TOWARDS A BENCHMARK EO SEMANTIC SEGMENTATION DATASET FOR UNCERTAINTY QUANTIFICATION

Dawood Wasif [1,2], *Yuanyuan Wang* [1], *Muhammad Shahzad* [1,2], *Rudolph Triebel* [3], *Xiao Xiang Zhu* [1]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[3] Institute of Robotics and Mechatronics, German Aerospace Center, Weßling, Germany

ABSTRACT

In order to achieve the objective of accurate and reliable use of deep neural networks for Earth Observation in large-scale scene understanding and interpretation, a large and diverse dataset with proper quantification of uncertainty is required. In this work, we exemplify the lack of a benchmark dataset and present the progress of a novel benchmark dataset for uncertainty quantification of deep learning models in the classic problem of building segmentation from overhead imagery. We present a synthetic dataset where synthetic UAV images were rendered from 3D mesh models of Berlin, Germany. The building masks were extracted from precise LoD-2 building models of the same area. We compare and contrast the performances of baseline methods for semantic segmentation and various uncertainty quantification techniques on this dataset. The experiments show that U-Net is the most accurate model with mIoU of 0.812. Moreover, the Bayesian model is found to be the most reliable uncertainty quantification method on our dataset, with the least ECE.

***Index Terms*—** uncertainty, building segmentation, synthetic, mesh models, Bayesian

1. INTRODUCTION

Accurate and reliable segmentation of buildings from overhead imagery is crucial for many applications such as automatic mapping [1]. With the advent of unmanned aerial vehicles (UAVs) in remote sensing, the resolution and coverage of imagery have significantly improved, leading to a significant boost in the accuracy of building segmentation methods. However, there is still a need to assess the reliability and uncertainty of these results carefully. While epistemic (model) uncertainty can be mitigated with more data or improved models, aleatoric (data) uncertainty, arising from noise inherent in the data, poses a significant challenge, potentially impacting the performance and reliability of segmentation models [2].

Many building segmentation datasets are annotated manually or using existing legacy data or crowd-sourced maps [3], which exhibit different levels of uncertainty. Crowd-sourced maps are prone to errors and inconsistencies due to the subjectivity and variability of the contributors. On the other hand, manual labeling can be time-consuming and may have varing uncertainty depending on the labeling person, making it non-scalable and inefficient for creating large benchmark datasets. Hence, we wonder if automated methods are more suitable for producing accurate, scalable, and controlled segmentation masks that can aid in more effective uncertainty quantification.

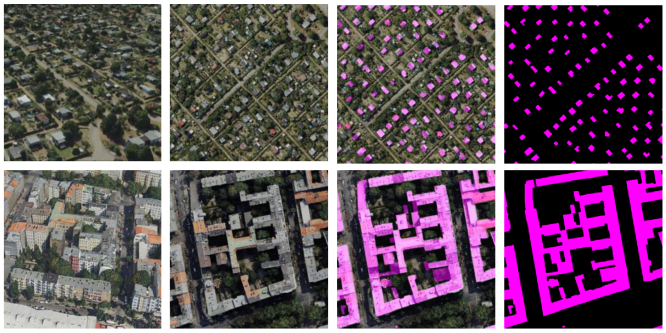

Fig. 1. Sample imagery from our dataset with each first and second row representing a residential and an urban area respectively, comparing the 3D mesh model from different perspectives including an oblique view (col 1) and a nadir view (col 2) with its overlaid ground truth (col 3) and plain ground truth (col 4).

We approach this research question using synthetic UAV datasets, which are artificial images that mimic the appearance and distribution of real-world UAV images. We created a synthetic dataset by capturing nadir views of models downloaded from the Business Location Center Berlin [4]. In this version of the dataset, we provide a consistent semantic segmentation benchmark dataset for accurate and precise building segmentation. This dataset covers different types of urban morphology, as shown in Fig. 1. Future version of this dataset

will include synthetic images with different types of noise.

This dataset is tested using various state-of-the-art semantic segmentation models, such as U-Net, Deeplabv3+, and Feature Pyramid Network (FPN), and uncertainty quantification techniques, such as Bayesian Neural Networks [5], Monte Carlo Dropout [6], Deep Ensembles [7], and Test Time Augmentation [8].

2. DATASET GENERATION

Our dataset consists of 10,000 synthetic overhead RGB images patches covering more than half of the area of Berlin, with a total area of 460 km^2. The dataset has two classes: building and background. The dataset consists of 10,000 RGB image patches and masks, each with a dimension of 214m by 214m with a spatial resolution of 0.2 meters. The data generation pipeline is explained below.

2.1. Data Acquisition

The 3D mesh models were downloaded from Business Location Center Berlin download portal, which allows downloading 3D mesh models as OBJ tiles. . The 3D mesh models utilized for creating this dataset are based on an imaging campaign from August 2020.A total of 10,000 OBJ tiles were downloaded. The corresponding high quality LoD2 models for each of the 3D mesh models were also acquired in order to generate the reference building footprint. These models were combined temporarily in patches of 5x5 to optimize the rendering process.

2.2. Blender Rendering

We simulate the baseline render settings similar to the real-world environment. A camera object of sensor width 0.36 mm is used and placed orthogonally above 3D mesh models at a distance of 150m to simulate a UAV. The format resolution for the output properties setting was set at 8192 x 8192 pixels with a 1:1 aspect ratio. The Cycles rendering engine in Blender was employed for rendering because of its realistic rendering capabilities and performance. The adaptive sampling noise threshold was set at 0.01 for the render properties, and the maximum sample was set at 2048. The exposure and gamma settings were set at a default of 0 and 1, respectively.

3. EXPERIMENTS

The high building densities in urban areas and diverse building sizes, shapes, structures, and textures impose substantial challenges on semantic segmentation methods applied to our dataset, potentially acting as underlying sources of associated uncertainties. Hence, we use three semantic segmentation models and evaluate uncertainty quantification methods on the dataset.

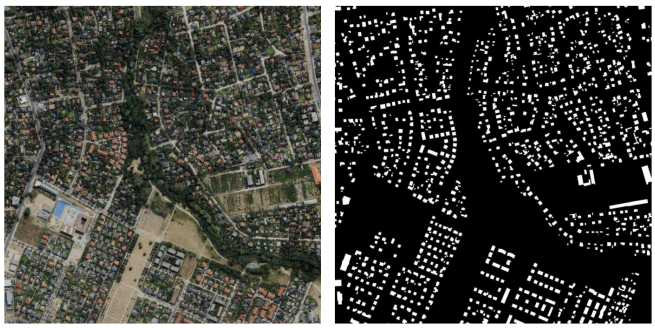

Fig. 2. A sample output pair of the rendering process, depicting an orthographic image render (left) and its corresponding LOD2 model mask (right).

3.1. Segmentation Model Benchmarking

In our efforts to benchmark the performance of different semantic segmentation models, we trained three segmentation models, U-Net, DeepLabv3+, and FPN, on our synthetic dataset. The models were trained using an Adam optimizer with a batch size of 16 and a learning rate of 0.002 for 100 epochs. We utilized a BCE-Dice loss that uses the binary cross-entropy (BCE) loss in conjunction with the Dice loss with a weight factor α, as defined in (1) :

$$L = \alpha L_{BCE} + (1 - \alpha) L_{Dice} \quad (1)$$

All the images were resized to 512 x 512 pixels, and input images were normalized using image mean and variance from ImageNet. Data augmentations such as randomly shifting, scaling, rotating, and adjusting the brightness and contrast of the images were used. To evaluate the performance we used metrics such as Intersection over Union (IoU), F1-score, and pixel accuracy.

Our preliminary experiments are summarized in Table 1, which shows that U-Net is the best-performing model on synthetic data, but with only small margin. Figure 3 shows an example of a predicted mask from each trained model.

Table 1. Comparison of Three Segmentation Models: Deeplabv3+, FPN, and Unet, using mIoU, F1, and Pixel Accuracy Metrics

Model	mIoU	F1	Pixel Acc.
U-Net	0.821	0.892	0.954
DeepLabv3+	0.811	0.885	0.950
FPN	0.813	0.886	0.950

3.2. Uncertainty Quantification Comparison

In the second stage of our experimentation to estimate the uncertainty, we utilized four widely recognized uncertainty

Fig. 3. A comparison of an example image from our dataset and its corresponding ground truth and the predicted mask from three different segmentation models.

Table 2. Comparison of uncertainty metrics from different methods.

Method	Predictive Entropy	Aleatoric Entropy	Mutual Information	Variation Ratio	Total Variance	ECE
BNN	0.359	0.281	0.078	0.0622	1.65 x 10^{-3}	0.156
MC Dropout	0.314	0.260	0.054	0.0492	0.52 x 10^{-3}	0.194
Deep Ensemble	0.287	0.254	0.033	0.032	7.36 x 10^{-4}	0.324
TTA	0.391	0.146	0.245	0.0388	9.78 x 10^{-4}	0.431

quantification methods: Bayesian Neural Network (BNN) with Variational Inference, Monte Carlo (MC) Dropout, Deep Ensemble, and Test Time Augmentation. BNNs apply the principles of Bayesian inference to neural networks by incorporating a prior distribution over their weights. MC Dropout involves randomly deactivating neurons during training and testing phases, generating an ensemble of different network configurations, thereby providing an approximation to Bayesian inference. Deep Ensembles involve training multiple models from different initializations and averaging their predictions, thereby leveraging model diversity to estimate uncertainty. Test Time Augmentation applies modifications to input data during inference to estimate uncertainty from the variability of predictions.

To quantify the model's uncertainty, we calculated Predictive Entropy, Mutual Information, Variation Ratio, and Total Variance [9, 10, 11].

Predictive entropy measures the entropy of the predicted class probabilities for a given input. The equation for predictive entropy is given by:

$$H(y|x, D_{train}) = -\sum_{c=1}^{C}\left(\frac{1}{N}\sum_{n=1}^{N} p(y=c|x,w_n)\right) \left(\frac{1}{N}\sum_{n=1}^{N}\log(p(y=c|x,w_n))\right), \quad (2)$$

where C is the number of classes, N is the number of samples in the dataset, and $p(y=c|x,w_n)$ is the predicted probability of class c for sample n with weights w_n.

Aleatoric entropy measures the uncertainty due to the inherent randomness in the data. The equation for aleatoric entropy is given by:

$$AE(y|x, D_{train}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\left(-p(y=c|x,w_n)\log(p(y=c|x,w_n))\right), \quad (3)$$

where $p(y=c|x,w_n)$ is the predicted probability of class c for sample n with weights w_n.

Mutual information measures the amount of information that the predicted class probabilities provide about the true class. The equation for mutual information is given by:

$$I(y|x, D_{train}) = H(y|x, D_{train}) - AE(y|x, D_{train}). \quad (4)$$

The variation ratio is a measure of the dispersion of a nominal variable. The equation for variation ratio is given by:

$$v := 1 - \frac{f_m}{N}, \quad (5)$$

where (f_m is the number of predictions falling into the modal class category).

Total variance is another measure that is used to estimate the spread of predicted confidences from each sample. The equation for total variance is given by:

$$\sigma^2 = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{n=1}^{N}\left(p(y=c|x,w_n) - \frac{1}{N}(p(y=c|x,w_n)\right)^2, \quad (6)$$

where $p(y=c|x,w_n)$ is the probability of class c given input x and weights w_n, for the n^{th} Monte Carlo sample.

In addition to these metrics, we also computed the Expected Calibration Error (ECE) with predictive entropy. ECE

quantifies the reliability of a model's predictive uncertainty, reflecting the degree of alignment between the model's confidence in its predictions and the actual correctness of those predictions. The equation for ECE is given by:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |\text{acc}(B_m) - \text{conf}(B_m)|, \qquad (7)$$

where B_m represents the set of instances (predictions) that fall into the m^{th} bin and n is the total number of instances.

The results are listed in Table 2. The comparison of the performance of each uncertainty quantification methods will be detailed in the future work, because of the lack of uncertainty ground truth in the dataset. However, the result indicates a low ECE for the BNN, which signifies a high level of reliability in its uncertainty estimation. This could potentially be attributed to the Bayesian model's robust posterior approximation to yield more calibrated uncertainty estimates closer to the true outcomes.

4. CONCLUSION

This paper presents an innovative approach to semantic segmentation and uncertainty quantification in remote sensing using synthetic datasets. The proposed synthetic dataset, derived from 3D models of Berlin city, offers enhanced control and variability compared to traditional real-world datasets. Berlin's distinctive and diverse architectural landscape makes it an excellent choice for building-related computer vision tasks. We experimented with multiple state-of-the-art semantic segmentation models and uncertainty quantification methods with relevant metrics to constitute a future research baseline. U-Net emerged as the best-performing model on our dataset and thus most suited for synthetic datasets, whereas BNN is the most reliable uncertainty quantification method due to its inherent probabilistic nature. Through our proposed dataset and its demonstrated capabilities, we encourage further research in building analysis using synthetic datasets and the development of robust and reliable semantic segmentation models with accurate uncertainty estimates.

5. REFERENCES

[1] Guangming Wu, Xiaowei Shao, Zhiling Guo, Qi Chen, Wei Yuan, Xiaodan Shi, Yongwei Xu, and Ryosuke Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sensing*, vol. 10, no. 3, pp. 407, 2018.

[2] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al., "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[3] Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[4] "Berlin 3d - downloadportal des business location centers," `https://www.businesslocationcenter.de/downloadportal`, 2023.

[5] Igor Kononenko, "Bayesian neural networks," *Biological Cybernetics*, vol. 61, no. 5, pp. 361–370, 1989.

[6] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[9] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1184–1193.

[10] Jishnu Mukhoti and Yarin Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

[11] Yarin Gal, Riashat Islam, and Zoubin Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.