A Comparison of Uncertainty Estimation Methods for Building Footprint Change Detection from Sentinel-2 Imagery

Jonathan Prexl*, Anton Baumann*, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich, Germany - {firstname.surname}@unibw.de
*These authors contributed equally to this work

Keywords: Deep Learning, Earth Observation, Semantic Segmentation, Building Footprints

Abstract

This manuscript investigates the effects of uncertainty methods applied to the problem of deep learning-based semantic segmentation of building footprints on moderate-resolution satellite imagery. While the recent efforts of big corporations to add information about building locations and sizes on a global or continental scale are generally valuable, still the overall challenge persists in identifying the spatial-temporal patterns of growing urbanization. In this work, we extend UNet-type architectures to perform binary building footprint classification based on *Sentinel-2* imagery resulting in five different models. While previous studies focused on urban areas in the Western world we conduct all training and evaluation in India. All models are trained on Microsoft building footprint products while for evaluation purposes high-quality reference data is manually selected from regions with especially good open-street-map coverage. Quantitative and qualitative experiments are conducted where a significant performance gain is found for a model trained with a mixture of aleatoric and epistemic uncertainty measures. The performance gain is even more pronounced for subsequent quantitative multi-temporal change detection experiments.

1. Introduction

The quantity and variety of freely accessible, medium-resolution Earth observation satellite data is expanding quickly due to the open data policy of many governmental space organisations. The primary reason for these organizations to provide open access to this data is to enable researchers worldwide to track environmental changes, such as variations in biomass (Sibanda et al., 2015) and forest cover (Nguyen et al., 2022), as well as to study the effects of increasing urbanization on our planet (Corbane et al., 2021). This study focuses in detail on monitoring urban areas, and in particular, the aspect of building footprints, even though most concepts are generally applicable to any semantic segmentation task on optical Earth observation data.

Building footprints are a key parameter for describing urban areas and serve as foundational data for various analyses, including population estimation (Huang et al., 2019) and research on land use and land efficiency. In recent years, many efforts have been made to close the information gap of building footprints on a global scale. Next to community-driven projects like open-street-map (OSM), cooperations like *Microsoft* and *Google* contributed with freely accessible building footprint maps on large scales (Microsoft Building Fooptprints, n.d., Sirko et al., 2021). Those maps are based on the underlying high-resolution imagery of their mapping products, such as *Bing-Maps* or *Google-Maps*. While these datasets represent significant advancements and are a valuable contribution to our understanding of global building footprints, there are still important questions that need to be addressed, including those related to temporal dynamics and the overall quality measurement of the data. We will touch on some of the problems later in this manuscript. Especially due to the unknown timestamp of the underlying imagery, it is still desirable to derive information about urbanisation, its temporal patterns and especially building footprint information from satellite imagery that is captured on a regular, scheduled basis. Here, in contrast to the above-mentioned products, the exact temporal information is precisely known, which allows for a detailed analysis of the occurring changes that will be the subject of this study.

One revolutionizing satellite with an open access policy is the *Sentinel-2* mission, specifically crafted to monitor a wide range of change dynamics on the Earth’s surface. Especially in the urban context, it has been shown that *Sentinel-2* exceeds the capabilities of the heritage Landsat missions with respect to the mapping of urban structures due to a sensor resolution closer to the typical urban morphology length scale. There are various approaches to mapping urbanization and its temporal change patterns. Traditionally, many studies treat the entire urban area as a single class in an N-class classification setup, which is a common approach in land cover mapping (Schmitt et al., 2020). However, more recent methods aim for a more granular classification scheme to unlock additional possibilities. In a recent study, it has been shown that mapping capabilities of *Sentinel-2* imagery for building footprints reach up to building instance level for Western-type city morphologies. (Prexl and Schmitt, 2023, Prexl et al., 2023).

The task of mapping building structures from moderate resolution Earth observation data such as *Sentinel-2* is of varying degrees of difficulty due to various major factors. The first and foremost characteristic is the size and arrangement of the buildings. In Western cities, often houses are relatively large and arranged in an ordered pattern along streets. Those structured patterns make it an easier task for a deep learning-based model. In contrast to that, in economically poorer regions of the world, buildings are statistically smaller and within urban environments often densely packed. This makes the analysis of building-related information from a (relatively) low-resolution information source such as *Sentinel-2* a challenging task.

Therefore, in this manuscript, we employ and compare methods for predicting uncertainty in deep neural networks specifically for the task of building footprint segmentation and sub-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-339-2024

sequent change detection. These methods offer the potential to improve the calibration of model output probabilities in relation to ground truth data. Well-calibrated models serve as a foundation for more accurate post-classification change detection. The focus lies on generating predictions with calibrated uncertainties that better reflect actual data uncertainties. By integrating uncertainty estimates into the change detection process, these models establish a more robust framework for identifying changes in building footprints. This study explores diverse uncertainty estimation approaches and showcases their effective integration into change detection algorithms, ultimately enhancing their precision and dependability.

The manuscript is structured as follows. Section Uncertainty Estimation in Deep Learning provides an overview of the uncertainty methods that are used within this study, as well as an introduction to the basic terminology and concepts. The section Experimental Setup gives an overview of the technical deep learning aspect of the work, as well as evaluation procedures. In section Results we will quantitatively evaluate the different uncertainty methods and give a qualitative overview of the resulting change detection product. We will discuss the findings in section Discussion before we conclude.

2. Uncertainty Estimation in Deep Learning

We investigate the impact of applying uncertainty estimation methods to the task of semantic segmentation of building footprints on *Sentinel-2* imagery via deep learning. Models that are tailored to predict a correctly calibrated model uncertainty have intrinsically better-calibrated class probability vectors, and hence are expected to provide more accurate change detection results during temporal inference and comparison steps. In order to investigate the effect, we build on (Prexl and Schmitt, 2023, Prexl et al., 2023, Ayala et al., 2022) and expand the models with various methods for uncertainty prediction. Whereas the previously mentioned works choose study areas in the USA and Europe, where building footprints are easier to segment, we will train and evaluate the models in India. Here, more challenging patterns are common due to dense urbanisation or small building structures, and hence the generation of reliable building footprint change maps needs well-calibrated class probabilities. In the following, we provide the reader with a short recap of the commonly used approaches for uncertainty estimation, before diving into the technical details for the model training and evaluation.

Two main types of uncertainties often considered are *aleatoric uncertainty* and *epistemic uncertainty*. The former arises from inherent randomness in the data, such as variations in lighting, occlusions, and diverse building materials, which introduce noise into building footprint segmentation. This type of uncertainty is irreducible regardless of data quantity. The latter stems from incomplete knowledge or insufficient training data. For example, a model trained mostly in urban settings might struggle with rural building styles, revealing a knowledge gap that could be reduced by expanding the training dataset to include a broader range of environments.

Aleatoric uncertainty is commonly addressed with Maximum a Posteriori (MAP) estimation (Nix and Weigend, 1994), i.e., in the regression setting the mean squared error loss function is adapted to not only predict the mean of the posterior distribution but additionally the variance parameter. For classification, on the other hand, the model output is often transformed into a probability vector through the softmax function. While this output already fully defines a categorical distribution, it should not be used to infer uncertainty information, as in practice, deep neural networks tend to be overconfident (Hendrycks D., 2017). To this end, Kendal and Gal (Kendall and Gal, 2017) proposed a custom softmax layer that models logits as Gaussian random variables with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ and utilize Monte Carlo sampling to estimate $p(y|\mathbf{x})$:

$$\hat{\mathbf{z}}_j \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$$

$$p(y|\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} \mathit{softmax}(\hat{\mathbf{z}}_\mathbf{j})$$

This method is referred to as *heteroscedastic classification*.

Instead of placing a distribution over the output of a model, epistemic uncertainty is modelled by placing a prior distribution over a model's weights. However, obtaining the full posterior distribution over the parameters is computationally intractable (Gawlikowski et al., 2021). To approximate this distribution, several techniques have been introduced:

Deep Ensembles (Lakshminarayanan et al., 2017) employ several neural networks with diverse initial weights to estimate epistemic uncertainty, with each model in the ensemble producing unique predictions. This approach effectively simulates the predictive posterior distribution, capable of capturing multiple modes of the distribution (Fort et al., 2019). However, it necessitates greater computational resources due to the multiple models and incurs delayed inference from the need to collate outputs across the ensemble. This requires repeating the training process and the inference step multiple times, hence is the most computationally expensive approach. Therefore, we won't utilise this method in our study. Rather, we apply a comparable method (compare next section) by utilising Monte Carlo Dropout, which is still expensive during inference (multiple forward passes) but only one training run for the model is required.

Monte Carlo Dropout (MC Dropout) (Gal and Ghahramani, 2016) approximate the posterior distribution of weights in Bayesian Neural Networks by repurposing dropout regularization (Srivastava et al., 2014). This technique activates dropout during inference to generate predictions from multiple "submodels," each representing a different dropout configuration from the approximate posterior. While this requires multiple passes during inference, slowing down prediction times, it benefits from the efficiency of training only a single model.

The **MIMO framework** (Havasi et al., 2020) and extensions like MIMO U-Net (Baumann et al., 2023) leverage the inherent overparameterization of deep neural networks (Molchanov et al., 2016), which allows for the removal of numerous connections without significant loss in performance, suggesting the feasibility of multiple independent subnetworks within a single network. MIMO trains these "winning tickets" simultaneously, facilitating the concurrent evaluation of all subnetworks in a single forward pass during testing, thus combining the ensemble methods' ability to explore multi-mode distributions with the efficiency of a single-pass evaluation (Havasi et al., 2020).

In the following, we will implement Monte Carlo Dropout and the MIMO framework to estimate epistemic uncertainty, as well

as heteroscedastic classification techniques and their combinations for evaluating aleatoric uncertainty in the context of building footprint segmentation. We will then qualitatively compare the results in terms of prediction quality and the calibration of the predicted probabilities, using a standard U-Net as the baseline for our analysis.

3. Experimental Setup

Used Data Similar to (Prexl and Schmitt, 2023, Prexl et al., 2023) we train UNet-based (Ronneberger et al., 2015) architectures on the task of building footprint segmentation, whereas the major contribution of this study is the extension of uncertainty estimation methods specifically in the context of compact, irregularly shaped buildings. We use 21 full-size *Sentinel-2* scenes over 6 locations in India, where different seasons are covered, to get a robust model with respect to seasonal change. All scenes are centered around urban areas, which ensures dense and sparse urban patterns are present in the dataset. All testing is done on a further scene over New Delhi and the surrounding suburbs. Figure 2 displays the locations of the training and testing areas.

Since *OpenStreetMap* (OSM) building footprint data is only sparsely available in India, we train on the Microsoft building footprints dataset (Microsoft Building Fooptprints, n.d.). In comparison to the underlying studies (Prexl and Schmitt, 2023, Prexl et al., 2023) we have to account for lower label quality representing a hurdle for learning a robust model as mentioned in Section 1.

During training, we extract 20k random samples of the size $10 \times 128 \times 128$ (all ten and twenty-meter bands of *Sentinel-2*) from the 21 potential *Sentinel-2* scenes. We ensure a balanced training set by choosing 10k samples with more than 4% building pixels (dense urbanization) in the corresponding label and 10k with less than 4% building pixels (sparse or no urbanization).

Used Models In our study, we evaluate five models designed to quantify aleatoric and epistemic uncertainties, consistent with the methods discussed in Section 2.

As a baseline, we use **Deterministic UNet (Det-UNet)**, a conventional implementation of UNet (Ronneberger et al., 2015) with no specific features for uncertainty estimation. Following this, we use the **HC-UNet**, a variation of UNet enhanced with a heteroscedastic classification head (Kendall and Gal, 2017), where logits are modelled as Gaussians, and a sampling softmax method is utilized as described previously. The third model is the **Monte Carlo UNet (MC-UNet)**. This version modifies the standard UNet by incorporating dropout layers before the final two blocks of the encoder and the initial two blocks of the decoder, implementing a dropout probability of 0.1 as specified in (Kendall et al., 2015). Additionally, we assess the **MIMO UNet (MIMO)**, which follows the default configuration detailed in (Baumann et al., 2023) and includes three subnetworks, all without dropout. Finally, we examine **HC-MIMO**, which combines the MIMO UNet structure with the heteroscedastic classification approach, retaining the same hyperparameters as the standard MIMO UNet. This model aims to integrate both types of uncertainties within a single framework.

Each model produces two types of predictions: $P(Y_i = 1|\mathbf{x})$ and $P(Y_i = 0|\mathbf{x})$, which represent the probabilities of a pixel being classified as part of a building or not, respectively. Analog to (Prexl and Schmitt, 2023, Prexl et al., 2023) we upsample (bi-cubic) all input data to 2.5m GSD and use the same GSD for rasterizing the ground truth maps. Therefore, the network architecture predicts $P(Y_i = 1|\mathbf{x}) \in [0, 1]^{512\times512}$ building probability vector for each $\mathbf{x} \in \mathbb{R}^{10\times128\times128}$ input *Sentinel-2* sample.

Abbreviation	Description	Uncertainty
Det-UNet	Deterministic UNet	-
MC-UNet	Monte Carlo dropout	Epistemic
HC-UNet	Heteroscedastic Classification	Aleatoric
MIMO	Multiple Independent Subnetworks	Epistemic
HC-MIMO	MIMO + Heteroscedastic Classification	Alea. + Epis.

Table 1. An overview and the corresponding abbreviations for all five models and corresponding uncertainty measures used throughout this study.

We employ the Adam optimizer with a learning rate of 1.5×10^{-4}, no weight decay, and train for 30 epochs.

Model Evaluation A major challenge when creating building footprint extraction models in regions outside the Western world is the quality of the openly available labels. While in countries like the USA (as well as in many central European countries), ground truth data is either sufficiently available provided by the governmental agencies or through community-driven databases like OSM. For many countries, including our study area of India, this is not the case. Still, if OSM data available,ble it usually holds higher quality building footprint information in comparison to the freely available data sources such as the MBF dataset. In order to give the reader a better intuition, Fig. 1 visualizes the typical occurring label situation in our study area. Therefore, for evaluation purposes, we manually select 15 scenes within the urban area of New-Delhi (total area of $8.63\,\mathrm{km}^2$) where OSM-contributed building footprints are available. Following this strategy, we can avoid taking the lesser quality MBF data into account for evaluation, even though this limits the area for evaluation dramatically and only covers dense build-up areas.

Evaluation Calibration Plots Calibration plots are a basic tool to investigate a model's capability of correctly aligning the predicted probability width with the actual ground truth and represent the main metric to investigate the effect of the different uncertainty methods in this study. To construct calibration plots, we first categorize the predictions from a test set into M distinct bins. These bins are organized based on the predicted class probabilities, denoted as $P(Y = 1|\mathbf{x})$, where $\mathbf{x}$ represents the input features. For each bin B_m with number of samples in the bin $|B_m|$, we assess two key metrics:

Confidence: The frequency with which the model predicts the class $Y = 1$ for the inputs in the bin.

$$\mathrm{conf}(B_m) \coloneqq \frac{1}{|B_m|} \sum_{i \in B_m} P(Y_i = 1|\mathbf{x_i})$$

Frequency: The actual occurrence rate of class $Y = 1$ for these inputs, as determined by the ground truth labels.

$$\mathrm{freq}(B_m) \coloneqq \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}[Y_i = 1]$$

The overall calibration quality is often summarized in the *ex-*

Figure 1. A visual representation of a common data quality issue in developing countries. In a specific area (RGB overview on the left), OpenStreetMap (OSM) labels (shown in the middle image) exhibit the highest geometric precision but are limited in availability. On the other hand, globally accessible MFB labels are suitable for training purposes with moderate-resolution imagery but lack the necessary geometric accuracy for evaluation tasks.

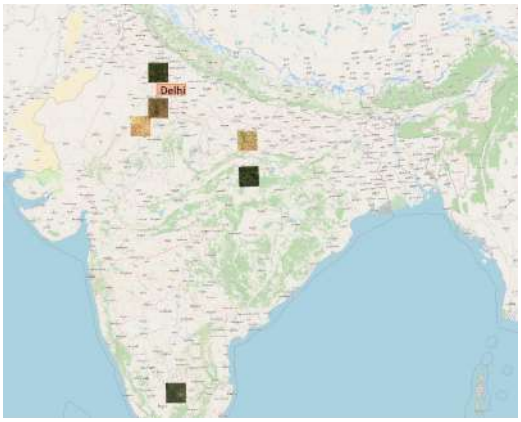

Figure 2. The six locations for the training of our models (in total 21 *Sentinel-2* scenes over different seasons) as well as the location for evaluation.

pected calibration score (ECE), which is defined as

$$\mathrm{ECE} \coloneqq \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{freq}(B_m) - \mathrm{conf}(b_m)\right|$$

where N is the size of the test set.

Further, as a measure of uncertainty, we use entropy, which for the categorical distributed random variable Y is defined as

$$H[Y|\mathbf{X}=\mathbf{x}] \coloneqq -\sum_{y\in\{0,1\}} p(y|\mathbf{x}) \log p(y|\mathbf{x})$$

This measure equals zero when the probability $p(y|\mathbf{x})$ is at either extreme (0 or 1), indicating no uncertainty in the prediction. Conversely, entropy reaches its maximum value when $p(y|\mathbf{x}) = \frac{1}{2}$, reflecting maximum uncertainty.

4. Results

In this section, we present the results obtained by training models with the four above-described uncertainty extensions (plus the naive baseline). We divide the analysis into two subsections where we first show all results for (single-timestep) building footprint segmentation where we in detail compare the results obtained with different uncertainty measures before we show the extension to multi-temporal change detection analysis.

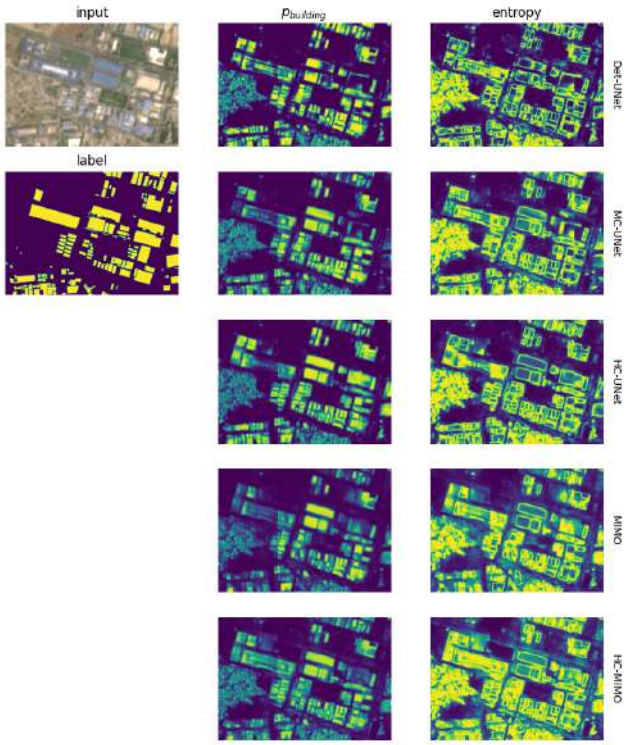

Figure 3. Illustrative results from five distinct UNet-based models are presented across five rows. The leftmost column shows the RGB view and, when available, the corresponding OSM label. In the middle column, you can see the five generated building footprint probability maps. Finally, the right column displays the corresponding entropy plots for each model.

4.1 Mono-Temporal Footprint Segmentation

We test the five models (compare Table 1) for the prediction of building footprint probabilities together with model uncertainty over our study area in India. We show a graphical evolution of the results in Fig. 3 and Fig. 4. The numerical results of the calibration test are provided in Fig. 5 and Table 2.

Figure 3 reveals distinct patterns of uncertainty associated with building sizes and urban layout. High uncertainty levels were observed in areas characterized by dense, unstructured development, particularly affecting small buildings. These buildings displayed lower probabilities of being correctly identified across all methods, with high entropy distributed throughout these densely built-up areas. Conversely, large buildings with

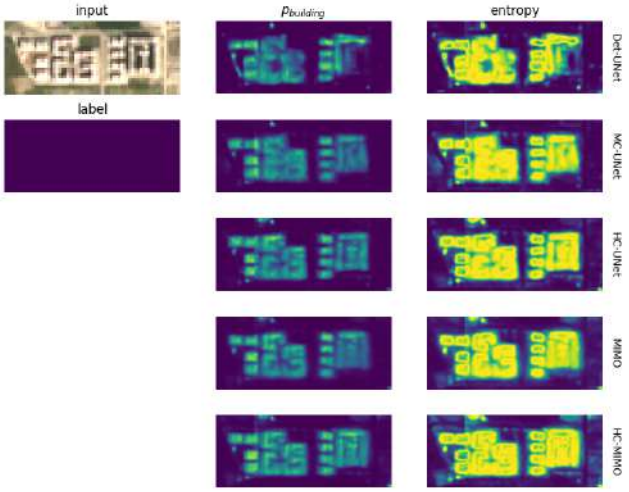

Figure 4. Illustrative results from five distinct UNet-based models are presented across five rows. The leftmost column shows the RGB view and, when available, the corresponding OSM label. In the middle column, you can see the five generated building footprint probability maps. Finally, the right column displays the corresponding entropy plots for each model.

simple shapes exhibited much lower uncertainty, where internal areas of these buildings showed low entropy indicating higher confidence in the predictions, while the boundaries displayed higher entropy (c.f. Figure 3).

Further, we observe that the building footprint predictions remain qualitatively similar when we take the argmax of the probabilistic outputs. The primary differences lie in the predicted building probability and entropy between the single-mode methods (Det-UNet, MC-UNet) and the multi-mode methods (MIMO, HC-MIMO). The single-mode methods generally produce high-quality predictions but tend to be overconfident, as exemplified in Figure 4, where Det-UNet recognizes only fragments of a larger building but doesn't assign high uncertainty to the misclassified areas. Conversely, the multi-mode methods assign lower probabilities to this building but maintain consistent estimations.

These qualitative observations are supported by the calibration plot shown in Figure 5, which indicates a general pattern of underconfidence at lower predicted probabilities and overconfidence at higher probabilities. Det-UNet, in particular, displays significant overconfidence when the predicted probabilities exceed 0.5 (indicative of a pixel predicted as a building). Methods that model only one type of uncertainty—either aleatoric or epistemic (such as MC-UNet, HC-UNet, MIMO)—demonstrate somewhat better calibration. However, HC-MIMO, which accounts for both aleatoric and epistemic uncertainties, exhibits the most accurate calibration, especially in the lower probability ranges.

In our experiments, we observed that the computational effort and training time were comparable across all tested methods. All methods demonstrated similar inference delays, except for MC-UNet. Notably, the inference time for MC-UNet scales linearly with the number of samples used in the model.

4.2 Change Detection Analysis

In this section, we will show the effect of the previously mentioned models with respect to quantitative multi-temporal

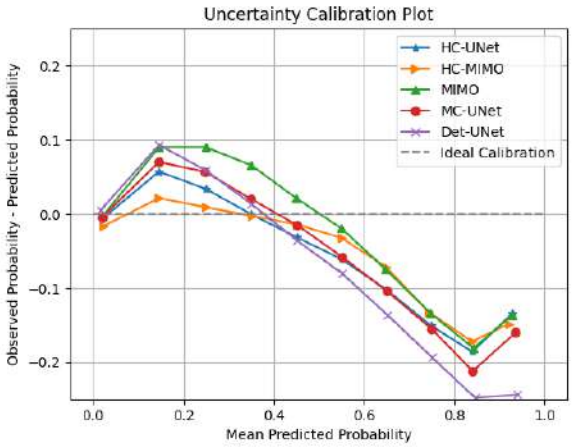

Figure 5. Calibration plot for uncertainty modelling methods tested in our experiments: The methods typically exhibit underconfidence for predicted building probabilities below 0.4 and overconfidence for probabilities above 0.5. HC-MIMO demonstrated the best calibration performance.

Model	↓ ECE
UNet	0.079
MC-UNet	0.057
HC-UNet	0.050
MIMO	0.050
HC-MIMO	**0.028**

Table 2. Expected calibration error of different uncertainty modelling methods used in our experiments shows that focusing solely on either aleatoric or epistemic uncertainty slightly improves calibration. The best calibration results were achieved using HC-MIMO, which integrates both aleatoric and epistemic uncertainties.

change analysis. We will conduct the change detection analysis based on imagery of the years 2019 and 2023 respectively. We manually scan the region of interest for meaningful and diverse sets of occurring changes. Figure 6 shows an example of a changing area and compares the change detection results obtained by the naive baseline **Det-UNet** in comparison to our study's best-performing model **HC-UNet**. A clear trend towards fewer change patterns can be observed when applying the better calibrated **HC-UNet**. With no multi-temporal ground truth available, we have to restrict the evaluation to a qualitative comparison with the pre- and post-imagery. It can be observed that for both cases, A and B for Fig. 6 the reduction in displayed changes is in line with observations in the raw image.

To facilitate the reader a better intuition of the performance differences, we list a set of diverse change situations and their corresponding change detection results in Fig. 7

5. Discussion

We argue that, given the result of the final change detection analysis in Section 4 the influence of the proposed uncertainty extensions is clearly meaningful due to better model calibration regarding the segmentation task on each independent timestep. Figures 6 and 7 clearly illustrate the failure case of change detection when relying on predictions that are not well-calibrated, which we aim to address. Utilizing a deterministic classifier in such scenarios could result in artifacts due to high vari-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-339-2024 | © Author(s) 2024. CC BY 4.0 License.

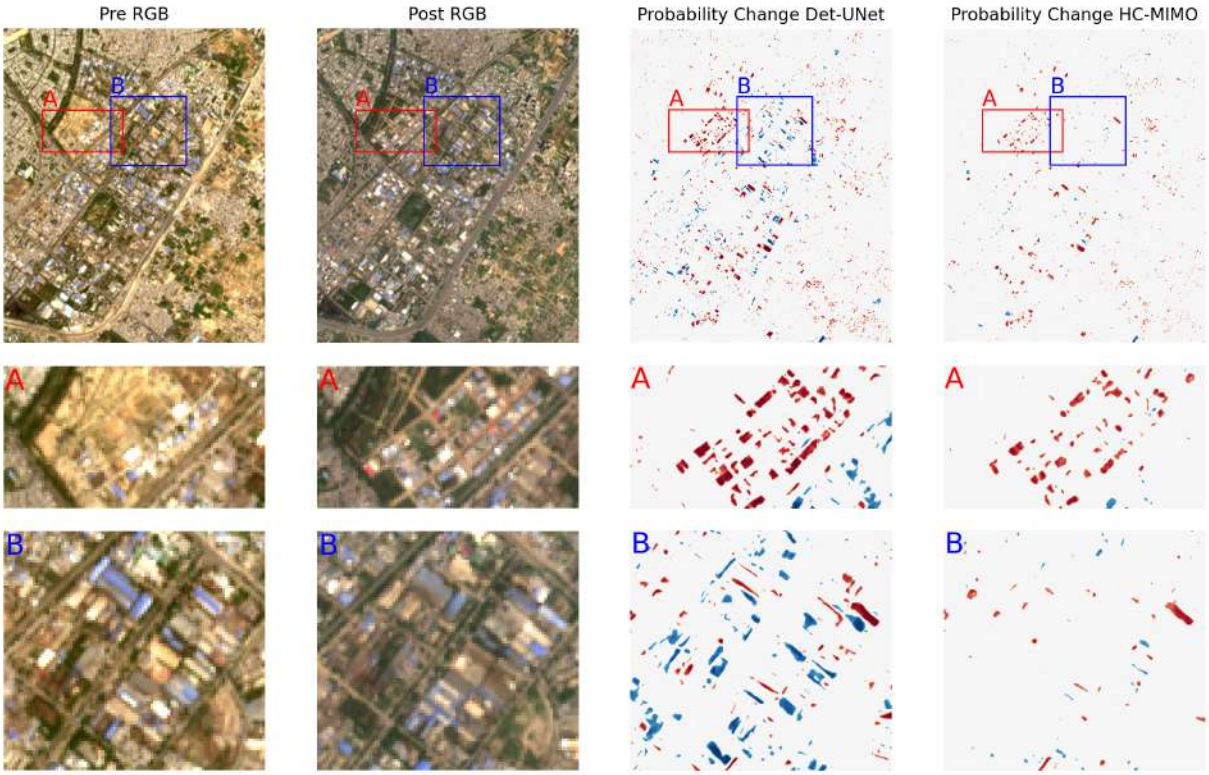

Figure 6. A quantitative change detection example for the New-Delhi area. From left to right: Pre-event RGB, post-event RGB, change in building probability for the naive baseline **Det-UNet** and the change in building probability for the best performing model **HC-MIMO**.

ance in predictions, particularly when processing images under slightly different conditions. A well-calibrated model, on the other hand, can more reliably handle these variations, leading to more stable and accurate predictions. This underscores the importance of integrating robust uncertainty estimation methods that can adapt to varying imaging conditions and ultimately enhance the overall effectiveness of change detection methods in practical applications.

6. Conclusion

This study investigates the extension of building footprint segmentation models by state-of-the-art uncertainty methods for deep learning-based models. The models are trained and evaluated on Indian cities and surroundings, which poses a challenging task due to dense urbanization and the frequent occurrence of buildings with small footprint areas. Evaluation of models was performed on selected areas with sufficient OpenStreetMap ground truth information, while the models generally got trained on freely available, lower quality, Microsoft footprint data. The lack of standardised freely available and large-scale ground truth data is one of the most severe bottlenecks for research towards urban structure monitoring via *Sentinel-2*. While all investigated uncertainty methods deliver reasonable quality, the study shows that the optimal tradeoff between inference speed and good uncertainty calibrations is given by the HC-MIMO architecture, which utilizes several independent subnetworks inside a single model to efficiently emulate a deep ensemble for efficient epistemic uncertainty prediction.

7. Acknowledgement

We thank our college Dr. Deepika Mann for generating the dataset which is used for training the models.

References

Ayala, C., Aranda, C., Galar, M., 2022. Pushing the limits of Sentinel-2 for building footprint extraction. *Proc. IGARSS*, IEEE, 322–325.

Baumann, A., Roßberg, T., Schmitt, M., 2023. Probabilistic MIMO U-Net: Efficient and accurate uncertainty estimation for pixel-wise regression. *Proc. ICCV*, 48, 4498–4506.

Corbane, C., Syrris, V., Sabo, F., Politis, P., Melchiorri, M., Pesaresi, M., Soille, P., Kemper, T., 2021. Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery. *Neural Computing and Applications*, 33, 6697–6720.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. *arXiv:1912.02757*.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proc. ICML*, 48, 1050–1059.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R. et al., 2021. A survey of uncertainty in deep neural networks. *arXiv:2107.03342*.

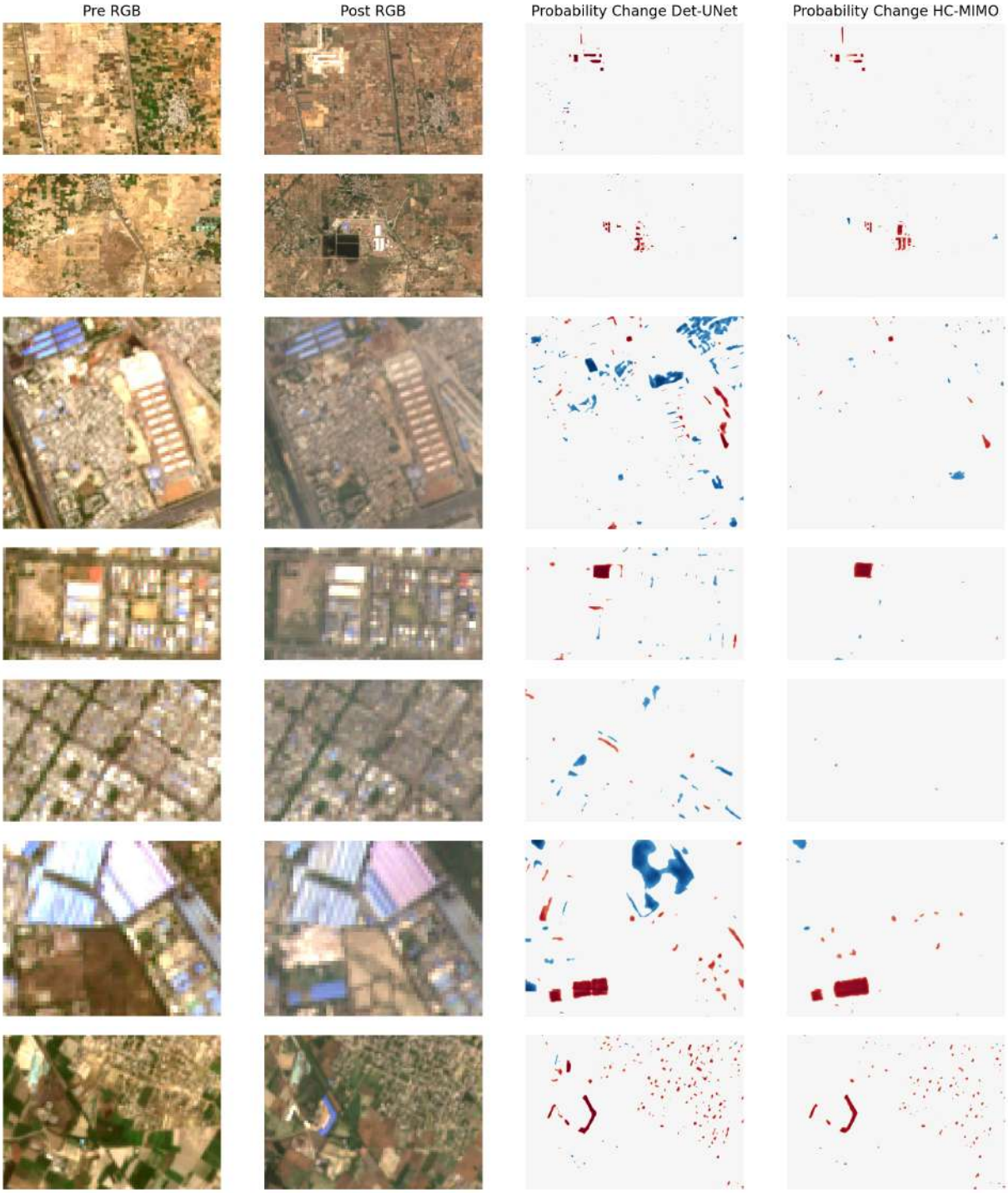

Figure 7. A quantitative change detection example for the New-Delhi area. From left to right: Pre-event RGB, post-event RGB, change in building probability for the naive baseline **Det-UNet** and the change in building probability for the best performing model **HC-MIMO**.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-339-2024 | © Author(s) 2024. CC BY 4.0 License.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., Tran, D., 2020. Training independent subnetworks for robust prediction. *arXiv:2010.06610*.

Hendrycks D., G. K., 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proc. ICLR*.

Huang, X., Wang, C., Li, Z., 2019. High-resolution population grid in the conus using microsoft building footprints: A feasibility study. *Proc. ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 1–9.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv:1511.02680*.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30, 5580–5590.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proc. NeurIPS*, 30, 6405–6416.

Microsoft Building Fooptprints, n.d. `https://github.com/microsoft/USBuildingFootprints`. Accessed: 2022-11-01.

Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2016. Pruning convolutional neural networks for resource efficient inference. *arXiv:1611.06440*.

Nguyen, T.-A., Kellenberger, B., Tuia, D., 2022. Mapping forest in the Swiss Alps treeline ecotone with explainable deep learning. *Remote Sensing of Environment*, 281, 113217.

Nix, D., Weigend, A., 1994. Estimating the mean and variance of the target probability distribution. *Proc. ICNN*, 1, 55–60.

Prexl, J., Saha, S., Schmitt, M., 2023. High precision mapping of building changes using Sentinel-2. *Proc. IGARSS*, IEEE, 6744–6747.

Prexl, J., Schmitt, M., 2023. The potential of Sentinel-2 data for global building footprint mapping with high temporal resolution. *Proc. JURSE*, IEEE.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Proc. MICCAI*, Springer, 234–241.

Schmitt, M., Prexl, J., Ebel, P., Liebel, L., Zhu, X. X., 2020. Weakly supervised semantic segmentation of satellite images for land cover mapping–challenges and opportunities. *arXiv:2002.08254*.

Sibanda, M., Mutanga, O., Rouget, M., 2015. Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 110, 55–65.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., Quinn, J., 2021. Continental-scale building detection from high resolution satellite imagery. *arXiv:2107.12283*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 15, 1929–1958.