SEMANTIC SEGMENTATION OF AIRBORNE IMAGES AND CORRESPONDING DIGITAL SURFACE MODELS – ADDITIONAL INPUT DATA OR ADDITIONAL TASK?

Matthias Schmitz*, Wolfgang Brandenburger, Helmut Mayer

Institute for Applied Computer Science, Bundeswehr University Munich, Neubiberg, Germany
-(matthias.schmitz, w.brandenburger, helmut.mayer)@unibw.de

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: Convolutional Network, Semantic Segmentation, Multi-Task Learning, Height Estimation

ABSTRACT:

We analyze the effects of additional height data for semantic segmentation of aerial images with a convolutional encoder-decoder network. Besides a merely image-based semantic segmentation, we trained the same network with height as additional input and furthermore, we defined a multi-task model, where we trained the network to estimate the relative height of objects in parallel to semantic segmentation on the image data only. Our findings are, that excellent results are possible for image data only and additional height information has no significant effect – neither when employed as extra input nor when used for multi-task training, even with differently weighted losses. Based on our results, we, thus, hypothesize that a strong encoder-decoder network implicitly learns the correlation of object categories and relative heights.

1. INTRODUCTION

Semantic segmentation, i.e., pixel-wise classification, using Convolutional Networks (ConvNets) has been shown to produce very good results. Many different approaches have been published in recent years, starting from adapting well-known architectures for image classification and fine-tuning them for semantic segmentation (Long et al., 2015), up to specific architectures that are trained directly for semantic segmentation without any pre-training (Jégou et al., 2017).

Other authors have presented a fusion model for semantic segmentation of aerial images, combining image data as well as height data in a single ConvNet (Zhang et al., 2017). In their experiments, they examined the influence of the height data when fusing it with the image data at different depths of the network based on the sensitivity for single classes.

Yet other authors have dealt with multi-task learning (Kendall et al., 2018), defining a ConvNet treating three tasks in parallel: semantic segmentation, instance segmentation and depth estimation from single images of road scenes.

Schmitz et al. (2019) have examined the behavior of different strategies for training a ConvNet on multiple similar datasets on the example of facade segmentation.

In this paper, we analyze how a ConvNet for semantic segmentation is affected when height data in the form of a (standardized) Digital Surface Model (DSM) is used additionally to image data (here: IR-R-G) for training. Contrary to Zhang et al. (2017), we do not examine the effects when additional height data is fused at different depths within the ConvNet, but when it is either used as additional input data or as an additional task. We employ FC-DenseNet56 (Jégou et al., 2017) as architecture for our experiments and train it as baseline directly with image data only on the Vaihingen 2D Semantic Labeling dataset (Cramer, 2010). Two modifications of the network architecture, depending on the usage of the height data, are the focus of our analysis reported in this paper.

First, we include the height as additional input data. I.e., we concatenate the image and the height to a 4-channel IR-R-G-H vector. The ConvNet is trained with this data and we analyze the effect of the additional height information regarding the overall accuracy (ratio of correctly classified pixels to all pixels).

Second, we use the DSM data for height regression in an additional task parallel to semantic segmentation. I.e., we define a multi-task learning problem and train both, semantic segmentation as well as height regression, in parallel on the same input data. This is realized in a way where both tasks share the main part of the network, but use independent parts for semantic segmentation and height regression. The assumption behind this is that both tasks are highly correlated, meaning that training one of them affects the other and vise versa. As it is hardly possible to estimate the absolute height of a scene from an image patch, we standardize (subtracting mean, dividing by the standard deviation) the DSM input to train the network on relative height differences.

In Figure 1 we present an abstract version of our network architecture with optional height information. Please notice that the height data (red dashed parts) is either used at the beginning or at the end of the network, depending on whether height is employed as additional input data or for height regression as an additional task.

The paper is organized as follows: After discussing related work we present our methodology in Section 3 and the experiments in Section 4. Section 5 gives the results and discusses them before the paper ends with the conclusion.

2. RELATED WORK

ConvNets outperform traditional approaches especially in image-level classification (Krizhevsky et al., 2012) and pixel-level semantic segmentation (Long et al., 2015). A review of

*Corresponding author

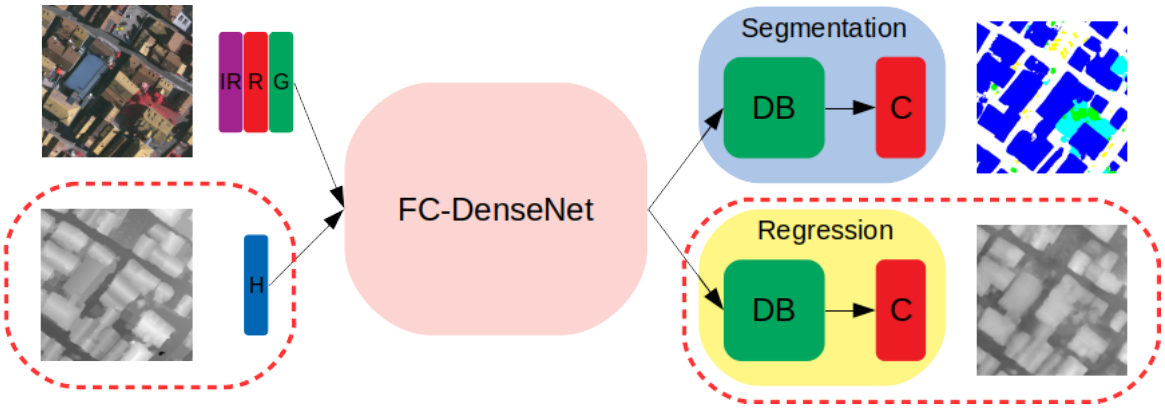

Figure 1. The architecture and its modifications. The red dashed parts are optional in an exclusive-or manner. Input channels: *IR-R-G*: infra-red, red and green, *H*: height. *DB*: dense block and *C*: convolutional layer. *FC-DenseNet* from (Jégou et al., 2017).

deep learning methods for semantic segmentation can be found in (Garcia-Garcia et al., 2017).

ConvNets have been successfully applied in remote sensing, e.g., for classification tasks like semantic labeling (Mnih and Hinton, 2010; Maggiori et al., 2017; Marmanis et al., 2018) as well as landcover mapping (Chen et al., 2014; Kussul et al., 2017). An overview of recent advances and an analysis of challenges of deep learning for remote sensing is given in (Zhu et al., 2017).

Besides for classification, deep neural network-based architectures can be used for regression to predict continuous variables, e.g., the depth of each pixel in a single (monocular) image of a scene (Eigen et al., 2014). Depth estimates and surface normals are together with semantic labels an important component for the understanding of geometric relations within a scene. The exploitation of these related but different factors describing the scene leads to essential improvements in recognition tasks (Silberman et al., 2012) as well as in remote sensing applications like classification of topographic objects and change detection. Monocular depth estimation is equivalent to height prediction in optical remote sensing images and, thus, an inherently ambiguous problem. In spite of this (Mou and Zhu, 2018; Ghamisi and Yokoya, 2018; Amirkolaee and Arefi, 2019) apply ConvNets to predict height values in single aerial images and use the resulting height maps as additional input data to significantly improve the classification accuracy for semantic segmentation.

The above discussion implies that semantic segmentation and depth estimation in monocular images are related tasks. Contextual information about the scene, the perspective and relations between objects are essential when depth is to be estimated in single images. The accuracy of semantic labeling may in turn benefit from the depth, when an additional objective acts as a regularizer during training, preserving object boundaries. A joint estimation in a multi-task deep learning model of these objectives leads to mutual improvements in accuracy and training time (Eigen and Fergus, 2015). An appropriate loss function is assigned to each task of the model and generates different target data for the branches. However, the loss functions are aggregated as linear combination without considering the different physical meaning of image intensity and depth.

Building upon this, Kendall et al. (2018) introduce a global weighting of the differently behaving loss functions based on the task-specific uncertainty. A convolutional encoder-decoder network is employed, enabling simultaneous semantic segmentation, instance segmentation and depth estimation for each pixel in a single RGB image. Opposed to this, Jiao et al. (2018) propose semantic and depth-aware objectives in a multi-task learning framework, estimating class labels as well as depth values for each pixel to take into account the imbalanced depth and category distributions of existing datasets. Both approaches significantly improve the accuracy of scene understanding tasks and show the necessity of adequately defined objectives of a multi-task model.

In their pioneering work, Srivastava et al. (2017) use a multi-task ConvNet, learning both, height prediction as well as semantic labeling for each pixel in aerial images. However, the multi-task training loss is defined as a weighted combination of the task-dependent loss functions. Hence, the estimated semantic labels and height maps do not differ significantly from those obtained by single-task models. Substantial improvements exploiting a multi-task model for semantic segmentation are shown in (Bischke et al., 2017). It preserves semantic segmentation boundaries in aerial images by introducing a globally weighted multi-task loss.

3. METHODOLOGY

The network architecture we have chosen for our experiments, FC-DenseNet56 (Jégou et al., 2017), is an extension of DenseNet (Huang et al., 2017) for semantic segmentation. It is composed of 56 layers in total, organized as an initial convolution, 11 dense blocks (5 followed by a transition down operation, a bottleneck block and another 5 headed by a transition up operation) and a final convolutional layer. Shallow and deep layers are directly connected by shortcuts to take low-level information, e.g., edges, into account for segmentation. DenseNet is defined by a set of dense blocks, in which already extracted features are directly used as input for all subsequent layers within the block. Training such a network is more efficient and also possible with smaller amounts of training data, because there is no need for a repeated encoding of information and due to a relatively small number of parameters.

As the data we use for our experiments consists of an infrared, a red and a green channel (and additionally of height data for one group of experiments) and common datasets usually contain RGB-images, pre-training the networks would be difficult. However, utilizing FC-DenseNet56 allows to train the networks directly on the desired data, without any pre-training.

We have defined three different settings to analyze the influence of additional height data, which are introduced in the following three sections.

3.1 Semantic Segmentation with Image Data

The baseline setting that we employ as reference for our experiments is semantic segmentation on the image data only. No height information is included and the network is trained on a 3-channel input on a single task.

3.2 Semantic Segmentation with Image Data and DSM

As an additional setting, the same network is trained on a 4-channel input, concatenating image data and height information. Except for increasing the input dimension from 3 to 4, the network is not changed, i.e., the number of layers and kernels is identical. Our goal is to determine if additional input data improves the results for semantic segmentation. A big drawback with this setting is that height data is necessary for producing semantic segmentation not only during training but also at inference.

3.3 Semantic Segmentation and DSM Estimation from Image Data

Finally, we define a setting where the image data alone serves as input and the DSM data is utilized for the additional task of height estimation. Both tasks, semantic segmentation and height estimation, are trained in parallel with independent, task-specific loss functions. The idea behind this is that the semantic class of an object is highly correlated to its (relative) height and training a network on these tasks in parallel converges faster and produces more stable results due to the mutual support. While most of the network is shared by the individual tasks, we decided to make the final convolutional layer as well as the last dense block independent for each task. The total loss is defined by the weighted sum of the individual losses. Details are specified in Section 4.

Training on an additional task, thus, acts like a regularization. Another benefit is that height information is only necessary for training, while at inference time the semantic segmentation can be computed using the image data only. Furthermore, a rough relative height estimate will be available for the input images.

4. EXPERIMENTS

4.1 Datasets

We trained and tested all networks on the ISPRS 2D semantic labeling – Vaihingen dataset. It consists of 33 True Orthophoto Mosaics (TOM), generated from aerial images (3 channels: infrared, red and green – IR-R-G) with corresponding DSMs and dense ground-truth annotations. Semantic classes are *impervious surface, building, low vegetation, tree, car* and *clutter/background*. The data has a ground resolution of 9 cm and is split into 16 training and 17 test images.

4.2 Training

For training, we randomly cropped patches with 448×448 pixels from the images and scaled them independently in both dimensions in the range $[0.95, 1.1]$. The image and the DSM were interpolated bilinearly while for the annotation we applied nearest neighbor interpolation. Furthermore, we augmented the patches by random rotation by $k \times 90^\circ$ ($k \in [0, 1, 2, 3]$) as well as random flipping, leading to eight different states.

Depending on the setting, optimization was performed by minimization of the cross-entropy loss for semantic segmentation or the combined loss, consisting of the weighted sum of cross-entropy loss for semantic segmentation and mean squared error loss for height regression. Adam (Kingma and Ba, 2014) with decoupled weight decay (Loshchilov and Hutter, 2017) was utilized as optimizer. Each network was trained for 150k iterations with a batchsize of 2.

When height data was included, the DSM patches were standardized to ensure that only the relative height was taken into account. As normalized DSM (nDSM) usually mean the "height above ground", we denote the standardized patches as sDSM:

$$sDSM_{x,y} = \frac{DSM_{x,y} - \mu_{DSM}}{\sigma_{DSM}}, \tag{1}$$

with μ_{DSM} the mean height and σ_{DSM} the standard deviation of the current patch.

4.3 Experiments

For a deeper analysis we trained each network with different weight decays: 10^{-4}, 10^{-5}, 10^{-6} and 0. The multi-task version additionally was trained with different weights for the task-specific losses (0.5:0.5, ... ,0.9:0.1).

Training time was about 14 hours for the single-task experiments (only a minimal difference between 3- and 4-channel input) and about 19 hours for the multi-task setting. Each network was trained on a single Nvidia GTX 1080 Ti.

5. RESULTS AND DISCUSSION

Table 1 presents the overall accuracy (the ratio of correctly classified pixels to all pixels) for semantic segmentation with different settings. The Vaihingen dataset provides ground-truth annotation with eroded object boundaries, to reduce the impact of uncertain border definitions. Pixels under eroded areas were ignored for evaluation.

Weight decay seems to have no notable effect with these settings and data. Furthermore, no setting clearly outperforms the others. However, it is surprising that the worst results come from the setting including height data as input, even if there is no big difference.

In contrast to Ghamisi and Yokoya (2018) and Amirkolaee and Arefi (2019) who use ConvNets in single-task settings to estimate height maps from single images which are then used as additional input for semantic segmentation, we cannot observe any significant improvement in terms of accuracy even when using ground-truth DSMs. In their studies, they observed positive effects of additional height information in semantic labeling, even when using estimated heights. The estimated height acts as a coarse global prediction (Eigen et al., 2014) and leads to an implicit refinement of the semantic segmentation. However, our findings do not support their observations as the accuracies of our baseline settings are slightly better than of the experiments with IR-R-G-H input.

wd	IR-R-G	IR-R-G-H	Multi-0.5	Multi-0.6	Multi-0.7	Multi-0.8	Multi-0.9
0	**88.83**	85.73	88.29	88.82	88.74	88.21	88.59
10^{-4}	88.09	87.23	87.13	88.49	87.77	88.55	86.85
10^{-5}	88.49	85.3	88.37	88.65	88.32	88.41	88.41
10^{-6}	88.09	88.23	87.56	88.37	88.43	88.26	88.14

Table 1. Overall test accuracy for semantic segmentation with different settings. 0.x in Multi-0.x denotes the weight applied to the semantic segmentation loss. For regression, the corresponding weight is (1 - 0.x).

Changing the weights of the individual losses in the multi-task models also has no clear effect on the result. Moreover, comparing the multi-task results to those from the baseline setting, no regularization on the semantic segmentation task is observable. The same observation has been made in (Srivastava et al., 2017).

On the other hand, Kendall et al. (2018) have shown, that a multi-task model with a shared encoder and separate decoders as well as dynamic weights for the specific tasks, based on their individual uncertainty, can lead to better results for each task, including semantic segmentation. A multi-task network with deeper independent parts might give better results, but would suffer from higher memory consumption and from higher computational complexity as more parameters have to be optimized.

Figure 2 gives results for each setting for the same tile. All versions fail to classify greater parts of lower flat roofs between higher buildings correctly, like in the upper left corner or near the big building in the center with the gray roof.

Some missclassifications are plausible, like not annotated but found cars in shadows and vice versa, trees/bushes annotated as low vegetation or wrongly classified railroads, which are annotated inconsistently in the ground-truth data. Due to the TOM projection, edges of segmented objects are not straight while the annotation always is.

In Figure 3, the color information from the input patch (top left) is mapped to 3D points with x, y (pixel position) and (relative) height from ground truth (top center) and our regression (top right), respectively. For a meaningful comparison and visualization, we shifted and rescaled the network's height output using the mean and the standard deviation of the ground-truth data.

The results demonstrate that the multi-task network learns not only semantic context, but also additional knowledge about object categories in the form of relative height. While detailed structures like ridges are imprecise, the rough 3D shapes of individual objects (buildings, cars and trees) are clearly observable.

Taking into account that our baseline results using only image data are comparable to other state-of-the-art methods, our observations from the other experiments lead us to the following hypotheses:

- The baseline model implicitly learns an adequate model of the (relative) height of semantic classes.
- The employed architecture and the individual loss functions do not consider the different physical meaning of intensity and height information.
- Due to the quality of ground-truth data (images as well as DSM and annotations), an upper bound of achievable accuracy may impede significantly better results.

6. CONCLUSION

We have applied FC-DenseNet56 for semantic segmentation of aerial images with and without additional height data and achieved state-of-the-art results regarding pixel-wise classification accuracy. The use of height data in addition to image data as well as the definition of a multi-task model, trained on semantic segmentation as well as height regression in parallel on single images, does not significantly affect the results.

Thus, we have come to the conclusion that ConvNets are able to implicitly learn the (relative) height of objects. This could be the reason for no significant differences in the versions.

For future work, we want to analyze how advanced loss functions and their adaptive weighting can improve results on semantic segmentation and on height estimation. Especially for the latter, we are interested in how one could include contextual information like shadows or the visible parts of facades in aerial images.

ACKNOWLEDGMENT

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [Cramer, 2010]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

REFERENCES

Amirkolaee, H. A., Arefi, H., 2019. Height Estimation from Single Aerial Images using a Deep Convolutional Encoder-Decoder Network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, 50–66.

Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A., 2017. Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks. *arXiv preprint arXiv:1709.05932*.

Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y., 2014. Deep Learning-based Classification of Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7, 2094–2107.

Cramer, M., 2010. The DGPF-Test on Digital Airborne Camera Evaluation – Overview and Test Design. *Photogrammetrie – Fernerkundung – Geoinformation*, 2010, 73–82.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *IEEE International Conference on Computer Vision*, 2650–2658.

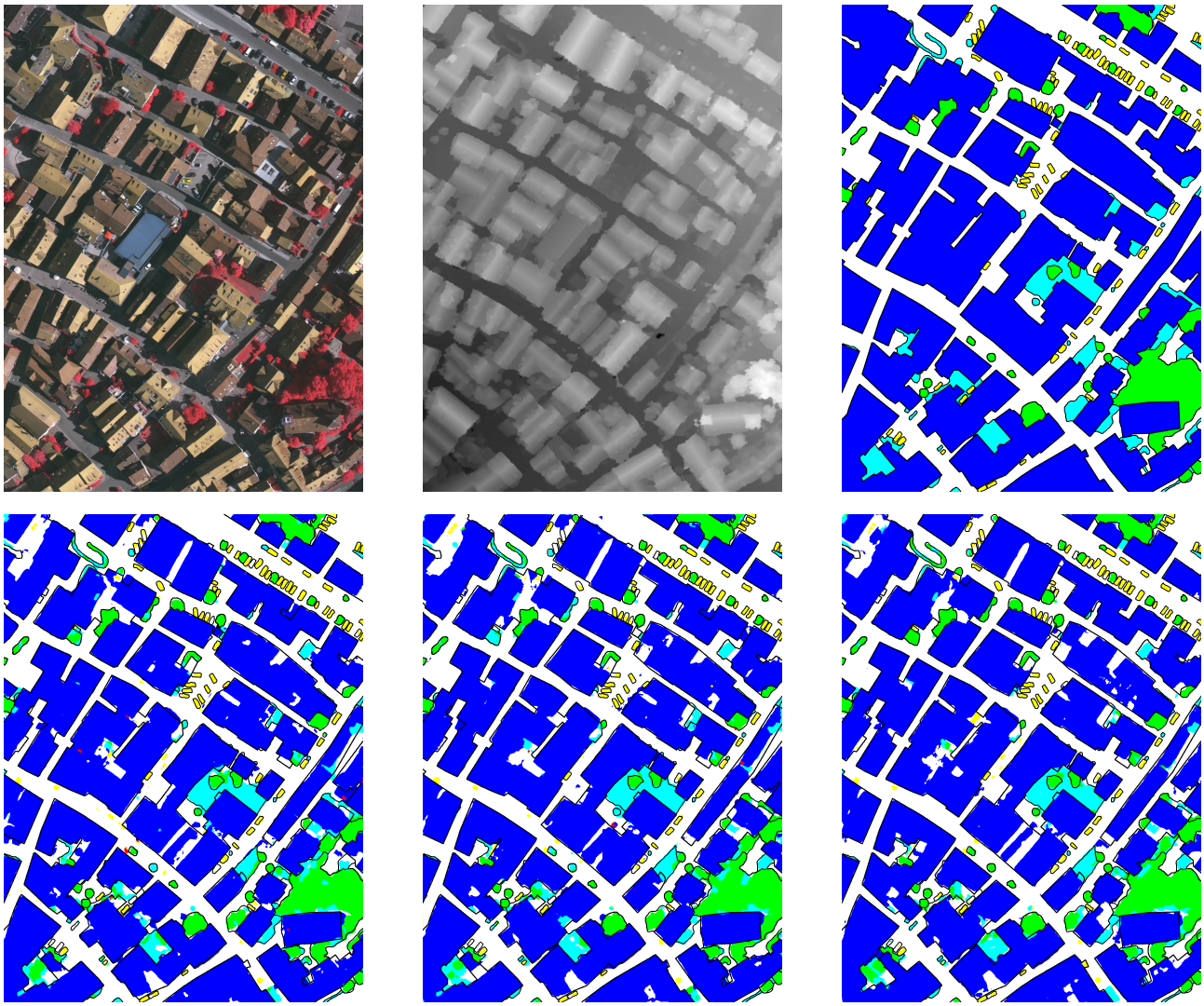

Figure 2. Top: image, DSM and ground-truth annotation. Bottom: results from single-task with image data, single-task with image and height data and from multi-task setting.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 2366–2374.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv preprint arXiv:1704.06857*.

Ghamisi, P., Yokoya, N., 2018. Img2dsm: Height Simulation from Single Imagery using Conditional Generative Adversarial Nets. *IEEE Geoscience and Remote Sensing Letters*, 15, 794–798.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q., 2017. Densely Connected Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2261–2269.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The One Hundred Layers Tiramisu: Fully Convolutional Densenets for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1175–1183.

Jiao, J., Cao, Y., Song, Y., Lau, R., 2018. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. *European Conference on Computer Vision*, 53–69.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *IEEE Conference on Computer Vision and Pattern Recognition*, 7482–7491.

Kingma, D. P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 1097–1105.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep Learning Classification of Land Cover and Crop Types using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14, 778–782.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. *IEEE Confer-*

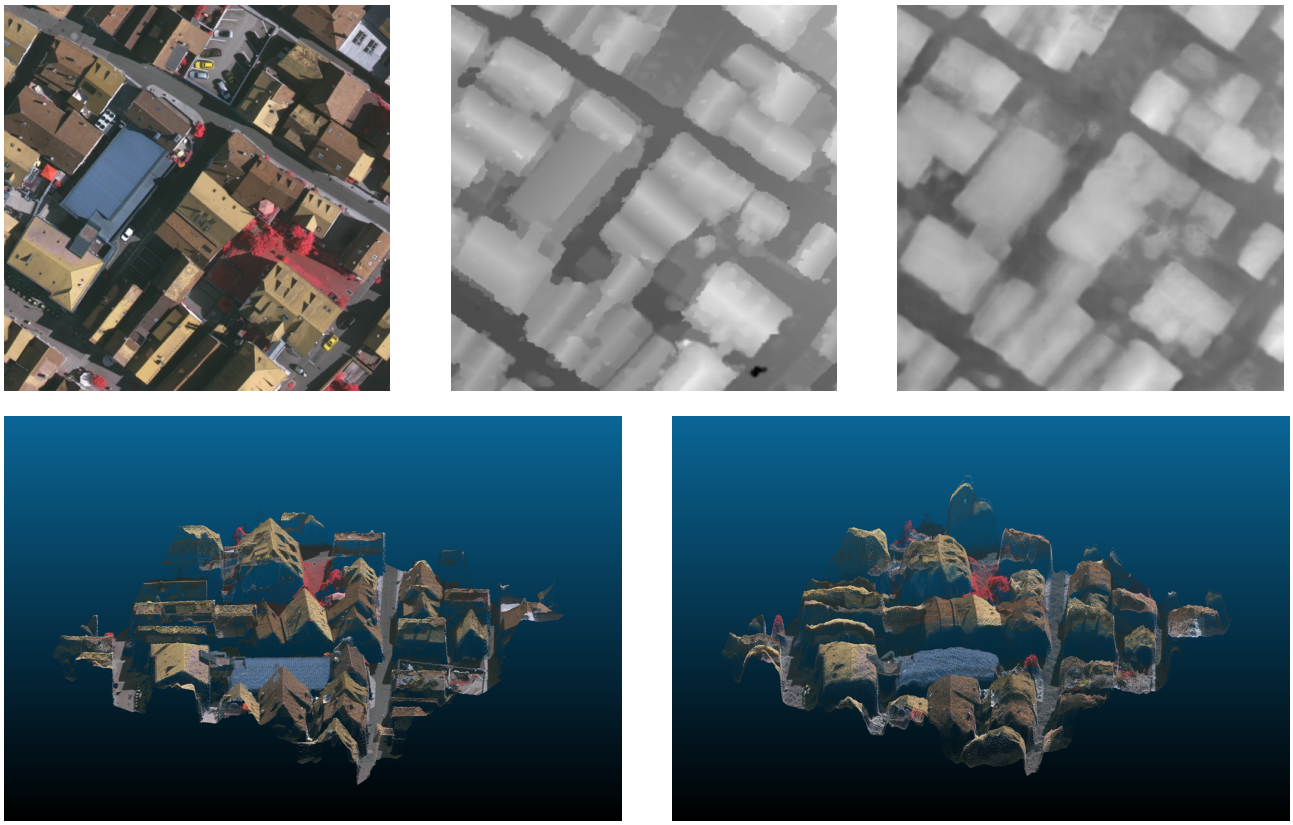

Figure 3. Color information from the image (top left), mapped to a point cloud with x, y (pixel position) and (relative) height from ground truth (top center) and our regression (top right). A screenshot from each pointcloud is shown in the bottom row: ground-truth left and our regression right.

ence on Computer Vision and Pattern Recognition, 3431–3440.

Loshchilov, I., Hutter, F., 2017. Decoupled Weight Decay Regularization. *arXiv e-prints*, arXiv:1711.05101.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *IEEE International Geoscience and Remote Sensing Symposium*, IEEE.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 158–172.

Mnih, V., Hinton, G. E., 2010. Learning to detect roads in high-resolution aerial images. *European Conference on Computer Vision*, 210–223.

Mou, L., Zhu, X., 2018. IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network. *arXiv preprint arXiv:1802.10249*.

Schmitz, M., Huang, H., Mayer, H., 2019. Comparison of Training Strategies for ConvNets on Multiple Similar Datasets for Facade Segmentation. *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W13, 111-117.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. *European Conference on Computer Vision*, 746–760.

Srivastava, S., Volpi, M., Tuia, D., 2017. Joint height estimation and semantic labeling of monocular aerial images with cnns. *IEEE International Geoscience and Remote Sensing Symposium*, 5173–5176.

Zhang, W., Huang, H., Schmitz, M., Sun, X., Wang, H., Mayer, H., 2017. A Multi-Resolution Fusion Model Incorporating Color and Elevation for Semantic Segmentation. *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 513-517.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5, 8–36.