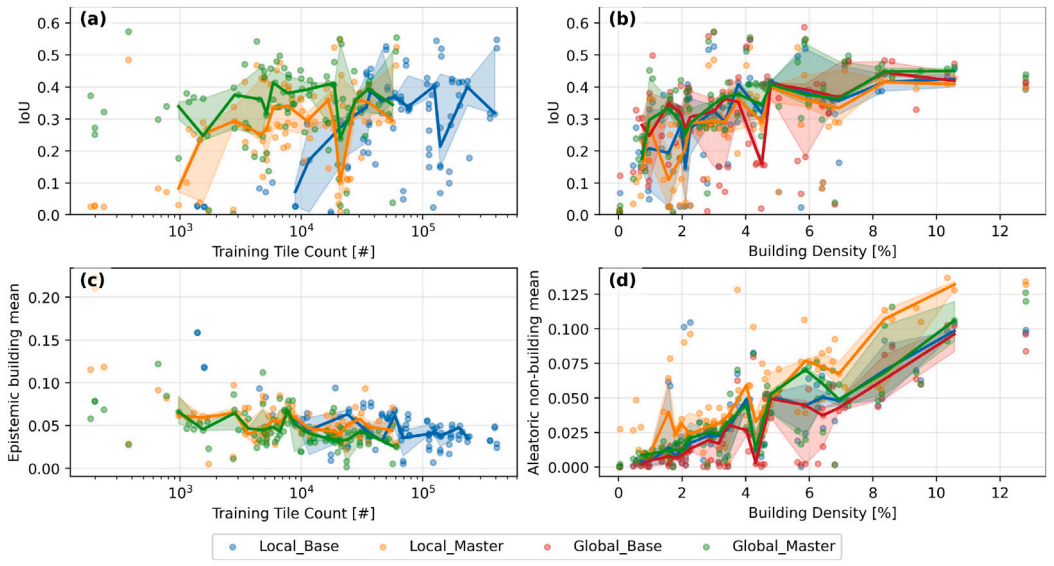

Fig. 5. This figure, shows histogram plots of Intersection over Union (IoU) scores for all testing locations across both time steps —2018 (green) and 2019 (red)— segmented by model type. Each histogram illustrates the distribution of IoU values, while the vertical dashed lines indicate the mean IoU for each year and model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Panel (a) shows the relationship between IoU and training tile count. For the global master model, the number of training tiles has little relevance in further improving IoU, confirming that generalization is mainly driven by the diversity of training locations rather than their sheer size. In contrast, the local master model exhibits a saturation effect at around 1500 tiles, while the local base model requires more data (approaching 50,000 tiles) before IoU gains level off. These thresholds are not absolute, as performance also depends on factors such as spectral variability

between training and test data, the balance of urban density classes in training, and label quality. Nonetheless, the patterns suggest that IoU can be controlled by training data quantity up to a point, after which diminishing returns occur.

Panel (b) compares building density with IoU, reflecting the earlier finding (Fig. 4) that higher degrees of urban density correlate with higher IoU. This suggests that models tend to perform better in dense urban areas, likely due to more distinct spectral and spatial patterns of buildings compared to rural environments. However, this comes at the cost of higher uncertainty, indicating a trade-off between accuracy and confidence. Notably, the lowest IoU values are observed for tiles with 0–2% urban density, likely because the models are trained on samples with a minimum of 5% urban density. This highlights the potential benefit of incorporating lower-density tiles into the training process to improve generalization.

Panel (c) shows a strong negative correlation between training tile count and epistemic building uncertainty across all models, confirming the well-established notion that epistemic uncertainty decreases with more data (Kendall and Gal, 2017; Ji and Tang, 2025).

Panel (d) shows a clear positive relationship between aleatoric non-building uncertainty and building density. This is consistent with the understanding that aleatoric uncertainty reflects data-inherent noise and class ambiguity (Kendall and Gal, 2017). In dense urban areas, mixed pixels (e.g., at building edges, narrow streets, or vegetation between structures) are more frequent in Sentinel-2 imagery, leading to higher uncertainties for both background and building pixels.

To further disentangle the distribution of aleatoric uncertainty, an additional analysis was conducted using NDVI bins derived from the corresponding Sentinel-2 imagery. Four NDVI classes were defined as land-cover proxies following established practice (Idrees et al., 2022): NDVI < 0.1 representing Clouds, Shadows, and Water (CSW), used as a proxy for non-vegetated dark surfaces such as water bodies, cloud shadows, or built-up areas with low reflectance; NDVI values between 0.1 and 0.25 representing Bare Soil or Sand, corresponding to dry soil, sand, or impervious surfaces; NDVI values between 0.25 and 0.5 representing Sparse Vegetation, typically associated with semi-arid grasslands, croplands, or transitional vegetation cover; and NDVI values greater than 0.5 representing Dense Vegetation, including forests, healthy vegetation, or irrigated agricultural areas with full canopy cover.

Fig. 6(a) illustrates how both the magnitude and distribution of aleatoric uncertainty vary across these NDVI classes and model types. Two key patterns emerge. First, across all models, aleatoric uncertainty is highest over bare soil regions. These areas are characterized by strong spectral mixing between built-up surfaces, soil, and surrounding land-cover types, which increases pixel-level ambiguity. In contrast, dense vegetation consistently exhibits the lowest uncertainty values, likely because its spectral signature is well separated from built-up areas. This confirms that aleatoric uncertainty effectively captures uncertainty arising from mixed pixels and spectral overlap. Second, the model type influences robustness across surface classes. The base models generally exhibit lower aleatoric uncertainty than the fine-tuned master models, particularly in ambiguous NDVI classes such as bare soil and sparse vegetation. This suggests that the spatio-temporal fine-tuning process increases the model's sensitivity to local spectral variability, resulting in higher—but more informative—aleatoric uncertainty estimates.

Fig. 6(b) further explains the relationship between mean softmax confidence and mean aleatoric uncertainty across NDVI classes and model types. The scatter plot shows an inverse relationship: higher aleatoric uncertainty corresponds to lower softmax confidence. In the context of binary building classification, higher softmax values indicate greater confidence that a pixel belongs to the building class. Consistent with expectations, the CSW class shows relatively high softmax values, reflecting the frequent presence of built-up structures in low-NDVI urban cores. In contrast, sparse and dense vegetation classes exhibit very low softmax values across all model types, demonstrating robust separation between vegetation and buildings. Bare soil emerges as the most challenging land-cover class for all models, characterized by both elevated softmax values and high aleatoric uncertainty. Notably, the master models show systematically higher aleatoric uncertainty than their respective base models while simultaneously assigning higher softmax values. Overall, the results indicate that aleatoric uncertainty estimates therefore provide valuable auxiliary information and should be explicitly leveraged in the subsequent temporal disaggregation step to improve robustness in spectrally ambiguous regions.

4.2. Generalization capability of global models over unseen test locations

To assess the generalization capability, a set of 10 test locations, unseen during training, is reserved for evaluation. Table 3 presents the IoU, uncertainty estimates, model cost and urban density across all model types for these locations.

The results demonstrate that the Global Master model consistently delivers the strongest performance, achieving the highest mean IoU across the test locations. Remarkably, its average IoU on these unseen locations even exceeds the global mean IoU, underscoring the model's robust generalization capacity. Across the ten test locations, which range between 1.8 and 9 percent urban density, no strong negative correlation between urban density and IoU is observed. This suggests that, beyond geographic generalization, the model is capable of handling varying levels of urbanization within unseen domains. In terms of

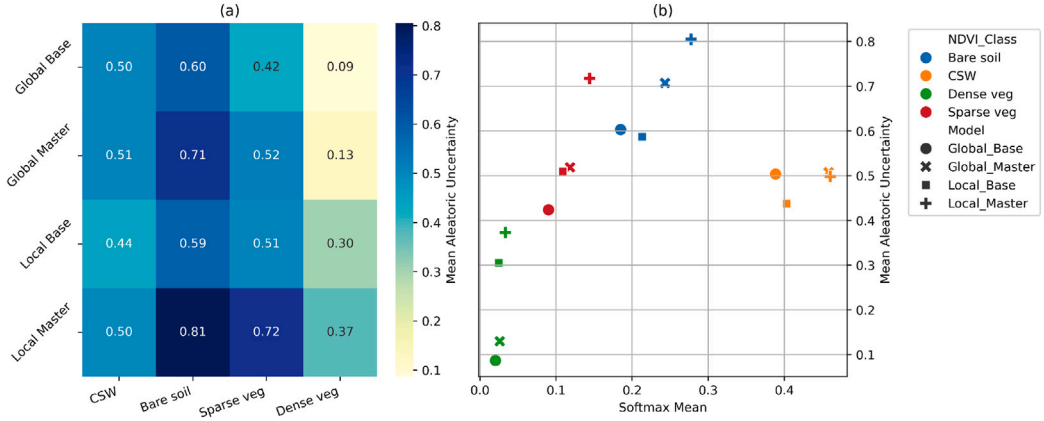

Fig. 6. Sub-Figure (a) shows aleatoric uncertainty across NDVI-based land-cover classes (CSW: NDVI < 0.1; Bare Soil: 0.1–0.25; sparse Vegetation: 0.25–0.5; dense Vegetation: > 0.5) for base and master models. Sub-Figure (b) shows mean softmax confidence versus mean aleatoric uncertainty across NDVI classes for all model types.

Table 3
Comparison of IoU and uncertainty-based cost metrics for 9 testing locations. (GB = global Base Model, GM = global master Model, T.Unc = total predictive uncertainty, Bld. = building pixel, Non-Bld. = None building pixel, Loc. = Location) Arrows denote (↑) higher values are preferable and (↓) lower values are preferable, with the bold value representing the better-performing model in each case.

GB IoU	GM IoU	GB T.Unc.Bld.	GM T.Unc.Bld.	GB T.Unc.Non-Bld.	GM T.Unc.Non-Bld.	GB Average Cost	GM Average Cost	Urban density	Loc.
0.33	**0.42**↑	**0.479**↓	0.592	**0.074**↓	0.077	0.146	**0.129**↓	9.4	2
0.30	**0.43**↑	**0.452**↓	0.706	**0.014**↓	0.021	0.029	**0.028**↓	1.8	8
0.46	**0.48**↑	**0.649**↓	0.677	**0.054**↓	0.056	**0.089**↓	0.087	7.1	10
0.23	**0.38**↑	**0.308**↓	0.607	**0.007**↓	0.027	**0.038**↓	0.041	2.7	14
0.01	**0.16**↑	**0.015**↓	0.231	**0.001**↓	0.014	0.081	**0.076**↓	4.3	26
0.22	**0.43**↑	**0.316**↓	0.737	**0.022**↓	0.053	0.068	**0.065**↓	4.1	31
0.40	**0.44**↑	**0.554**↓	0.615	**0.061**↓	0.066	0.119	**0.113**↓	8.2	32
0.39	**0.39**↑	**0.598**↓	0.612	**0.018**↓	0.019	**0.030**↓	0.031	2.1	44
0.30↑	0.29	**0.578**↓	0.648	**0.029**↓	0.039	**0.039**↓	0.045	2.3	56

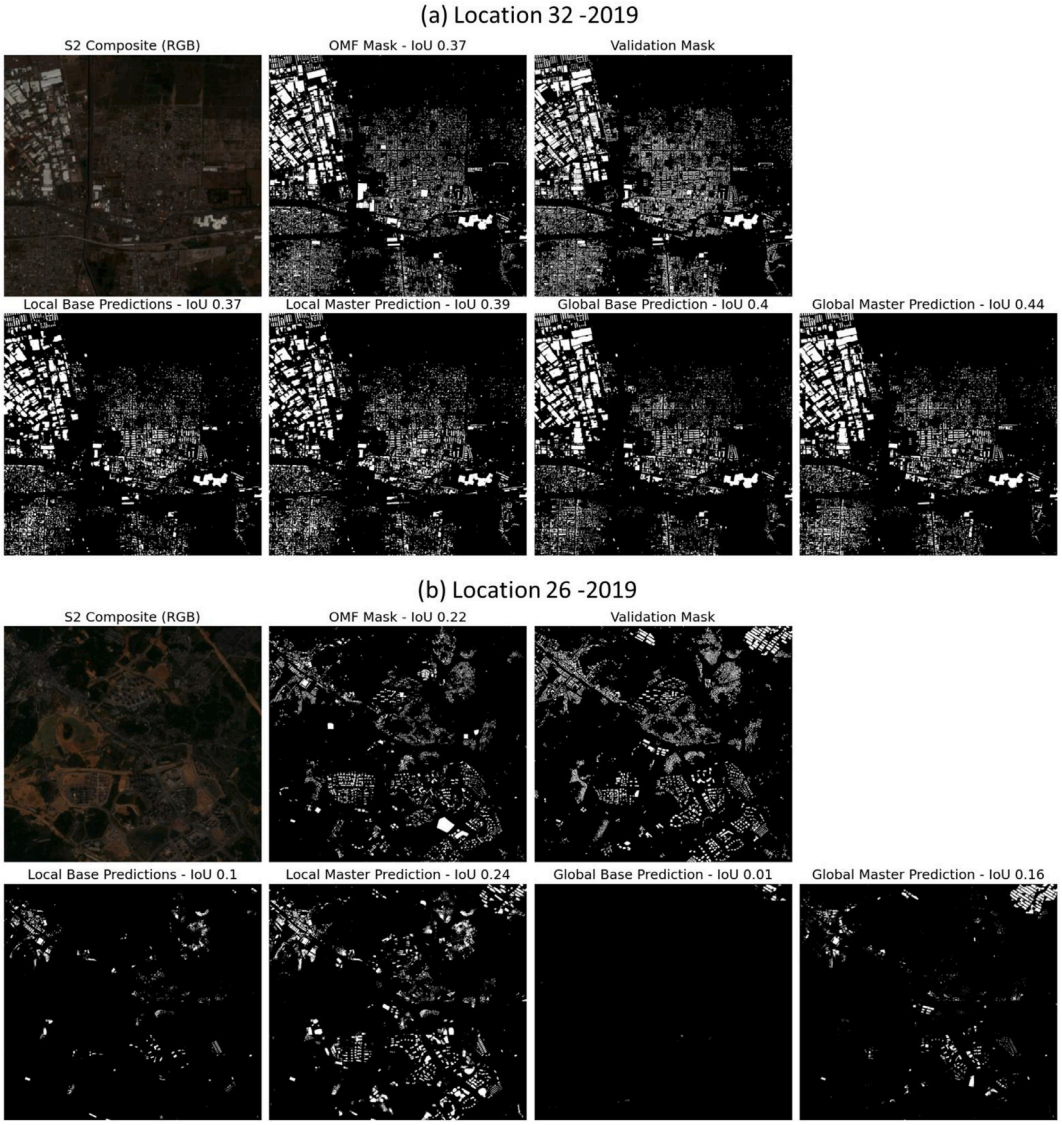

Fig. 7. This figure shows spatial prediction results for testing (a) location 32 (2018) and (b) location 26 (2019). The figure presents a false-color Sentinel-2 RGB composite, followed by the raw Overture Maps Foundation building footprint mask, the testing mask, and segmentation outputs from the local Base, local Master, global Base, and global master models. Each prediction map includes the respective Intersection over Union (IoU) value.

predictive uncertainty, the values are slightly higher compared to the global mean of the predictive uncertainties, both for the global base and Global Master models. Nevertheless, the recurring pattern of higher uncertainties being associated with higher IoU values, and vice versa, is confirmed. For example, location 26 exhibits very low IoU scores for both the global base and Global Master models, while also showing the lowest uncertainty values.

Fig. 7 further illustrates the prediction results, including both local models and the OMF input dataset. Two representative locations are presented: one corresponding to one of the best-performing examples

((a) Location 32, 2018) and another highlighting one of the weakest performances ((b) Location 26, 2019).

In both cases (Fig. 7(a), b), the input data, ground-truth testing masks, and prediction outputs for all model types are shown. In the best-performing case (Fig. 7(a), Location 32, 2018), the OMF footprint labels used as input are of relatively high quality (IoU = 0.37), an aspect that is not consistent across all regions. The predictions from the master models (both global and local) exhibit noticeably sharper and more distinct building edges compared to their respective base models, underlining the benefits of spatio-temporal domain adaptation. These improvements are particularly visible in the delineation of individual building boundaries. There the master models separate adjacent structures more accurately. There is also an improvement in the reduction of false positives such as roads, parking lots, and open areas misclassified as buildings. In contrast, the weakest-performing case (Fig. 7(b), Location 26, 2019), highlights several limitations. Here, the OMF input labels are of poorer quality (IoU = 0.22) compared to Location 32, with inconsistencies and missing footprints, which strongly influence the model predictions. Both global and local models struggle to capture the underlying building structures, with reduced IoU and higher uncertainty values.

Still, poor OMF annotations do not fully explain why the models perform worse, as model performance depends on multiple variables. Beyond training data quality, factors such as spectral differences between images or the contrast of built-up structures against their surrounding environment play a crucial role. This can be visually interpreted from the Sentinel-2 composites: at location 32, buildings are easy to separate from their surroundings, whereas at location 26 they blend in much more strongly, making segmentation considerably harder. Such cases highlight the difficulty of predicting in advance where the model will perform well or poorly. However, the strong correlations observed consistently between low uncertainty (particularly aleatoric and total uncertainty) and low IoU suggest a practical solution: very low uncertainty could serve as a proxy to flag potentially unreliable predictions. After analyzing all model performance, associated uncertainties, and generalization capabilities, master models achieve the strongest results, particularly in terms of segmentation accuracy. Hence, global and local master models and their predictions are used in the continuation of this paper, applying the proposed temporal disaggregation approach to assign time labels to static high-resolution building footprints.

4.3. How does the temporal disaggregation improve the final segmentation in regard to IoU, building count and built-up area?

As described in the methodology section, the temporal disaggregation procedure combines the model prediction outputs with a very-high-resolution building footprint dataset to assign temporal labels to static polygons. A supervised XGBoost classifier is employed to produce a binary decision for each polygon, integrating softmax probabilities and uncertainty estimates derived from the Bayesian U-Net. To ensure a robust evaluation of the classifier, a bootstrap analysis was conducted. In total, 20 bootstrap iterations were performed, where for each run a random 30% subset of the data was withheld for testing while the remaining data were used for training. This procedure resulted in a mean AUC of 0.862 with a standard deviation of 0.00008, indicating a highly stable decision boundary with respect to variations in the training data. In addition, a hyperparameter sensitivity analysis was conducted on the best-performing configuration using a grid search over 430 parameter combinations. The grid varied the number of estimators, maximum tree depth, learning rate, subsampling ratio, and column subsampling ratio. The resulting performance variability showed a standard deviation of 0.00168 in AUC, confirming that the model performance is largely insensitive to moderate hyperparameter changes. Based on this analysis, the final model configuration was set to: number of estimators: 200, maximum depth: 6, learning rate: 0.05, subsample: 0.6 and column sample by tree: 0.8. To provide practical guidance on model deployment, a single general (using all model outputs) trained XGBoost classifier was evaluated by comparing it to model-specific classifiers trained separately for each model type. The results demonstrate that the globally trained classifier yields only marginal performance differences (0.3% AUC deviations for the global base models and 1.3% for local models) compared to specialized models, supporting its use as a single general solution for downstream disaggregation.

Table 4
Ablation study comparing softmax-only, uncertainty-only, and combined feature configurations in regard to AUC, Accuarcy and F1 performances. Arrows denote (↑) higher values are preferable.

Model configuration	AUC	Accuracy	F1 Score
Softmax only	0.8105	0.7654	0.8084
Uncertainty only	0.7518	0.7238	0.7777
Softmax and uncertainty	**0.8329**↑	**0.7722**↑	**0.8167**↑

Table 5
Relative feature importance of the global XGBoost classifier expressed as percentage contribution to total gain.

Feature	Importance (%)
Softmax mean	48.7
Polygon area (m^2)	29.2
Aleatoric uncertainty mean	16.5
Epistemic uncertainty mean	7.4

Finally, an ablation study was conducted to quantify the contribution of uncertainty-related features relative to softmax outputs alone. Table 4 summarizes the AUC, accuracy, and F1 score for three feature configurations: softmax mean only, uncertainty features only (aleatoric and epistemic), and a combined setting using softmax and uncertainty features. The results demonstrate that incorporating uncertainty information consistently improves all performance metrics compared to using either feature group in isolation.

To further investigate the relative importance of individual features, gain-based feature importance scores were extracted from the trained XGBoost model. Table 5 reports the normalized contribution of each input variable. Aleatoric uncertainty accounts for approximately 16.5% of the total information gain, while epistemic uncertainty contributes about 7.5%, indicating that both uncertainty types provide meaningful but distinct signals for the classification task. Softmax Mean has the largest feature importance gain of 48.7% whereas polygon area contributes 29.2% of the total information gain. This suggests that geometric characteristics play a crucial role in the decision boundary, particularly for pixels with a lot of ambiguity, where prediction confidence alone may be insufficient.

The evaluated model was then applied to drive the temporal disaggregation, where the static building footprints are updated by removing false positives, retaining valid structures, and adding new detections. To evaluate the updating process at the building level, predicted and testing polygons are compared directly for each test location. Polygons with any spatial overlap are flagged as true, while those without intersection are flagged as false. This simple overlap criterion introduces a limitation, as it does not guarantee good agreement in shape or area. To address this, the evaluation is separated into two complementary components: (1) building count comparison and (2) building area comparison.

Fig. 8 presents the first part of this evaluation. The left panels (a, c) show the relative distribution across five categories using a boxplot:

1. New polygon correctly added (True Positive),
2. Original polygon correctly retained (True Positive),
3. New polygon falsely added (False Positive),
4. Original polygon falsely removed (False Negative),
5. Total updated building count relative to the ground truth.

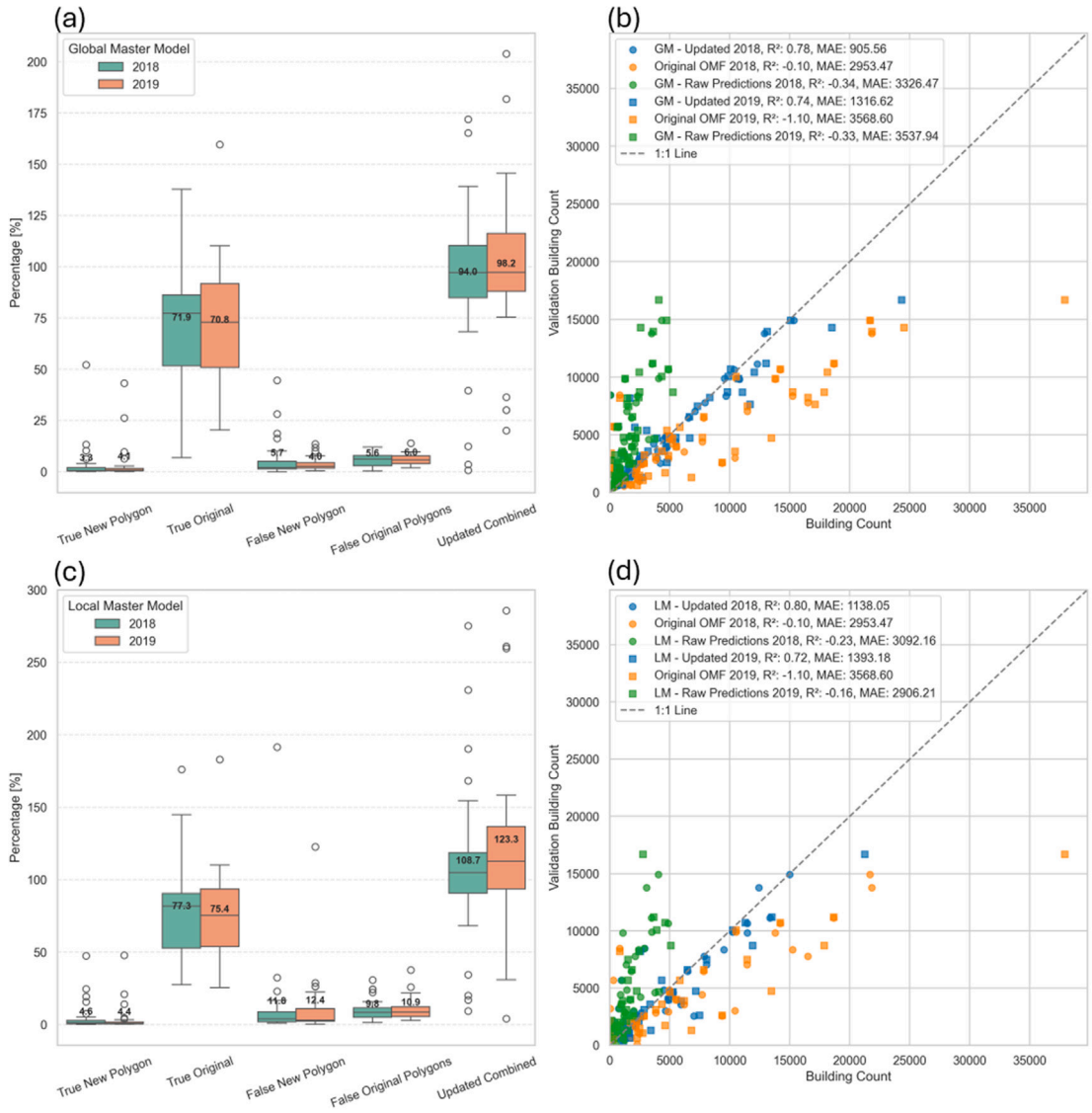

Fig. 8. This figure shows a building count assessment for both updated global (a, b) and local (c, d) master products. The right panels (b, d) show the building count of testing data vs. the building count of the updated file. On the left panels (a, c) the overall percentage per selection case to update the building footprints - per year. The x-label definitions are the following: 'True_New_Percent' and 'False_New_Percent' have the number of testing polygons as the denominator. 'True_Original_Percent', 'False_Original_Percent', 'Total_Prediction_Percent' have the number of updated polygons as the denominator.

The boxplots demonstrate that the removal process of original polygons is highly effective, as reflected by the very low percentage of falsely removed buildings. Only a small number of new polygons need to be added, which suggests that most test locations already contain comprehensive very-high-resolution building footprint data beyond 2019 - for both global and local master models. Between 70% and 77% of OMF footprints remained after applying the temporal disaggregation method. Overall, the updating process ((Selected Original + Added New Polygons)/Number of Testing Polygons) successfully retained 94–98% and 108–123% of all buildings for the global and local master model respectively. In general, the local master model overestimates the total building counts, which is reflected in the inclusion of more new polygons of which around 11% seem to be false detections. The global master model has half of that rate.

The right panels (b, d) of Fig. 8 compare building counts between the updated product, the original OMF footprints, and the raw vectorized predictions from the global and local master models. The rationale for including the raw predictions and original OMF is to evaluate whether the disaggregation process provides an actual advantage over using the predictions or OMF directly. The improvement is substantial: for 2018, the R^2 increases from −0.10 (original OMF) and −0.34 (raw predictions) to 0.78 (updated global master model) and 0.80 (updated local master model), while for 2019 it improves from −1.10 (original OMF), −0.33 (raw predictions global master model) and −0.23 (raw predictions local master model) to 0.74 (updated global master model) and 0.72 (updated local master model). In addition, the mean absolute error (MAE) in building counts is reduced by more than 50% for both years and model types, when comparing against either the original OMF or the raw predictions. These results confirm that the updated products provide a clear advantage in terms of individual building detection, outperforming both input datasets when considered separately.

As a next step, the built-up area of the testing locations is compared across the original OMF, the raw predictions, and the updated product. The left panels (a,c) of Fig. 9 present a direct area comparison

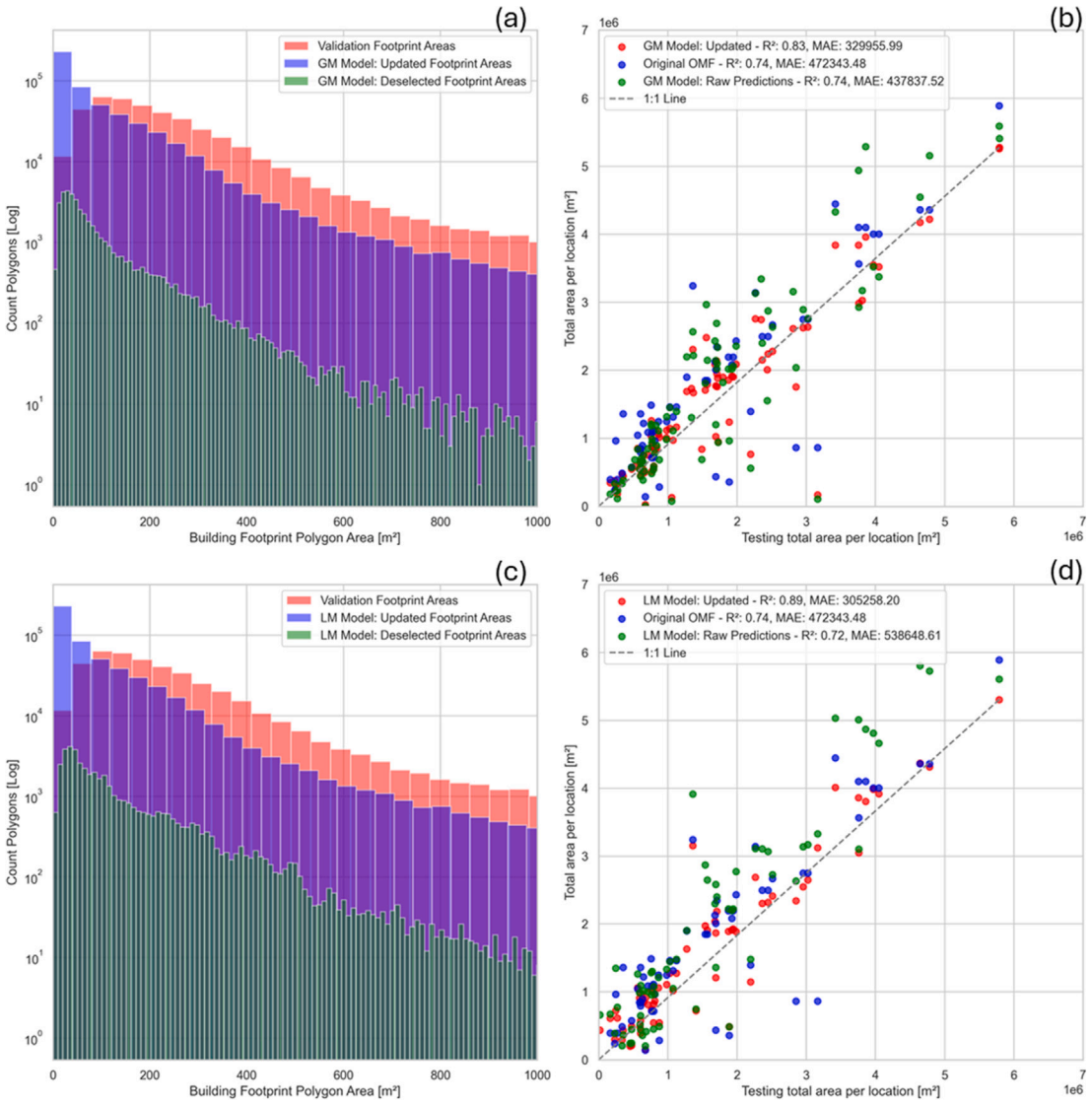

Fig. 9. This figure shows a build up area assessment for both updated global (a,b) and local (c,d) master products. On the left panels (a,c) the area comparison between intersected polygons from the testing and updated building footprints is displayed. The right panels (b,d) illustrate the total built up area per location compared to the testing area for both updated and original building footprints.

using histograms of all testing polygons, updated polygons, and deselected polygons. The red histogram shows the distribution of testing footprint areas, the blue histogram represents the updated footprint areas, and the green histogram indicates the deselected footprint areas. The y-axis is presented in log-scale to highlight the much smaller number of deselected polygons. Both sub-figures (a, c) show a significant discrepancy in average building sizes between the OMF input data and the testing data. Specifically, the OMF footprints exhibit a much higher number of small buildings compared to the testing dataset. This notable difference underscores the challenges in directly comparing the two building footprint sources. One contributing factor may be the difference in image resolution: OMF footprints are likely derived from higher-resolution imagery than the testing data, which is based on PlanetScope imagery with a spatial resolution of approximately 4 meters. Consequently, certain building structures that appear as distinct entities in high-resolution images may appear merged or indistinct in coarser-resolution imagery, potentially resulting in one large building in the testing set corresponding to multiple smaller buildings in the OMF dataset.

Interestingly, the issue of extremely small buildings is less pronounced in the deselected polygons. Unlike the OMF dataset, where the highest count is concentrated in the smallest building size classes, the deselected polygons do not exhibit this same peak (a small shift to the larger areas). This suggests that the U-Net segmentation model is capable of identifying and retaining a significant number of small buildings, even those smaller than the native Sentinel-2 resolution of 10 meters (approximately 100 m^2). This finding highlights the model's sensitivity and segmentation ability at sub-pixel or borderline-pixel scales, which may be attributed to contextual and spectral cues learned during training. The right panels (b,c) of Fig. 9 provide a less biased comparison by directly evaluating the total built-up area, without relying on the exact shape or perfectly georeferenced position of individual building footprints. The scatterplot shows the raw predictions (green), updated polygons (red), and original OMF footprints (blue), all compared against the testing data. The improvement from the original OMF footprints to the updated product is substantial, with the R^2 increasing from 0.74 to 0.83 (global master model), 0.74 to 0.89 (local master model) and the mean absolute error (MAE) reduced by nearly 25% for

Table 6
Comparison of IoU, average building count, and average building area across different datasets. OMF = OMF footprints, GM = global master model predictions, updated = updated footprints. Bold values represent the best performance.

	IoU	Avrg. Building count [%]	Avrg. Building area [%]
OMF	0.30	156.76	127.76
GM	0.35	40.95	104.50
GM updated	**0.36**	**95.81**	**100.76**
LM	0.32	50.59	138.10
LM updated	0.34	114.88	115.81

both global and local master models. A similar improvement is evident when comparing the raw predictions to the updated product: the raw predictions overestimate built-up area, resulting in an R^2 of 0.74 (global master model), and an R^2 of 0.74 (local master model). This clearly demonstrates the effectiveness and practical relevance of the proposed temporal disaggregation approach.

Table 6 summarizes the overall building area and count assessment. On average, the updated global master model captures approximately 96% of the total building area and 101% of the building counts relative to the reference data. In contrast, the local master model captures about 115% of the building area and 116% of the building counts. While the global master model slightly underestimates the total built-up extent, the local master model tends to overestimate both area and counts, reflecting its higher sensitivity to newly added polygons.

At last, Fig. 10 exemplifies an integrated view of spatial and temporal urban growth as predicted by the proposed temporal disaggregation approach. Subplots Fig. 10 (a–d) show four representative locations with half-yearly building footprint updates, using half year Sentinel 2 imagery composites, from 2018 to 2024. To improve interpretability, only newly added buildings are visualized for each half-year period. The color scheme follows a uniform gradient from purple to red, where each hue pair represents one year — with lighter shades corresponding to the first half (January-June) and darker shades to the second half (July-December). This figure highlights how the proposed framework can be extended to sub-annual resolution, allowing individual building footprints to be assigned precise temporal information.

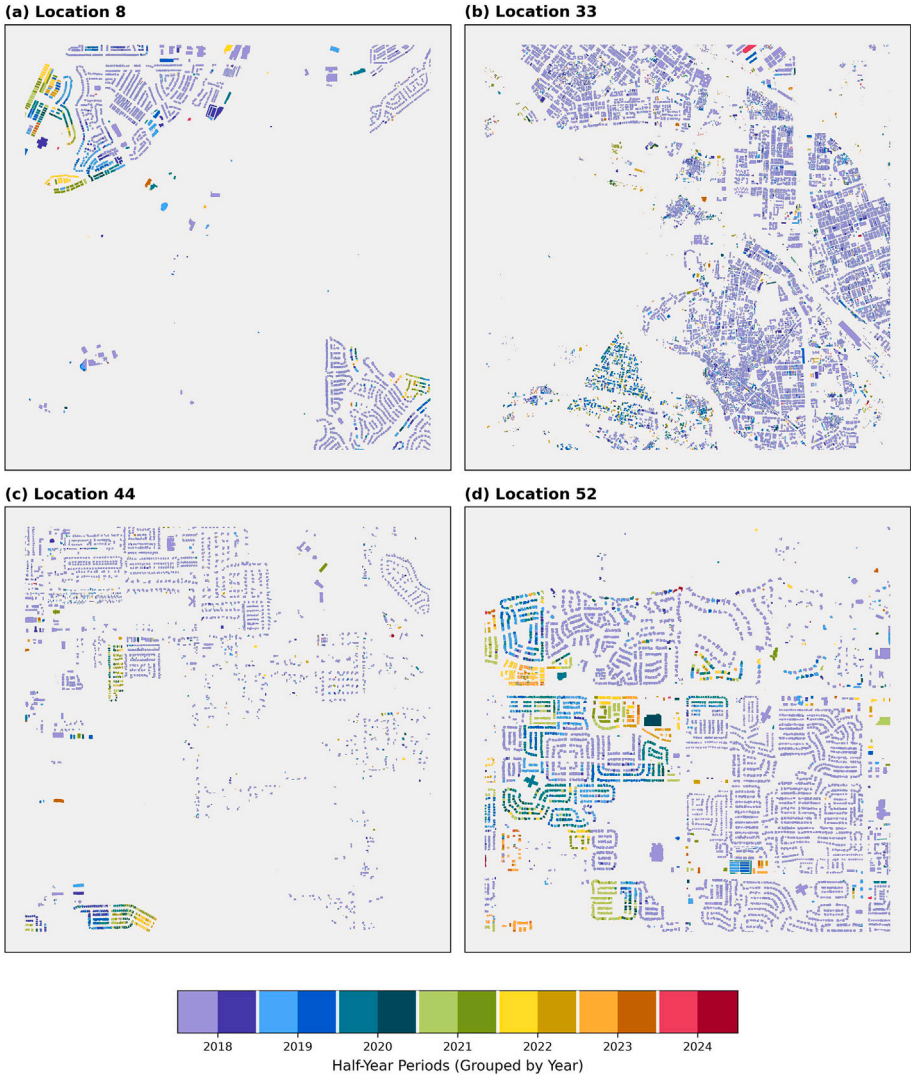

Fig. 10. This figure shows the spatio-temporal evolution of urban building footprints between 2018 and 2024 for four representative locations (a–d). Each map visualizes newly detected buildings for each half-year period, where colors transition from purple (2018) to red (2024), with lighter shades representing the first half (January-June) and darker shades representing the second half (July-December) of each year. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

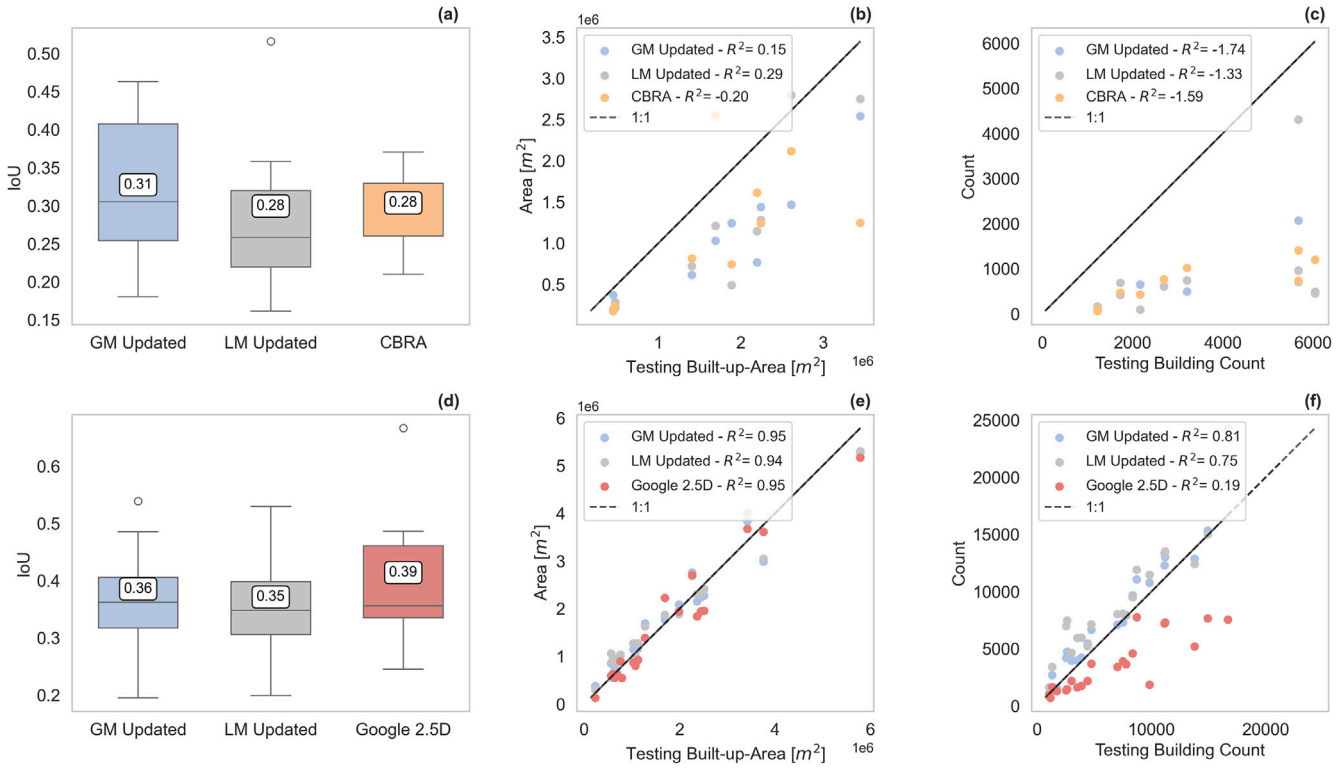

Fig. 11. This figure shows the benchmarking of the updated (global master (GM) and local master (LM)) building footprint products against state-of-the-art temporal datasets (CBRA and Google 2.5D). Metrics include IoU, building count, and total built-up area across all available SpaceNet 7 test locations and timestamps. (a) and (d) are boxplots showing IoU estimates over the relevant locations (including the mean value), (b) and (e) are scatter-plots showing build up area between testing location input datasets and (c) and (f) are scatter-plots showing building count between testing location input datasets. The black dotted line in each scatter plot indicates a 1:1 reference line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. How does the disaggregated product compare to existing state-of-the-art products?

Similarly to the previous evaluation, benchmarking against state-of-the-art temporal building footprint datasets is performed by comparing IoU, building count, and built-up area across all available testing locations and timestamps. The CBRA dataset covers the entire area of China with annual temporal resolution, while Google 2.5D provides annual building footprint layers for the Global South. The CBRA dataset is directly available in binary format, as a softmax threshold of 0.5 had already been applied during processing. In contrast, the Google 2.5D product is provided as softmax probability outputs (ranging from 0–1), which are binarized using an equal threshold of 0.5 to ensure a consistent basis for comparison. This does not represent a perfect evaluation scenario, however, it serves as a practical case study, since a fixed threshold was also applied in the CBRA product. Consequently, the comparison with the Google 2.5D building product is subject to the same inherent bias. As no explicit decision boundary is provided by the data providers, the use of a standardized threshold of 0.5 becomes a necessary and transparent choice to ensure comparability and reproducibility across datasets. These binary rasters, alongside the updated building footprint product, are compared against the testing rasters to calculate IoU. Building counts are derived by vectorizing the binary rasters and summing up the resulting polygons, which are then compared to the testing footprint annotations. Finally, the built-up area is computed by summing the polygon areas per location and comparing them to the testing footprints.

Fig. 11, shows the overall results comparing the mentioned datasets to the testing locations. Subplots (a) and (d) present boxplots of IoU values across locations, while (b) and (e) show scatterplots comparing built-up area with the testing data. Finally, (c) and (f) illustrate scatterplots of building counts relative to the testing dataset. The black dotted line in each scatter plot indicates a 1:1 reference line. The subplots (a–c) in Fig. 11 present the assessment of the CBRA dataset against the testing locations, alongside the updated global and local master footprints from this work. On average the updated global master product outperforms CBRA in terms of mean IoU by approximately 10%, whereas the local master product has very similar performance. Both updated products exhibit a wider IoU range (0.18–0.47 for the global master updated and 0.16–0.36 for the local master updated), whereas CBRA delivers a narrower but more stable range (0.21–0.37). These results are particularly promising given that CBRA relies on a more advanced processing chain, including a super-resolution module for Sentinel-2 imagery. By contrast, this work applies bicubic interpolation in combination with the proposed spatio-temporal adaptation strategy—yet achieves competitive results.

Considering subplot (b), which compares built-up area, both datasets under-perform relative to the global benchmark, with $R^2 = 0.15$ for the global master updated footprints, $R^2 = 0.29$ for the local master updated footprints and $R^2 = -0.20$ for CBRA. A similar pattern emerges for building count (subplot c), where performance levels between the two datasets are comparable but substantially lower than the global estimates. A likely explanation is the poor or missing high-resolution building footprints within OMF across China, which weakens the synergy between OMF and Sentinel-2 segmentations. In such cases, the method relies almost exclusively on Sentinel-2 predictions. Nevertheless, it is encouraging that, even under these conditions, the updated product remains competitive with state-of-the-art datasets. The subplots (d–f) in

Table 7
Approximate training and inference costs for the different model types. Training times are estimated from observed throughput and depend on hardware configuration. Inference speed refers to Bayesian inference with 32 Monte Carlo runs.

Model type	Training tiles	Epochs	Total seen tiles	Training time	Inference speed
Global Base	5,840,000	30	175,200,000	~88 h (~3.5 days)	~115 tiles/s
Local Base	100,000	30	3,000,000	~5 h	~115 tiles/s
Global Master	5000	10	50,000	~15 min	~115 tiles/s
Local Master	5000	10	50,000	~15 min	~115 tiles/s

Fig. 11 compare the updated products with Google 2.5D across test locations in the Global South. For IoU (subplot d), Google 2.5D slightly outperforms the updated product, with a mean IoU of 0.39 compared to 0.36 and 0.35 for the updated global and local master footprints respectively. In terms of built-up area (subplot e), both datasets perform very well, showing strong correlations with testing data (R^2 = 0.95 for the updated global master product, R^2 = 0.94 for the updated local master product and R^2 = 0.95 for Google 2.5D). Subplot (f) illustrates the differences in building count: here, the updated global and local master product performs significantly better (R^2 = 0.81, R^2 = 0.75 respectively) compared to vectorized Google 2.5D (R^2 = 0.19). This discrepancy arises because the updated method incorporates high-resolution vector footprints, which enable finer delineation and separation of individual buildings. In contrast, the coarser resolution of raster-based Google 2.5D tends to merge neighboring buildings, limiting its accuracy for building counts. This again highlights the strength of the synergistic approach, combining OMF vectors with Sentinel-2 segmentations.

5. Practical implications

The results indicate that model performance is primarily driven by two key factors: the quality of the input labels and the amount of available training data. While other influences exist and were discussed throughout the manuscript, these two factors consistently explain a significant part of the observed performance variability across model types. Importantly, the experiments reveal a saturation effect with respect to training data volume. Beyond a certain threshold, increasing the number of training samples yields only marginal performance gains, particularly for the local and master model variants. Based on these findings, recommended training data volumes for each model type are summarized in Table 7.

Overall, the master models, both global and local, are recommended for most practical applications. They consistently achieve higher IoU scores and, more importantly, provide improved robustness to label noise through their explicit modeling of aleatoric and epistemic uncertainty. This increased awareness of ambiguous pixels helps mitigate error propagation originating from incomplete or inaccurate building footprint labels. When sufficient computational resources are available, the global master model is preferred, as it benefits from broader spatial generalization while maintaining robustness to heterogeneous label quality. In cases where computational constraints or region-specific adaptations are required, the local master model offers a strong alternative with significantly reduced training time.

Assessing label quality remains a practical challenge, as no global completeness map exists for the OMF-based building footprints used in this study. However, existing global assessments of OpenStreetMap building completeness can serve as a useful proxy for estimating expected performance and uncertainty levels (Zhou et al., 2022). Such reference datasets provide actionable guidance for users when selecting an appropriate model configuration based on regional data quality. To further increase the robustness of the outputs, multiple Sentinel-2 acquisitions covering similar time periods can be incorporated when available. By aggregating predictions across several images and computing the mean response, an additional consistency check is introduced that reduces the influence of scene-specific noise. This temporal aggregation can act as an extra robustness layer, particularly in environments characterized by short-lived or highly dynamic built-up structures, such as temporary settlements, construction sites, or tent-based housing. This also implies that areas with less Sentinel 2 coverage, due to high cloud-cover, have less data for such a robustness check as well as a more limited temporal resolution.

6. Limitations and future work

Considering the proposed methodology the most apparent limitation is that the suggested disaggregation approach depends on the quality and completeness of the high-resolution building footprints. In regions with poor OMF coverage (e.g., China), the synergy effect is weakened, leaving performance largely reliant on Sentinel-2 predictions. Still, this work also highlights that even though those cases exist, Sentinel-2 predictions can still deliver a robust product, which is comparable to or better than other benchmark products. Moreover, it is important to note that the raw OMF building footprints are used without additional quality control, which may have negatively impacted model performance, particularly during fine-tuning. To address this, a pre-filtering step using auxiliary datasets such as GHSL or WSF to assess and further improve the quality of training labels is recommended for future work.

Another challenge remains regarding aleatoric uncertainty, particularly in dense urban areas where spectral confusion is high; future work could explore heteroscedastic loss attenuation to mitigate residual errors (Kendall and Gal, 2017). Additional research could investigate how uncertainty (especially aleatoric) can be operationalized as a user-facing proxy to flag potentially unreliable predictions, supporting risk and exposure assessments.

While this study provides a strong proof of concept, the framework can be extended in future work by exploring more advanced architectures such as Vision Transformers (ViTs) (Liu et al., 2023a; Thisanke et al., 2023), Mamba-based state space models (Dao and Gu, 2024; Bao et al., 2025; Zhu et al., 2024a), and other emerging state-of-the-art approaches. However, their integration into the proposed workflow requires the availability of robust probabilistic outputs, which are essential for the uncertainty-driven disaggregation strategy developed in this study.

Finally, further investigation is required to assess how the proposed temporal disaggregation framework handles edge cases, such as temporary structures (e.g., tents, construction scaffolding) or building demolitions. The current implementation assigns mono-temporal labels based on Sentinel-2 observations combined with high-resolution footprint geometries. However, short-lived built-up signals or temporary surface changes may lead to false temporal assignments if only single acquisitions are considered. To mitigate this effect, future work should incorporate temporal persistence checks, for example by requiring that built-up signals remain stable across multiple Sentinel-2 acquisitions before a structure is classified as permanent. Such multi-temporal consistency filtering would reduce the risk of labeling temporary objects as buildings.

7. Conclusion

This study introduces a novel method for the temporal disaggregation of building footprints, addressing a major limitation of existing open-source datasets such as Google, Meta, OSM, and OMF, which — despite their high spatial resolution — lack temporal metadata. Temporal information is essential for applications such as risk assessment, population estimation, and urbanization monitoring, and the proposed approach demonstrates how static footprints can be enriched with temporal labels using Earth Observation data.

The evaluation of multiple training strategies demonstrates that the spatio-temporal fine-tuned global model achieves the most robust performance across all test locations, including unseen test sites. Incorporating uncertainty into the assessment allows the evaluation of model strategies not only in terms of performance but also in terms of transparency and reliability. In this study, the fine-tuned global model delivers the best results, balancing uncertainty, average cost, and IoU. The results further demonstrate clear improvements over using only Sentinel-2 segmentations or raw high-resolution building footprints. The best performing product (global master model) successfully captures on average 96% of all individual buildings and 101% of the total building area, while Sentinel-2 segmentations greatly underestimate building count and OMF data overestimates both building count and built up area. Moreover, integrating uncertainty estimates alongside softmax outputs in the disaggregation method allows users to make more informed decisions by accounting for both trustworthiness and performance. Finally, benchmarking the updated product against CBRA (China) and Google 2.5D (Global South) confirms its competitiveness, as it matches or outperforms existing datasets.

The findings highlight the value of combining high-resolution static footprints with Sentinel-2–based segmentation, enabling the correction of high resolution static footprints and the detection of newly built structures. By integrating IoU, building counts, built-up area, and uncertainty-based metrics, the study establishes a comprehensive validation framework that extends beyond purely geometric accuracy and provides practical guidance for applying the proposed methodology.

CRediT authorship contribution statement

Manuel Huber: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christian Geiß:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Jessy Ribaira:** Writing – review & editing, Software. **Michael Schmitt:** Writing – review & editing. **Hannes Taubenböck:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This project has received funding from the European Union's Horizon 2021 Research and Innovation Programme under Grant Agreement N° 101073954. The authors acknowledge the support of the Deutsche Zentrum für Luft- und Raumfahrt (DLR) and partner institutions for providing research infrastructure and valuable collaboration.

Appendix A

Table A.8
Distribution of Tiles by Urban Density (UD) and Location. Row colors indicate the total number of tiles: for < 200, for 200–600, for 600–1500, and for > 1500.

	Tiles Low UD	Tiles Medium UD	Tiles High UD	Total Number of Tiles
Location 0	10	5	3	18
Location 1	35	38	12	85
Location 2	424	885	26	1335
Location 3	36	55	27	131
Location 4	915	1000	210	2135
Location 5	156	277	230	663
Location 6	24	19	5	48
Location 7	201	268	85	554
Location 8	95	192	88	375
Location 9	62	109	107	279
Location 10	878	1000	428	2306
Location 11	100	124	89	347
Location 12	295	831	367	1494
Location 13	559	1000	463	2022
Location 14	87	127	39	259
Location 15	380	459	139	978
Location 16	8	9	4	21
Location 17	863	631	150	1650
Location 18	312	609	382	1303
Location 19	341	228	50	619
Location 20	666	917	335	1919
Location 21	812	628	152	1598
Location 22	553	611	164	1329
Location 23	334	554	28	916
Location 24	83	143	65	296
Location 25	132	186	64	394
Location 26	623	1000	495	2119
Location 27	730	1000	388	2120
Location 28	62	73	18	153
Location 29	997	1000	703	2703
Location 30	57	53	18	128
Location 31	363	595	274	1310
Location 32	632	714	308	1720
Location 33	213	262	110	588
Location 34	185	240	33	458
Location 35	128	131	65	334
Location 36	139	231	76	455
Location 37	1000	1000	397	2408
Location 38	14	30	22	70
Location 39	187	244	113	547
Location 40	19	48	14	81
Location 41	42	56	17	115
Location 42	123	167	22	312
Location 43	632	1000	478	2111
Location 44	221	324	63	608
Location 45	1000	1000	507	2508
Location 46	173	133	30	337
Location 47	807	613	145	1571
Location 48	352	417	130	899
Location 49	218	717	671	1629
Location 50	538	586	148	1273
Location 51	347	440	130	917
Location 52	42	70	28	140
Location 53	179	588	593	1382
Location 54	98	91	58	257
Location 55	434	513	119	1068
Location 56	246	279	81	620
Location 57	286	258	112	691
Location 58	53	77	22	152
Location 59	251	235	69	555

Data availability

Data will be made available on request.

References

Ahmad, M.N., Almutlaq, F., Huq, M.E., Islam, F., Javed, A., Skilodimou, H.D., Bathrellos, G.D., 2025. Enhanced urban impervious surface land use mapping using a novel multi-sensor feature fusion method and remote sensing data. Environ. Earth Sci. 84, 228.

Ahn, Y., Leyk, S., Uhl, J.H., McShane, C.M., 2024. An integrated multi-source dataset for measuring settlement evolution in the united states from 1810 to 2020. Sci. Data 11, 275. https://doi.org/10.1038/s41597-024-03081-x

Alshehhi, R., Marpu, P.R., Woon, W.L., Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 130, 139–149. https://doi.org/10.1016/j.isprsjprs.2017.05.002

Bao, M., Lyu, S., Xu, Z., Zhou, H., Ren, J., Xiang, S., Li, X., Cheng, G., 2025. Vision mamba in remote sensing: A comprehensive survey of techniques, applications and outlook. arXiv preprint arXiv:2505.00630.

Baumann, A., Roßberg, T., Schmitt, M., 2023. Probabilistic MIMO U-Net: efficient and accurate uncertainty estimation for pixel-wise regression. In: IEEE-CVF International Conference on Computer Vision, pp. 4498–4506, https://doi.org/10.1109/ICCVW60793.2023.00484

Chen, H., Cheng, L., Zhuang, Q., Zhang, K., Li, N., Liu, L., Duan, Z., 2022. Structure-aware weakly supervised network for building extraction from remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2022.3217830

Dao, T., Gu, A., 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060.

Dechesne, C., Lassalle, P., Lefèvre, S., 2021. Bayesian U-Net: estimating uncertainty in semantic segmentation of earth observation images. Remote Sens. 13, 3836. https://doi.org/10.3390/rs13193836

European Commission, Joint Research Centre, 2023. Global human settlement layer (ghsl): built-up surface. https://developers.google.com/earth-engine/datasets/catalog/JRC_GHSL_P2023A_GHS_BUILT_C (Accessed: 4 June 2025).

European Space Agency (ESA) and Google Earth Engine, 2023. Copernicus/s2_sr_harmonized: sentinel-2 surface reflectance. https://developers.google.com/earth-engine/datasets/catalog/copernicuss2srharmonized (Accessed: 4 June 2025).

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059, https://doi.org/10.48550/arXiv.1506.02142

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56, 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E., 2018. Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. IEEE Geosci. Remote Sens. Lett. 15, 607–611.

Gevaert, C.M., Buunk, T., Van Den Homberg, M.J.C., 2024. Auditing geospatial datasets for biases: using global building datasets for disaster risk management. IEEE J. Sel. Top. Appl. Earth Obs. https://doi.org/10.1109/JSTARS.2024.3422503

Gibril, M.B.A., Al-Ruzouq, R., Shanableh, A., Jena, R., Bolcek, J., Shafri, H.Z.M., Ghorbanzadeh, O., 2024. Transformer-based semantic segmentation for large-scale building footprint extraction from very-high resolution satellite images. Adv. Space Res. 73, 4937–4954.

Gong Peng, G.P., Wang Jie, W.J., Yu Le le, Y., Zhao YongChao, Z.Y., Zhao YuanYuan, Z.Y., Liang Lu, L.L., Niu ZhenGuo, N.Z., Huang XiaoMeng, H.X., Fu HaoHuan, F.H., Liu Shuang, L.S., et al., 2013. Finer resolution observation and monitoring of global land cover: first mapping results with landsat tm and etm+ data. Int. J. Remote Sens. https://doi.org/10.1080/01431161.2012.748992

Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021. Deep building footprint update network: a semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sens. Environ. 264, 112589. https://doi.org/10.1016/j.rse.2021.112589

Hafner, S., Ban, Y., Nascetti, A., 2022. Unsupervised domain adaptation for global urban extraction using sentinel-1 SAR and sentinel-2 MSI data. Remote Sens. Environ. 280, 113192. https://doi.org/10.1016/j.rse.2022.113192

Haghighi Gashti, E., Delavar, M.R., Guan, H., Li, J., 2024. Semantic segmentation uncertainty assessment of different U-Net architectures for extracting building footprints. ISPRS Ann. Photogramm. Remote Sens. 10, 141–148. https://doi.org/10.5194/isprs-annals-X-4-2024-141-2024

He, W., Jiang, Z., 2023. Uncertainty quantification of deep learning for spatiotemporal data: Challenges and opportunities. arXiv preprint arXiv:2311.02485.

Hertel, V., Wani, O., Geiß, C., Wieland, M., Taubenböck, H., 2025. Bayesiammtl: uncertainty-aware multitask learning for post-disaster building damage assessment. Int. J. Appl. Earth Obs. Geoinf. 143, 104759. https://doi.org/10.1016/j.jag.2025.104759

Hu, Q., Zhen, L., Mao, Y., Zhou, X., Zhou, G., 2021. Automated building extraction using satellite remote sensing imagery. Autom. Constr. 123, 103509. https://doi.org/10.1016/j.autcon.2020.103509

Huang, X., Wang, C., Li, Z., 2019. High-resolution population grid in the conus using Microsoft building footprints: a feasibility study. In: ACM SIGSPATIAL International Workshop on Geospatial Humanities, pp. 1–9, https://doi.org/10.1145/3356991.3365469

Huang, X., Yang, J., Wang, W., Liu, Z., 2022. Mapping 10-m global impervious surface area (gisa-10m) using multi-source geospatial data. Earth Syst. Sci. Data Discuss. 2022, 1–39. https://doi.org/10.5194/essd-14-3649-2022

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110, 457–506. https://doi.org/10.1007/s10994-021-05946-3

Idrees, M.O., Omar, D.M., Babalola, A., Ahmadu, H.A., Yusuf, A., Lawal, F.O., 2022. Urban land use land cover mapping in tropical Savannah using landsat-8 derived normalized difference vegetation index (ndvi) threshold. S. Afr. J. Geomat. 11.

Ji, C., Tang, H., 2025. Towards reliable land cover mapping under domain shift: an overview and comprehensive comparative study on uncertainty estimation. Earth-sci. Rev. 105070. https://doi.org/10.1016/j.earscirev.2025.105070

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30, https://doi.org/10.5555/3295222.3295309

LaBonte, T., Martinez, C., Roberts, S.A., 2019. We know where we don't know: 3d bayesian cnns for credible geometric uncertainty. arXiv preprint arXiv:1910.10793.

Liu, C., Albrecht, C.M., Wang, Y., Zhu, X.X., 2024. Task specific pretraining with noisy labels for remote sensing image segmentation. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 7040–7044.

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., Zhang, Y., 2019. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. Remote Sens. 11, 830. https://doi.org/10.3390/rs11070830

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z., 2023a. A survey of visual transformers. IEEE Trans. Neural Netw. Learn. Syst. 35, 7478–7498. https://doi.org/10.1109/TNNLS.2022.3227717

Liu, Z., Tang, H., 2023. Learning sparse geometric features for building segmentation from low-resolution remote-sensing images. Remote Sens. 15, 1741. https://doi.org/10.3390/rs15071741

Liu, Z., Tang, H., Feng, L., Lyu, S., 2023b. China building rooftop area: the first multi-annual (2016–2021) and high-resolution (2.5 m) building rooftop area dataset in China derived with super-resolution segmentation from sentinel-2 imagery. Earth Syst. Sci. Data 15, 3547–3572. https://doi.org/10.5194/essd-15-3547-2023

Lu, T., Ming, D., Lin, X., Hong, Z., Bai, X., Fang, J., 2018. Detecting building edges from high spatial resolution remote sensing imagery using richer convolution features network. Remote Sens. 10, 1496. https://doi.org/10.3390/rs10091496

Maps, O., 2025. Overture maps foundation. https://overturemaps.org/ (accessed: 8 April 2025).

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., et al., 2020. Outlining where humans live, the world settlement footprint 2015. Sci. Data 7, 242. https://doi.org/10.1038/s41597-020-00580-5

Mas, J.-F., Filho, B.S., Pontius Jr, R.G., Gutiérrez, M.F., Rodrigues, H., 2013. A suite of tools for ROC analysis of spatial models. ISPRS Int. J. Geo-inf. 2, 869–887. https://doi.org/10.3390/ijgi2030869

Meka, M. and Google Research, 2024. Open buildings temporal dataset, earth engine app. https://mmeka-ee.projects.earthengine.app/view/open-buildings-temporal-dataset (Accessed: 4 June 2025).

Microsoft, 2025. Microsoft building footprints. https://github.com/microsoft/GlobalMLBuildingFootprints (accessed: 8 April 2025).

Mirpulatov, I., Illarionova, S., Shadrin, D., Burnaev, E., 2023. Pseudo-labeling approach for land cover classification through remote sensing observations with noisy labels. IEEE Access 11, 82570–82583. https://doi.org/10.1109/ACCESS.2023.3300967

Neupane, B., Aryal, J., Rajabifard, A., Pelizari, P.A., Geiß, C., 2025. Multi-label learning with VIT for building footprint extraction from off-nadir aerial images. IEEE Geosci. Remote Sens. Lett. https://doi.org/10.1109/LGRS.2025.3532589

Palacios-Lopez, D., Bachofer, F., Esch, T., Heldens, W., Hirner, A., Marconcini, M., Sorichetta, A., Zeidler, J., Kuenzer, C., Dech, S., et al., 2019. New perspectives for mapping global population distribution using world settlement footprint products. Sustainability 11, 6056. https://doi.org/10.3390/su11216056

Prexl, J., Schmitt, M., 2023. The potential of sentinel-2 data for global building footprint mapping with high temporal resolution. In: 2023 Joint Urban Remote Sensing Event (JURSE). IEEE, Heraklion, Greece, pp. 1–4, https://doi.org/10.1109/JURSE57346.2023.10144166

Razzak, M.T., Mateo-García, G., Lecuyer, G., Gómez-Chova, L., Gal, Y., Kalaitzis, F., 2023. Multi-spectral multi-image super-resolution of sentinel-2 with radiometric consistency losses and its effect on building delineation. ISPRS J. Photogramm. Remote Sens. 195, 1–13. https://doi.org/10.1016/j.isprsjprs.2022.10.019

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. version Number: 1. arxiv:1505.04597.

Sale, Y., Hofman, P., Löhr, T., Wimmer, L., Nagler, T., Hüllermeier, E., 2024. Label-wise aleatoric and epistemic uncertainty quantification. arXiv preprint arXiv:2406.02354.

Sapena, M., Kühnl, M., Wurm, M., Patino, J.E., Duque, J.C., Taubenböck, H., 2022. Empiric recommendations for population disaggregation under different data scenarios. PLOS ONE 17, e0274504. https://doi.org/10.1371/journal.pone.0274504

Sarma, A., Kay, M., 2020. Prior setting in practice: strategies and rationales used in choosing prior distributions for Bayesian analysis. In: 2020 Chi Conference on Human Factors in Computing Systems, pp. 1–12, https://doi.org/10.1145/3313831.3376377

Shi, Y., Li, Q., Zhu, X., 2019. Building footprint extraction with graph convolutional network. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5136–5139, https://doi.org/10.1109/IGARSS.2019.8898764

Shridhar, K., Laumann, F., Liwicki, M., 2019. A comprehensive guide to bayesian convolutional neural network with variational inference. arXiv preprint arXiv:1901.02731.

Sirko, W., Brempong, E.A., Marcos, J.T.C., Annkah, A., Korme, A., Hassen, M.A., Sapkota, K., Shekel, T., Diack, A., Nevo, S., Hickey, J., Quinn, J., 2023. High-Resolution Building and Road Detection from Sentinel-2. version Number: 3. arxiv:2310.11622.

SpaceNet, 2025. Spacenet7 challenge. https://github.com/SpaceNetChallenge/ (accessed: 8 April 2025).

Taubenböck, H., 2025. Urban Science with a View from Space. Technical Report. German Aerospace Center (DLR). https://elib.dlr.de/217314/.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., Herath, D., 2023. Semantic segmentation using vision transformers: a survey. Eng. Appl. Artif. Intell. 126, 106669. https://doi.org/10.1016/j.engappai.2023.106669

Uhl, J.H., Leyk, S., 2022. Mtbf-33: a multi-temporal building footprint dataset for 33 counties in the united states (1900–2015). Data Br. 43, 108369. https://doi.org/10.1016/j.dib.2022.108369

Usmani, M., Bovolo, F., Napolitano, M., 2023. Remote sensing and deep learning to understand noisy openstreetmap. Remote Sens. 15, 4639.

Van Etten, A., Hogan, D., Martinez Manso, J., Shermeyer, J., Weir, N., Lewis, R., 2021. The multi-temporal urban development spacenet dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6398–6407, https://doi.org/10.1109/CVPR46437.2021.00633

Yang, G., Destercke, S., Masson, M.-H., 2016. The costs of indeterminacy: how to determine them? IEEE Trans. Cybern. 47, 4316–4327. https://doi.org/10.1109/TCYB.2016.2607237

Yu, H., Hu, H., Xu, B., Shang, Q., Wang, Z., Zhu, Q., 2023. Superpixelgraph: semi-automatic generation of building footprint through semantic-sensitive superpixel and neural graph networks. Int. J. Appl. Earth Obs. Geoinf. 125, 103556. https://doi.org/10.1016/j.jag.2023.103556

Zhou, D., Xiao, J., Bonafoni, S., Berger, C., Deilami, K., Zhou, Y., Frolking, S., Yao, R., Qiao, Z., Sobrino, J.A., 2018. Satellite remote sensing of surface urban heat islands: progress, challenges, and perspectives. Remote Sens. 11, 48. https://doi.org/10.3390/rs11010048

Zhou, Q., Zhang, Y., Chang, K., Brovelli, M.A., 2022. Assessing osm building completeness for almost 13, 000 cities globally. Int. J. Digit. Earth 15, 2400–2421.

Zhou, Y., Tan, Y., Wen, Q., Wang, W., Li, L., Li, Z., 2023. Deep multimodal fusion model for building structural type recognition using multisource remote sensing images and building-related knowledge. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 9646–9660.

Zhu, Q., Cai, Y., Fang, Y., Yang, Y., Chen, C., Fan, L., Nguyen, A., 2024a. Samba: semantic segmentation of remotely sensed images with state space model. Heliyon 10, https://doi.org/10.1016/j.heliyon.2024.e38495

Zhu, X.X., Li, Q., Shi, Y., Wang, Y., Stewart, A., Prexl, J., 2024b. Global OpenBuildingMap – Unveiling the Mystery of Global Buildings arxiv:2404.13911.

Zhu, Z., Zhou, Y., Seto, K.C., Stokes, E.C., Deng, C., Pickett, S.T.A., Taubenböck, H., 2019. Understanding an urbanizing planet: strategic directions for remote sensing. Remote Sens. Environ. 228, 164–182. https://doi.org/10.1016/j.rse.2019.04.020