3D SEMANTIC SEGMENTATION FROM MULTI-VIEW OPTICAL SATELLITE IMAGES

Pablo d'Angelo, Daniele Cerra, Seyed Majid Azimi, Nina Merkle, Jiaojiao Tian,
Stefan Auer, Miguel Pato, Raquel de los Reyes, Xiangyu Zhuo, Ksenia Bittner,
Thomas Krauss, Peter Reinartz
{Pablo.Angelo, Firstname.Lastname}@dlr.de
Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen

ABSTRACT

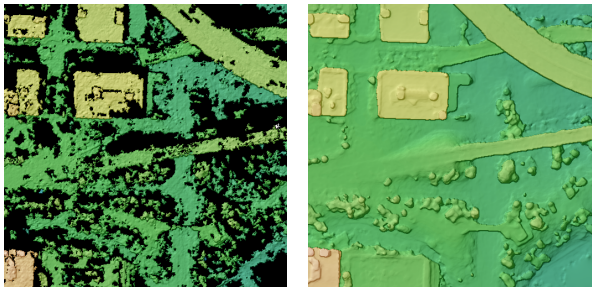
This paper describes the winning contribution to the 2019 IEEE GRSS Data Fusion Contest Multi-view Semantic Stereo Challenge. In this challenge, a digital surface model (DSM) and a semantic segmentation should be derived from a large number of multi-spectral WorldView-3 images. Results from 50 stereo pairs matched using Semi-Global Matching (SGM) are fused into a DSM. Semantic segmentation is performed with an ensemble of FCN networks taking as input RGB, multi-spectral and height data. Their results are then merged with pixel-wise detectors for the classes water and high vegetation. Compared to the second and third placed teams (mIOU-3 scores of 0.73 and 0.7295), our contribution reached a significantly higher score of 0.745.

1. INTRODUCTION

The 2019 IEEE GRSS Data Fusion Contest Multi-view Semantic Stereo Challenge [1] aims to promote Semantic 3D Reconstruction from satellite imagery. The challenge is based on the Urban Semantic 3D data set [2], which includes incidental satellite images, airborne lidar, and semantic labels covering approximately 20 square kilometers over two cities. Source data consists of 26 images collected between 2014 and 2016 over Jacksonville, Florida, and 43 images collected between 2014 and 2015 over Omaha, Nebraska, United States. This paper describes the contest winning approach developed at the German Aerospace Center (DLR).

2. METHOD

Using image orientation refined by bundle block adjustment, Semi-Global Matching (SGM) was used to produce height maps, digital surface models (DSM) and normalized DSM (nDSM). Convolutional Neural Network (CNN) based semantic segmentation was performed on the RGB, multi-spectral images (MSI) and height maps and projected into UTM coordinates. Pixel-wise detectors were applied to the orthorectified MSI images, deriving binary maps for the classes *high vegetation* and *water*. An ensemble of 3 CNN classifiers was merged with the ad hoc detectors to obtain the final semantic segmentation maps, after an additional step of morphological filtering.



(a) (b)

Fig. 1: DSM after matching (a) single stereo pair and (b) merging of 50 stereo pairs.

2.1. Image Orientation

Before performing dense matching, a good relative image orientation is required. As the contest data set was only coarsely aligned to the reference data, additional relative orientation was required in order to avoid systematic height offsets between individual stereo pairs. For image orientation and dense matching, a synthetic panchromatic image was generated by averaging the red, green and blue channels of the MSI images. Multi-ray tie points were matched using SIFT and refined and transferred to unmatched images using local least squares matching. Bias corrected RPCs were then obtained using bundle block adjustment [3].

2.2. Multi-View 3D Reconstruction

Following [4], we performed dense stereo matching using pairwise SGM using CENSUS as matching cost. Due to the difference in image acquisition time, dense matching of single stereo pairs yields incomplete results, particularly in areas with changes and vegetation, cf. Fig. 1. All possible stereo pairs with a convergence angle above a predefined threshold were ranked based on the number of tie points found in the image orientation step, matching the 50 pairs with the highest amount of tie points. Each pair is matched in both directions,

resulting in 100 height maps. We computed height clusters for every pixel in the final DSM and selected the mean height of the cluster with the highest number of points. In addition to the DSM heights, we produced quality layers containing the number of matches and standard deviations of all height values. Remaining holes were filled using interpolation. Finally, all images were orthorectified using the DSM.

We generated a digital terrain model (DTM) from the DSM using an adapted version of the method reported in [5], which analyzes height steps and slopes along multi-directional trajectories at each DSM pixel. Holes in the DTM were closed based on interpolation, and a normalized DSM (nDSM) was generated for obtaining relative heights (DSM minus DTM). In addition to the DSM in UTM coordinates, we computed height maps for each input image by reprojecting the point cloud obtained from all stereo pairs into the original satellite images. These height maps allow the additional use of dense height information during semantic segmentation of the input images in sensor geometry.

2.3. Semantic Classification

Semantic classification is performed by utilizing three different neural network architectures, plus two ad hoc approaches for the classes *high vegetation* and *water*. The provided RGB and MSI images, along with the dense height maps generated during our stereo matching, are used as input for the classification. Note that the third network of choice is the provided baseline U-Net network [1] and is therefore not described below.

2.3.1. Multi-modal Fusion Network

The first network used is a multi-modal fusion network and is based on the fully convolutional network (FCN) architecture proposed in [6]. It is trained on a stack of dense height maps (generated during stereo matching), RGB and NIR image triplets. The basis of the network forms two consecutive parts, where the first one acts as an encoder, down-sampling the input images in order to extract high-level features, and the second one as a decoder, gradually up-sampling the encoded features to obtain a feature map having the required output size. In contrast to the original FCN architecture, 4 convolution layers are employed instead of 2. We carried out the experiments using a NVIDIA Titan XP GPUs, training the network for 20 epochs using the Adam optimizer with a learning rate of 0.0001.

2.3.2. Small and large structure-sensitive CNN

Inspired by the works of [7, 8, 9, 10], we use a second network named "Small and large structure-sensitive CNN" (SLSS-CNN). It consists of two streams and is trained on RGB images only. The first stream is a small-structure-sensitive one consisting of several sub-blocks, where each block contains convolution layers, but includes no pooling operation in order to preserve features related to small objects. This stream uses batch-normalization and a drop-out layer to attenuate overfitting. Furthermore, we use several residual paths inside each block to allow the flow of input data to the last layers. Although removing pooling layers decreases the growth rate of the receptive field, it refines the object boundaries. Moreover, this prevents data loss during the sub-sampling steps. As a drawback, the removing of the pooling layers causes a loss of depth in the network and leads to lower-level features. This results in poor performance, as high-level features are one of the main reasons of deep CNNs' success. To alleviate this, we consider a large-structure-sensitive stream containing several blocks which contain pooling layers in contrast to the first stream, preserving the convolution and batch-normalization layers. The features extracted in parallel from the two streams are combined after each block, from the input layer to the output layer.We use 5 max-pooling and up-sampling operations in the pooling stream to extract and decode rich semantics. We apply a convolution layer with 1×1 kernel size to the concatenation of the output of both streams to reduce the feature maps, followed by argmax to create the final results. We normalize the input images before feeding them to the network. The experiment configurations during the training phase are the same as for the network in Subsection 2.3.1.

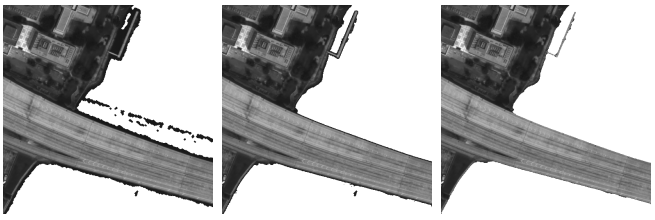
2.3.3. Water Detection

The water masks have been detected in three steps on the average images of all available orthorectified acquisitions. In the first step, all pixels with low consistency in the DEM have been selected as initial water candidates. The consistency has been estimated as the number of DEMs in which an image element has been matched, by selecting pixels matched in less than 5 stereo pairs. Among these, the selected water pixels have a Normalized Difference Water Index (NDWI) [11] above $\max(0.35, OtsuT)$, where 0.35 is an empirical fixed value, and $OtsuT$ represents an adaptive Otsu threshold. The adaptive threshold plays an important role, as areas with water bodies usually exhibit a bimodal distribution of the water index. In a second step, the average spectrum of the water pixels is computed, and spectrally similar image elements are added to the water mask, by thresholding the Spectral Angle (SA) [12] distance map to 0.072. In the final step, a larger SA distance (0.12) is applied to complete the water mask by adding similar pixels at a maximum distance of 50 m from the already detected water bodies. Morphological opening and closing operations are applied to refine the final results. For an example of the three water detection steps see Figure 2.

2.4. Classification Fusion

The classifiers described in the previous sections were merged using the SLSS-CNN as base classifier for the ground, high

[1] https://github.com/pubgeo/dfc2019



(a) (b) (c)

Fig. 2: Illustration of the three water detection stages.

(a) (b)

Fig. 3: Illustration of the detected elevated roads (a) before and (b) after the refinement.

vegetation and building classes. Water and elevated road pixels were removed and replaced with the closest valid label or ground, if the nDSM height was smaller than 5 m. Then elevated road pixels from the SLSS, Multi-modal Fusion, and the baseline U-Net network with a nDSM height value higher than 1.5 m were added. High vegetation was completed by adding the results of the described high vegetation detectors. The water mask was added in the final step.

2.5. Classification Refinement

Deep CNN classifications have problems with objects larger than the network's receptive field, resulting in few instances of incomplete labels (false negatives) for very large buildings. The buildings labels have been then refined by morphological closing, using a large square structuring element (29×29) applied to large structures only (buildings containing more than 5000 pixels). An additional opening with a smaller structuring element (square, 7×7) was then performed, and only changes on pixels higher than 11 meters in the nDSM where kept.

The bridge labels where refined by adding to the class elevated pixels with similar gradient to the detected objects, and having a spectral distance (SA) smaller than 0.1 from the average spectrum of the detected bridges. Finally, small objects (with less than 500 pixels) were removed from the class. An example showing the effects of the refinement is provided in Figure 3.

2.5.1. High Vegetation Detection

The results from two tree detectors have been overlaid on the CNN results, respectively the output of a random forest (RF) classifier and a hard threshold of nDSM and NDVI.

The RF classifier uses as input Gabor features [13] extracted from the spectrally adjusted RGB images, together with the multispectral images, DSM, nDSM, and DSM quality indicators. A high vegetation class probability map was then generated and a threshold ($T = 0.7$) selected to produce the binary masks for each image. An additional overlay of trees for the images in Jacksonville has been performed by selecting all objects higher than 4 meters with an NDVI larger than 0.5. Finally, all pixels belonging to the class underwent two cycles of morphological opening and closing, using as structuring element a disk of radius 2 pixels.

2.6. DSM Refinement

Due to the multi-temporal input data, both ground and top of canopy heights are retrieved, especially for deciduous vegetation. The multi-stereo DSM is based on the height cluster supported by most stereo pairs but this often represents the ground, not deciduous high vegetation. As the LiDAR reference data contains mostly top of canopy heights, the highest height cluster supported by at least two stereo pairs is used for pixels classified as high vegetation. The LiDAR ground truth was acquired several years before the satellite imagery, leading to systematic vegetation height differences for Jacksonville. The high vegetation growth between the two acquisitions was compensated by subtracting 65 cm in the DEMs from the pixels labeled as high vegetation in Jacksonville.

3. RESULTS

Each entry was evaluated based on the mean intersection over union filtered by the height error (mIoU-3), where only pixels with a DSM height error of less than 1 meter counted as true positive. Additionally, per class IoU and IoU-3 values as well as Z accuracy an completeness of the DSM were provided by the evaluation server.

Table 2 report DSM statistics for the different processing options. The cluster based merging leads to significant improvement in height accuracy, and improves IoU-3 in the ground and building classes by 0.5% and 0.4%.

The impact of different classification and post-processing steps is shown in Table 1. The results show that the basic CNN Ensemble and water masking without much post-processing would have been sufficient to win the contest, but the cluster based height merging, water mask, high vegetation and elevated roads post-processing lead to a further improvement of 1%.

4. CONCLUSIONS

The final contest results shows that, while CNNs are indispensable for high quality semantic segmentation, they still can be improved by traditional methods for specific tasks such

DSM Fusion	Semantic Segmentation	mIOU-3	mIOU	Ground	High Vegetation	Building	Water	Bridges
Median-Fusion	UNet + WM	0.718	0.782	0.819	0.509	0.809	0.949	0.823
Median-Fusion	CNN-Ensemble + WM	0.736	0.798	0.827	0.564	0.803	0.953	0.843
Cluster-Fusion	All	0.746	0.806	0.831	0.571	0.814	0.958	0.855

Table 1: Evaluation scores of different classifiers and DSM combinations. The first row used a basic median fusion for the DSM generation and the baseline U-Net and NDWI based water detection without morphological refinement. The second row reports the result of the CNN-Ensemble described in Sect. 2.4, but without classification refinement, cf. Sec. 2.5. The last row shows the results of the complete process.

Method	Postproc.	Height accuracy	Completeness
Median		0.411	0.654
Median	VegHeight	0.408	0.658
Cluster		0.356	0.671
Cluster	VegHeight	0.355	0.675

Table 2: Height statistics for different DSM fusion and post-processing algorithms. The cluster based algorithm performs better than median while the systematic vegetation height difference correction only has a small impact.

as water detection. For DSM generation from multi-view data, classical non-deep learning approaches based on SGM were used by the top 3 entries, indicating that more work needs to be done on CNN based stereo algorithms to reach the accuracy of traditional methods, especially when many stereo pairs are available. While our work includes some integration between semantic segmentation and DSM generation in the form of using height information in the multi-modal fusion network, and semantic segmentation results during DSM merging, future work could further benefit from a tighter integration of both semantic segmentation and stereo matching. Our entry won the competition by a margin of 1.46%, out of which 1% was the result of post-processing aimed at improving the semantic segmentation on large buildings, bridges, and high vegetation. While the final semantic segmentation score of the top two teams was quite similar, the better DSM due to bundle adjustment and mature implementation of SGM lead to a final difference of 1.46% in mIoU-3.

5. ACKNOWLEDGEMENT

The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest.

6. REFERENCES

[1] B. Le Saux, N. Yokoya, R. Hänsch, M. Brown, and G. Hager, "2019 data fusion contest [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 103–105, March 2019.

[2] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Z. Brown, "Semantic stereo for incidental satellite images," *CoRR*, vol. abs/1811.08739, 2018.

[3] P. d'Angelo, "Automatic orientation of large multitemporal satellite image blocks," in *International Symposium on Satellite Mapping Technology and Application*, November 2013, pp. 1–6.

[4] P. d'Angelo and G. Kuschk, "Dense multi-view stereo from satellite imagery," in *IGARSS 2012*. November 2012, number DOI: 10.1109/IGARSS.2012.6352565, pp. 6944–6947, IEEE Press.

[5] R. Perko, H. Raggam, K. Gutjahr, and M. Schardt, "Advanced DTM generation from very high resolution satellite stereo images," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015, number Volume II-3/W4.

[6] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[7] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz, "Aerial lanenet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, 2018.

[8] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[9] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3309–3318.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[11] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[12] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," *Remote sensing of environment*, vol. 44, no. 2-3, pp. 145–163, 1993.

[13] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 8, pp. 837–842, 1996.