Improving Building Segmentation Using Uncertainty Modeling and Metadata Injection

Hanxiang Hao
Video and Image Processing Lab (VIPER), Purdue University
West Lafayette, Indiana, USA

Sriram Baireddy
Video and Image Processing Lab (VIPER), Purdue University
West Lafayette, Indiana, USA

Kevin LaTourette
Optical Payload Center of Excellence, Lockheed Martin Space
Littleton, Colorado, USA

Latisha Konz
Optical Payload Center of Excellence, Lockheed Martin Space
Littleton, Colorado, USA

Moses Chan
Advanced Technology Center, Lockheed Martin Space
Sunnyvale, California, USA

Mary L. Comer
Video and Image Processing Lab (VIPER), Purdue University
West Lafayette, Indiana, USA

Edward J. Delp
Video and Image Processing Lab (VIPER), Purdue University
West Lafayette, Indiana, USA

ABSTRACT

Automatic building segmentation is an important task for satellite imagery analysis and scene understanding. Most existing segmentation methods focus on the case where the images are taken from directly overhead (*i.e.*, low off-nadir/viewing angle). These methods often fail to provide accurate results on satellite images with larger off-nadir angles due to the higher noise level and lower spatial resolution. In this paper, we propose a method that is able to provide accurate building segmentation for satellite imagery captured from a large range of off-nadir angles. Based on Bayesian deep learning, we explicitly design our method to learn the data noise via aleatoric and epistemic uncertainty modeling. Satellite image metadata (*e.g.*, off-nadir angle and ground sample distance) is also used in our model to further improve the result. We show that with uncertainty modeling and metadata injection, our method achieves better performance than the baseline method, especially for noisy images taken from large off-nadir angles[1].

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**.

KEYWORDS

satellite imagery analysis, building segmentation, Bayesian deep learning, uncertainty modeling

[1] An extended version of the this paper can be found at [6]

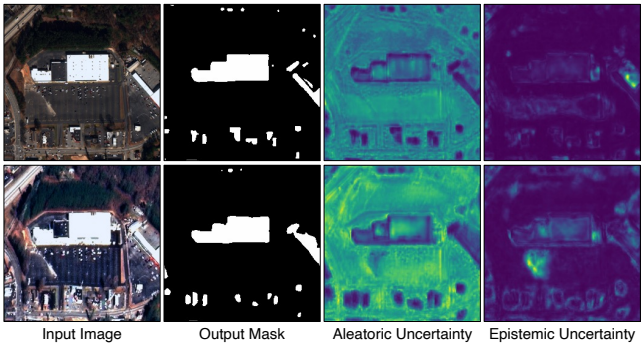

Figure 1: Building segmentation results of the proposed method for the small off-nadir angle (first row) and large off-nadir angle (second row). Aleatoric and epistemic uncertainty maps highlight the noisy regions, especially for the case with large off-nadir angle, in order to guide the model ignoring these noisy regions during training.

ACM Reference Format:
Hanxiang Hao, Sriram Baireddy, Kevin LaTourette, Latisha Konz, Moses Chan, Mary L. Comer, and Edward J. Delp. 2021. Improving Building Segmentation Using Uncertainty Modeling and Metadata Injection. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3474717.3483918

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

SIGSPATIAL '21, November 2–5, 2021, Beijing, China

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8664-7/21/11.
https://doi.org/10.1145/3474717.3483918

1 INTRODUCTION

Object segmentation for satellite imagery has been studied extensively because of the availability of computational resources and large-scale datasets [1, 4]. Although many existing methods achieve accurate segmentation results, using them in real-world applications is still challenging. Unlike many segmentation tasks for natural images, real-world object segmentation for satellite imagery often faces challenges in identifying small, visually heterogeneous

objects (*e.g.*, cars and buildings) with varying orientation and density in images [14]. For example, it is even hard for humans to recognize the small buildings from the images in Figure 1, because of the low lighting condition and image blur/noise. Furthermore, due to changes in the satellite off-nadir angle, the appearance of target objects can vary dramatically, including changes in lighting intensity, object resolution, and image noise level. The satellite off-nadir angle (*i.e.*, viewing angle) is the angle between the nadir point directly below the satellite and the center of the imaged scene [14]. As shown in Figure 1, from a small off-nadir angle (first row) to a large off-nadir angle (second row), the overall image intensity and image quality drops significantly.

In order to address these challenges, we present a building segmentation method with uncertainty modeling and satellite image metadata injection. Our method is able to provide accurate segmentation results when training with noisy images. More specifically, based on Bayesian deep learning, the proposed method is designed to capture both model and data uncertainty to ignore the image regions with a higher uncertainty level. For example, as shown in Figure 1, our uncertainty maps highlight the areas with larger image noise (*e.g.*, building boundaries or forest due to the image noise). As the off-nadir angle increases, the uncertainty level increases, indicating a higher data noise. Furthermore, satellite image metadata is also considered in our method. In this paper, we use ground sample distance (GSD) and off-nadir angle as input metadata. GSD describes the spatial resolution of the image and a larger GSD usually indicates noisier images. As mentioned earlier, different off-nadir angles can also cause changes in image quality. In this paper, we propose two metadata injection methods in Section 2.2 to show the effectiveness of using metadata in building segmentation. The main contributions of this paper are summarized as follows:

- we design a building segmentation model that is able to capture both model uncertainty (*i.e.*, epistemic uncertainty) and data uncertainty (*i.e.*, aleatoric uncertainty);
- two metadata injection methods are developed for using satellite image metadata to improve building segmentation;
- based on our experimental analysis, we show that the proposed method is able to achieve a better performance than the baseline method, especially for noisy images taken at large off-nadir angles.

2 METHOD

In this section, we will introduce our building segmentation method with uncertainty modeling and satellite image metadata injection. As shown in Figure 2, the proposed method is based on U-Net [12] and has multiple outputs. As described in Section 2.1, modeling uncertainty enables our method to ignore the noisy pixels that are caused by blurry or noisy images. Injecting satellite image metadata such as ground sample distance (GSD) and off-nadir angle provides the model with more information to improve its performance. We will provide two metadata injection approaches in Section 2.2.

2.1 Modeling Uncertainty via Bayesian Deep Learning

Unlike standard deep learning methods, Bayesian deep learning (Bayesian DL) provides a model with the ability to ignore certain data points based on uncertainty. In Bayesian DL, there are two types of uncertainty one can model:

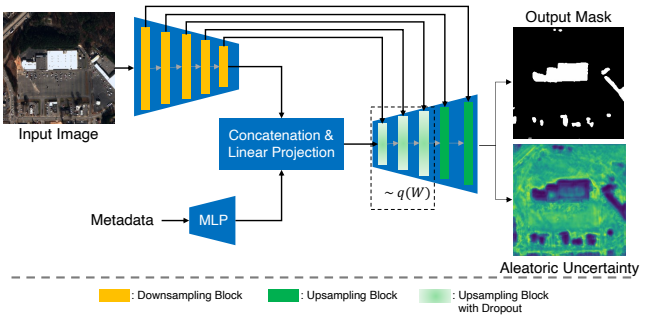

Figure 2: The block diagram of the proposed method with uncertainty modeling and concatenation-based metadata injection. $q(W)$ is the dropout variational distribution for modeling epistemic uncertainty.

- *Epistemic Uncertainty* describes the uncertainty that is caused by the model ignoring some training data. For example, a segmentation model might miss some building areas with certain colors/textures. Usually, this type of uncertainty can be reduced as more training data is made available.
- *Aleatoric Uncertainty* describes the uncertainty that is inherited from data (*e.g.*, image/sensor noise). Aleatoric uncertainty can be further categorized as *homoscedastic uncertainty*, which is the uncertainty based on the entire dataset, and *heteroscedastic uncertainty*, which is the uncertainty for each input data point (*i.e.*, each pixel in our case). In this paper, we will consider heteroscedastic aleatoric uncertainty to accurately model the data noise for different input images.

Following the work [3], we use Monte Carlo dropout to model the epistemic uncertainty (*i.e.*, the dropout layers from the decoder in Figure 2). For aleatoric uncertainty, as proposed in [8], we directly output the uncertainty map from the last layer of our model as shown in Figure 2 and use Gaussian corruption during training to model the aleatoric uncertainty. Please refer to [3, 8] for the implementation details of uncertainty modeling.

2.2 Metadata Injection

Satellite image metadata contains useful information to support many computer vision tasks, such as using solar and satellite azimuth and elevation angles for shadow detection and building height estimation [5, 11, 13]. In this paper, we consider two types of metadata to improve the building segmentation result: (1) ground sample distance (GSD); and (2) off-nadir angle. GSD describes the spatial resolution of the image; a larger GSD indicates blurrier and noisier images due to lower image resolution. Off-nadir angle describes the viewing angle of the satellite camera and a larger off-nadir angle can also cause lower image resolution. In the following sections, we will provide two metadata injection approaches to improve the baseline U-Net model.

2.2.1 *Metadata Injection via Feature Concatenation.*

As shown in Figure 2, we first pass the metadata vector to multi-layer perceptrons (MLP) to obtain the output vector ($\mathbf{h} \in \mathbb{R}^D$) for feature extraction and dimension expansion. Then we combine the metadata feature vector with the image features ($\mathbf{v} \in \mathbb{R}^{H \times W \times D}$) obtained from the last CNN encoder layer. To combine metadata and image features, we repeat the metadata feature vector to match the shape of image features, getting $\mathbf{h}' \in \mathbb{R}^{H \times W \times D}$. Then we concatenate the features along the channel dimension as $\mathbf{h_v} \in \mathbb{R}^{H \times W \times 2D}$. The final features can be obtained by linearly projecting the channel dimension back to the input channel dimension: $\mathbf{o} = \mathbf{F}(\mathbf{h_v}) \in \mathbb{R}^{H \times W \times D}$, where $\mathbf{F} : \mathbb{R}^{2D} \rightarrow \mathbb{R}^D$ is applied for each input element and it can be implemented by a convolutional layer with kernel size of 1. We refer to this concatenation-based approach as *MetaCat*.

2.2.2 *Metadata Injection via Affine Combination Module.*

As described above, the previous concatenation-based metadata injection method combines the metadata and image features by channel-wise concatenation following a linear projection layer. By doing so, we augment the image features using the metadata features for every location in the H and W dimensions evenly. However, intuitively, not all image features need to be modified. For example, since we focus on building segmentation, a large forest area should not be considered and modified. To effectively locate the desired regions that need to be modified, we use the Affine Combination Module (ACM) [10] for metadata injection. As the name indicates, ACM is based on affine transforms and can be formulated as follows:

$$\mathbf{v}' = \mathbf{h} \odot W(\mathbf{v}) + b(\mathbf{v}), \tag{1}$$

where $\mathbf{v}$ is the image features obtained from the CNN encoder, $\mathbf{h}$ is either the repeated metadata features $\mathbf{h}'$ as described in Section 2.2.1 or the features from the previous decoder layer, and $W(\cdot)$ and $b(\cdot)$ are convolutional layers as proposed in [10]. Modified from Figure 2, we replace the concatenation operators from both metadata injection (*i.e.*, the concatenation operator that combines the metadata feature with the encoder feature from the bottleneck layer) and U-Net skip connections (*i.e.*, concatenation operators that combine the encoder features with decoder features) with multiple ACMs. From Equation 1, we can consider the $W(\mathbf{v})$ term as the metadata-relevant information, since it can directly interact with the metadata features (or the previous decoder features). The $b(\mathbf{v})$ term can be considered as a metadata-irrelevant information that is not modified by the metadata features (or the previous decoder features). We refer to this ACM-based approach as *MetaACM*.

3 EXPERIMENT

3.1 Dataset and Experiment Setting

In this paper, we use the SpaceNet 4 dataset [14], which is designed for building segmentation with a larger range of off-nadir angles. There are 1,064 distinct locations in the dataset, with 27 images captured at each location at off-nadir angles ranging from $-32.5°$ to $54°$, which totals to 28,728 images. We partition the dataset into training, validation, and testing sets with the ratio of $6:2:2$. As mentioned in [14], the building annotations from the SpaceNet 4 dataset are obtained from the images with the smallest (in magnitude) off-nadir angle ($-7.8°$), and the same annotations are used for the other images with different off-nadir angles. This will cause inaccurate annotations due to the changes in building appearance caused by different off-nadir angles. To have an accurate evaluation, we manually label the testing images with off-nadir angles greater than $40°$. Note that we do not relabel the training data in order to evaluate if uncertainty modeling can handle the both image noise and annotation noise.

To ensure fair comparison between the proposed method and the baseline U-Net, all of our experiments used the same setting, which we will now describe. The downsampling blocks (yellow blocks) in Figure 2 are the residual blocks from a ResNet-34 model [7] pre-trained on ImageNet [2]. The upsampling blocks (dark green blocks) consist of *bilinear upsampling* → *convolution* → *batch normalization* → *ReLU*. The upsampling blocks with dropout (light green blocks) consist of *bilinear upsampling* → *dropout* → *convolution* → *batch normalization* → *ReLU*. Following [8], the dropout rate is set as 0.2. The MLP for metadata feature extraction consists of three blocks, where each block is a fully-connected layer following by a leaky ReLU layer with a slope of 0.2. During training, to allow for a larger batch size as required by batch normalization, we resize the input image to 256 with batch size as 64. ADAM optimizer [9] with learning rate 0.0001 (linear decay) is used and all experiments are trained for 1 million iterations. We use weight decay with the factor of 0.0001 for all experiments. For the Monte Carlo integration during inference, following [8], we set the number of samples as 50.

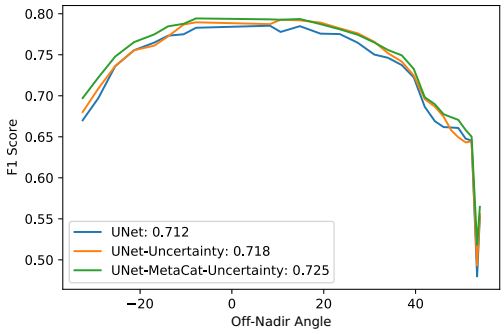

Figure 3: Testing F1 scores with different off-nadir angles. The average F1 scores of all off-nadir angles are shown in the legend.

3.2 Experimental Result and Analysis

We start with evaluating the use of uncertainty modeling and metadata injection (we consider MetaCat first and then compare MetaCat with MetaACM later). Figure 3 shows the F1 scores with different off-nadir angles in the testing set. Compared with the baseline U-Net, with uncertainty modeling, there is a slight improvement across most of the off-nadir angles. Adding the metadata injection layer can further improve the performance, especially for the cases with larger off-nadir angles ($> 40°$) and negative off-nadir angles.

As mentioned in [14], due to the data collection process, the images with large negative off-nadir angles have very different lighting conditions and shadows. Since most of the images are collected from positive off-nadir angles, the baseline method will suffer from unbalanced data during training. With metadata injection and uncertainty modeling, the proposed method is able to deal with the changes of lighting and shadows.

Figure 1 shows the prediction difference of two images with same scene but different off-nadir angles. Based on the ground truth, we can see that the proposed method can accurately detect building area even under this high noise-level condition. We can see that overall, both uncertainty maps increase the highlighted areas from small to large off-nadir angles due to higher noise in the input image. The aleatoric uncertainty has higher data noise around the forest region compared to the building region. This is due to the larger appearance variance of forests compared to buildings. Unlike aleatoric uncertainty, epistemic uncertainty focuses more around the buildings. It highlights the area where the predictions are not reliable, such as the boundary of buildings, due to the image noise.

Table 1: F1 scores for ACM-based and concatenation-based metadata injection. All of the listed experiments are based on U-Net with uncertainty modeling of both aleatoric and epistemic uncertainties. *None* means no metadata injection.

Experiment	Nadir	Off-Nadir	Very Off-Nadir	Overall
None	0.7752	0.7359	0.6347	0.7180
MetaCat	**0.7822**	**0.7429**	0.6415	**0.7249**
MetaACM	0.7758	0.7382	**0.6419**	0.7197

We compare the two metadata injection methods in Table 1. As defined in [14], we group the images into three categories based on the absolute off-nadir angle. Overall, MetaCat achieves better performance than MetaACM. Compared with the method without metadata injection, MetaCat has significant improvement for all three off-nadir angle categories. Although MetaACM does not have a major improvement for the lower off-nadir angle images, it achieves the best performance under the *Very Off-Nadir* category. In our experiments, MetaACM does not achieve a better performance than MetaCat. As mentioned in Section 2.2.2, since MetaACM has higher flexibility to modify the intermediate features in different spatial locations, we believe it has a greater potential to achieve better segmentation performance than MetaCat. [6] provides more results and analysis to show the effectiveness of MetaACM model.

4 CONCLUSION

In this paper, we propose a method that can provide accurate building segmentation despite the data noise that is caused by large off-nadir angles. Both aleatoric uncertainty and epistemic uncertainty are modeled by our method to enable our model to learn from noisy training data. Based on the level of predicted uncertainty, the proposed method learns to ignore the area with larger uncertainty and focus on the area with less uncertainty. Satellite image metadata is also considered to further improve the performance. We propose concatenation-based and ACM-based metadata injection methods to effectively use metadata for the building segmentation task. With our experimental analysis, we show that the proposed method is able to achieve a clear improvement compared to the baseline method, especially for the noisy images taken from large off-nadir angles.

ACKNOWLEDGMENTS

This material is based on research sponsored by Lockheed Martin Space. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of Lockheed Martin Space.

REFERENCES

[1] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G. Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. 2020. Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2020), 2825–2835. https://doi.org/10.1109/CVPR42600.2020.00290 Seoul, Korea.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848 Miami, FL.

[3] Yarin Gal and Zoubin Ghahramani. 2016. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *International Conference on Machine Learning* (May 2016). https://arxiv.org/abs/1506.02158 San Juan, Puerto Rico.

[4] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (June 2019). https://arxiv.org/abs/1911.09296 Long Beach, California.

[5] Hanxiang Hao, Sriram Baireddy, Emily Bartusiak, Mridul Gupta, Kevin LaTourette, Latisha Konz, Moses Chan, Mary L. Comer, and Edward J. Delp. 2021. Building Height Estimation via Satellite Metadata and Shadow Instance Detection. *Automatic Target Recognition XXXI* 11729 (April 2021), 175–190. https://doi.org/10.1117/12.2585012

[6] Hanxiang Hao, Sriram Baireddy, Kevin LaTourette, Latisha Konz, Moses Chan, Mary L. Comer, and Edward J. Delp. 2021. Improving Building Segmentation for Off-Nadir Satellite Imagery. *arXiv* 2109.03961 (September 2021). https://arxiv.org/abs/2109.03961

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2016), 770–778. https://doi.org/10.1109/CVPR.2016.90 Las Vegas, NV.

[8] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Conference on Neural Information Processing Systems* 30 (December 2017). https://arxiv.org/abs/1703.04977 Long Beach, CA.

[9] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (May 2015). http://arxiv.org/abs/1412.6980 San Diego, CA.

[10] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H.S. Torr. 2020. ManiGAN: Text-Guided Image Manipulation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* 30 (June 2020), 7877–7886. https://doi.org/10.1109/CVPR42600.2020.00790 Seattle, WA.

[11] Gregoris Liasis and Stavros Stavrou. 2016. Satellite images analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing* 19 (2016), 437–450. https://doi.org/10.1016/j.isprsjprs.2016.07.006

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention* 9351 (June 2015), 234–241. https://arxiv.org/abs/1505.04597

[13] Alexey Trekin, Vladimir Ignatiev, and Pavel Yakubovskiy. 2020. Deep Neural Networks for Determining the Parameters of Buildings from Single-Shot Satellite Imagery. *Computer and Systems Sciences Internationa* 59 (2020), 755–767. https://doi.org/10.1134/S106423072005007X

[14] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Etten, Varun Kumar, Sean Mcpherson, Jacob Shermeyer, and Hanlin Tang. 2019. SpaceNet MVOI: A Multi-View Overhead Imagery Dataset. *IEEE/CVF International Conference on Computer Vision* (October 2019), 992–1001. https://doi.org/10.1109/ICCV.2019.00108 Seoul, Korea.