

Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks

Benjamin Bischke^{1,2} Patrick Helber^{1,2} Joachim Folz² Damian Borth² Andreas Dengel^{1,2}

¹University of Kaiserslautern, Germany

²German Research Center for Artificial Intelligence (DFKI), Germany

{Benjamin.Bischke, Patrick.Helber, Joachim.Folz, Damian.Borth, Andreas.Dengel}@dfki.de

Abstract—The increased availability of high resolution satellite imagery allows to sense very detailed structures on the surface of our planet. Access to such information opens up new directions in the analysis of remote sensing imagery. However, at the same time this raises a set of new challenges for existing pixel-based prediction methods, such as semantic segmentation approaches. While deep neural networks have achieved significant advances in the semantic segmentation of high resolution images in the past, most of the existing approaches tend to produce predictions with poor boundaries. **In this paper, we address the problem of preserving semantic segmentation boundaries in high resolution satellite imagery by introducing a new cascaded multi-task loss.** We evaluate our approach on Inria Aerial Image Labeling Dataset which contains large-scale and high resolution images. Our results show that we are able to outperform state-of-the-art methods by 8.3% without any additional post-processing step.

Index Terms—Deep Learning, Semantic Segmentation, Satellite Imagery, Multi Task Learning, Building Extraction

I. INTRODUCTION

THE increasing number of satellites constantly sensing our planet has led to a tremendous amount of data being collected. Recently released datasets such as the EuroSAT [1] and Inria Building Dataset [2] contain images which cover a large surface of our earth including numerous cities. Today, labels employed for land-use classification and points-of-interest detection such as roads, buildings, agriculture areas, forests, etc. are primarily annotated manually. Building upon the recent advances in deep learning, we show how automated approaches can be used to support and reduce such a laborious labeling effort.

In this paper, we focus on the segmentation of building footprints from high resolution satellite imagery. This is of vital importance numerous domains such as Urban Planning, Sociology, and Emergency Response (as observed by the necessity to map buildings in Haiti during the response to hurricane ‘Matthew’ in 2016). The task of automatically segmenting building footprints at a global scale is a challenging task since satellite images often contain deviations depending up on the geographic location. Such deviations are caused by different urban settlements which can be densely or sparsely populated, having different shapes of buildings and varying illuminations due to local atmospheric distortions.

To address the problem of the global variation, Maggiori et. al. [2] created a benchmark database of labeled imagery covering multiple urban landscapes, ranging from highly dense

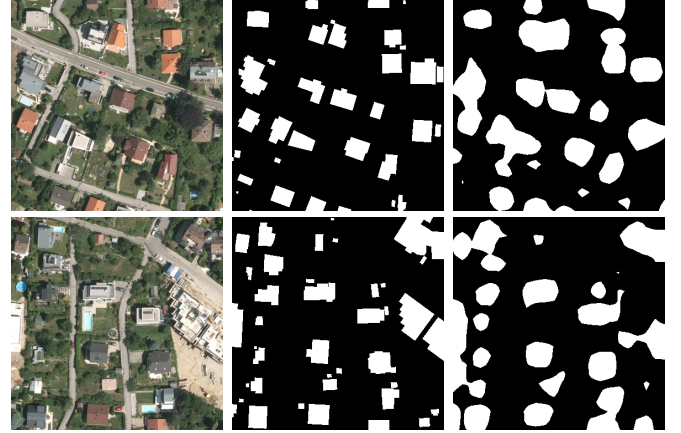


Fig. 1. Two examples of RGB satellite image (left), ground truth masks for building footprints (middle), and corresponding predictions by a FCN network [5] (right). It can be seen that the prediction overlap well with the ground truth but often fail to reflect the boundaries which are present in the ground truth masks.

metropolitan financial districts to alpine resorts. The authors observed that the shape of the building predictions on high resolution images is often rounded and does not exhibit straight boundaries which buildings usually have (see Fig. 3).

The problem of “blobby” predictions has been addressed in the work of Maggiori et. al. [3], [2] where they suggested to train an MLP on top of an FCN to improve the segmentation prediction. Similarly, Marmanis et. al. [4] proposed to combine feature maps from multiple networks at different scales and make the final predictions on top of these concatenated feature maps. One drawback of such approaches is that the model complexity and number of parameters is increased by the different networks. Furthermore, it leads to a very high memory consumption since feature maps at each resolution of the network have to be up-sampled to the full size of the output image.

In this paper, we propose a cascaded multi-task loss along with using a deeper network architecture (than used in [2], [3]), to overcome the problem of “blobby” predictions. Our approach incorporates boundary information of buildings and improves the segmentation results while using less memory at inference time compared to [2]. In this regard, the contributions of this paper can be summarized as follows:

- Focusing on the Inria Aerial Image Labeling Dataset, we

perform detailed experiments with VGG16 as encoder for segmentation networks. We show that features learned by VGG16 are more powerful than the ones extracted from networks proposed in [3], [2]. We additionally evaluate the importance of different decoder architectures.

- We introduce an uncertainty weighted and cascaded multi-task loss based on distance transform to improve semantic segmentation predictions of deep neural networks. Thereby, we achieved increased accuracy for the segmentation of building footprints in remote sensing imagery. Learning with our multi-task loss, we are able to improve the performance of the network by 3.1% without any major changes in the architecture.
- We show that our approach outperforms current state-of-the-art accuracy for the Intersection over Union (IoU) on the validation set significantly by 8.3% without any post-processing step.

II. RELATED WORK

Semantic Segmentation is one of the core challenges in computer vision. Convolutional neural networks such as fully convolutional networks or encoder-decoder based architectures have been successfully applied to this problem and outperformed traditional computer vision approaches marginally. A detailed survey of deep neural network based architectures for semantic segmentation can be found in [6]. One of the main problems when applying CNNs on semantic segmentation tasks is the down-sampling with pooling layers. This increases the field of view of convolutional kernels but loses at the same time high-frequency details in the image. Past work has addressed this issue by reintroducing high frequency details via skip-connections [5], [7], [8], dilated convolutions [9], [10], [11], [12] and expensive post-processing with conditional random fields (CRF's) [9], [11], [13]. While these approaches are able to improve the overall segmentation results, the boundaries between two different semantic classes can often not be well segmented. The importance of segmenting correct semantic boundaries is also considered in recent segmentation datasets such as the *MIT Places Challenge 2017* [14] which evaluates besides the predicted segmentation masks also the accuracy of semantic boundary predictions. Furthermore, it can be observed that more recent network architectures focus on the incorporation of boundary information in the models. This is often achieved on the architecture level by introducing special boundary refinement modules [13], [15], on the fusion level by combining feature maps with boundary predictions [4] or by using a different output representation in the training [16], [17]. Closest to our work are the approach of Hayder et. al. [16] and Yuan [17] which train the network to predict distance classes to object boundaries instead of a segmentation mask. Our work differs from this work, that we additionally predict semantic labels through a multi-task loss and further improve the segmentation results.

In context of remote sensing applications the extraction of building footprints has been extensively studied in the past decade. The problem has been addressed by traditional computer vision techniques using hand-crafted features such

as vegetation indices [18], [19], texture and color features [20], [18], [19] along with traditional machine learning classifiers (e.g., AdaBoost, support vector machines (SVM), random forests (RF)). Often an additional post-processing step is applied to refine the segmentation results [18], [21]. More recent work uses pixel level convolutional neural networks for building detection. Zhang et. al. [22] trained a CNN on Google Earth images and applied an additional post processing step using maximum suppression to remove false buildings. Yuan [17] used a FCN to predict the pixel distance to boundaries and thresholded the predictions to get the final segmentation mask. One of the first approaches which does not rely on an additional post-processing step was proposed by Huang et. al [23]. They trained a deconvolutional network with two parallel streams on RGB and NRG band combinations and fused the predictions of the two streams. A similar two stream network was used by Marmanis et. al. [4] which processed RGB and DEM information in parallel. Similar to our approach, the focus of Marmanis's et. al. [4] work is to preserve boundary information on segmentation classes. This was achieved by first using SegNet as feature extractor (as in our work) and applying additionally an edge detection network (HED) to extract edge information. The boundary predictions are injected into the network by concatenating the feature maps of SegNet with the edge prediction. Our work is different from this approach, that we do not want to extract boundary and semantic information by two different networks and fuse this information at later stages. Our goal is to rather train a single network such that a shared representation for boundary and segmentation prediction can be learned. This reduces the overall complexity of the model and avoids problems such as class-agnostic edge predictions. Recently, Maggiori et. al. [2] showed that previously trained CNN's based on the Massachusetts dataset, generalize poorly to satellite images taken from other cities. Therefore, they released a new large-scale dataset containing high-resolution satellite images. We evaluate our method on this new dataset and compare the results against the best performing methods.

III. CASCADED MULTI-TASK NETWORK

In our approach, we use multi-task learning to improve the segmentation predictions of building footprints. **The goal is to rely besides the semantic term also a geometric term which incorporates the boundary information of the segmentation mask into a single loss function.** We achieve this shared representation of semantic and geometric features by training the network on two different tasks. We train our segmentation network parameterized by θ with a set of training images x along with their ground truth segmentation masks S and corresponding truncated distance class labels D represented by $(x(n), S(n), D(n)); n = 1, 2, \dots, N$. In the following we explain details about the output representation, network architecture and multi-task loss function.

A. Output Representation

The goal of our multi-task approach is to incorporate besides semantic information about class labels also geometric

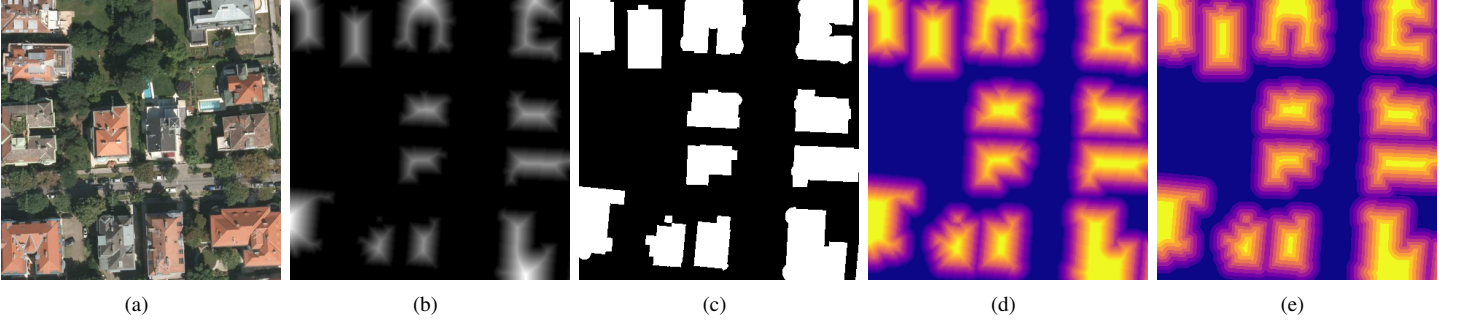


Fig. 2. One example input image with different output representations. (a) Satellite Image (RGB), (b) Semantic Segmentation Mask, (c) Distance Transform, (d) Truncated Distance Mask, (e) Truncated and Quantized Distance Mask (Best viewed in electronic version)

properties in the network training. Although there are multiple geometric properties which can be extracted such as shape and edge information, we extract on the distance of pixels to boundaries of buildings. Such a representation has the advantages that (1) it can be easily derived from existing segmentation masks by means of the distance transform and (2) neural networks can be easily trained with the representation using existing losses like the mean squared error or the negative log likelihood loss. Using the representation, we bias the network to learn per pixel information about the location of the boundary and capture implicitly geometric properties. We truncate the distance at a given threshold to only incorporate the nearest pixels to the border. Let Q denote the set of pixels on the object boundary and C_i the set of pixels belonging to class i . For every pixel p we compute the truncated distance $D(p)$ as

$$D(p) = \delta_p \min(\min_{q \in Q} d(p, q), R),$$

$$\delta_p = \begin{cases} +1 & \text{if } p \in C_{\text{building}} \\ -1 & \text{if } p \notin C_{\text{building}} \end{cases} \quad (1)$$

where $d(p, q)$ is the Euclidean distance between pixels p and q and R the truncation threshold. The pixel distances are additionally weighted by the sign function δ_p to represent whether pixels lie inside or outside the building masks. The continuous distance values are then uniformly quantized to facilitate training. Similar to Hayder et. al. [16] we one-hot encode the distance map into a binary vector representation $b(p)$ as:

$$D(p) \sum_{k=1}^K r_k b_k(p) \sum_{k=1}^K b_k(p) = 1 \quad (2)$$

with r_n as distance value corresponding to bin k . The k resulting binary pixel-wise maps can be understood as classification maps for each of the k th border distance.

B. Encoder-Decoder Network Architecture

The network in this work is based on the fully convolutional network SegNet [7]. SegNet has an encoder-decoder architecture which is commonly used for semantic segmentation. The encoder has the same architecture as VGG16 [24], consists of 13 convolutional layers of 3x3 convolutions and five layers of 2x2 max pooling. The decoder is a mirrored version of

the encoder which uses the pooling indices of the encoder to upsample the feature maps. A detailed illustration of the architecture is shown in Fig. III-B. We add one convolutional layer H_{dist} to the last layer of the decoder to predict the distance to the border of buildings. The final segmentation mask of building footprints is computed by a second convolutional layer H_{seg} . H_{seg} uses the concatenated feature maps of the last decoder layer and the feature maps produced by H_{dist} . Thereby the network can leverage semantic properties present in feature maps of the decoder and the geometric properties extracted by H_{dist} . Please note, that before the concatenation we pass feature maps of H_{dist} additionally through a ReLU. We finally squash the outputs of H_{seg} and H_{dist} through a softmax layer to get the probabilities for the class labels.

C. Uncertainty Based Multi-Task Loss

We define the multi-task loss as follows:

$$L_{\text{total}}(x; \theta) = \sum_{i=1}^T \lambda_i L_i(x; \theta) \quad (3)$$

where T is the number of tasks and L_i the corresponding task loss functions to be minimized with respect to the network parameters θ . Each task loss L_i is weighted by a scalar λ_i to model the importance of each task on the combined loss L_{total} . The weighting terms λ_i in the multi-task loss introduce additional hyper-parameters which are usually equated or found through an expensive grid-search. Motivated by Kendall et. al. [25], we learn the relative task weights λ_i by taking the uncertainty in the model's prediction for each task into consideration. The aim is to learn a relative task weight depending on the confidence of the individual task prediction. Within the context, we define the multi-loss function L_{total} as a combination of two pixel-wise classification losses. We write the total objective as follows:

$$L_{\text{total}}(x; \theta, \sigma_{\text{dist}}, \sigma_{\text{seg}}) = L_{\text{dist}}(x; \theta, \sigma_{\text{dist}}) + L_{\text{seg}}(x; \theta, \sigma_{\text{seg}}) \quad (4)$$

where $L_{\text{dist}}, L_{\text{seg}}$ are the classification loss functions for the prediction of the distance-classes and the segmentation mask with $\sigma_{\text{dist}}, \sigma_{\text{seg}}$ as corresponding task weights for λ_i .

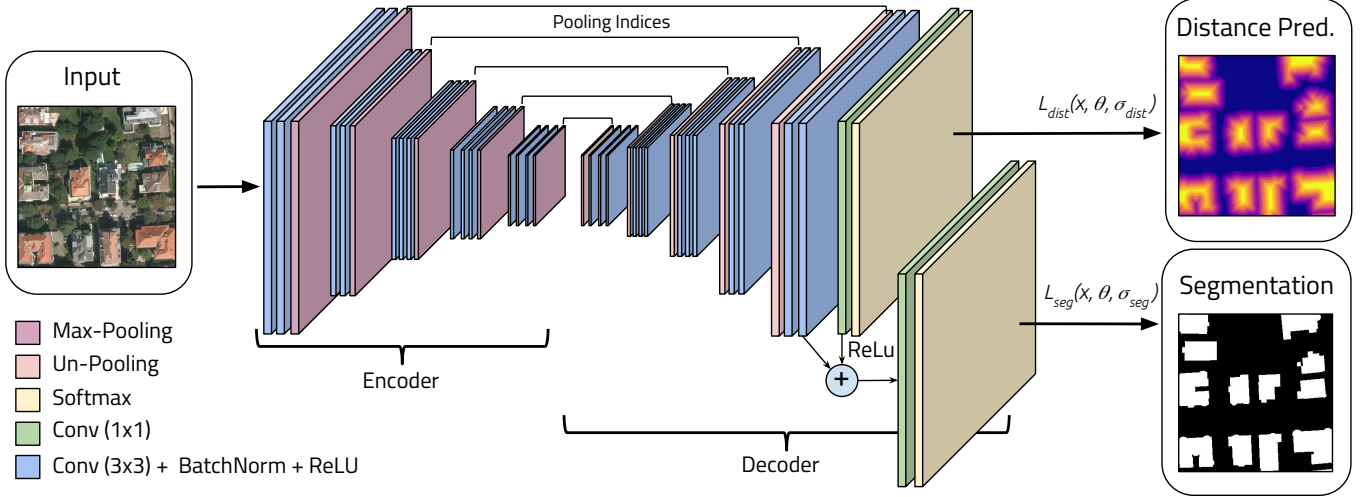


Fig. 3. An illustration of the proposed multi-task cascade architecture for semantic segmentation. The encoder is based on the VGG16 architecture. The decoder upsamples its input using the transferred pooling indices from its encoder to densifies the feature maps with multiple successive convolutional layers. The network uses one convolutional layer H_{dist} after the last decoder layer to predict the distance classes. Feature maps produced by H_{dist} and the last layer of the decoder are concatenated and passed to second convolutional layer H_{seg} to compute the final segmentation masks.

We represent the likelihood of the model for each classification task as a scaled version of the model output $f(x)$ with the uncertainty σ squashed through a softmax function:

$$P(C = 1|x, \theta, \sigma_t) = \frac{\exp(\frac{1}{\sigma_t^2} f_c(x))}{\sum_{c'=1}^C \exp(\frac{1}{\sigma_t^2} f_{c'}(x))} \quad (5)$$

Using the negative log likelihood, we express the classification loss with uncertainty as follows:

$$\begin{aligned} L_t(x, \theta, \sigma_t) &= \sum_{c=1}^C -C_c \log P(C_c = 1|x, \theta, \sigma_t) \\ &= \sum_{c=1}^C -C_c \log \left(\exp(\frac{1}{\sigma_t^2} f_c(x)) \right) + \log \sum_{c'=1}^C \exp(\frac{1}{\sigma_t^2} f_{c'}(x)) \end{aligned} \quad (6)$$

Applying the same assumption as in [25]:

$$\frac{1}{\sigma^2} \sum_{c'} \exp(\frac{1}{\sigma^2} f_{c'}(x)) \approx \left(\sum_{c'} \exp(f_{c'}(x)) \right)^{\frac{1}{\sigma^2}} \quad (7)$$

allows to simplify Eq. 6 to:

$$L_t(x, \theta, \sigma_t) \approx \frac{1}{\sigma_t^2} \sum_{c=1}^C -C_c \log P(C_c = 1|x, \theta) + \log(\sigma_t^2) \quad (8)$$

We use the approximated form of Eq. 8 in both classification tasks L_{dist} and L_{seg} for the prediction of segmentation classes and distance classes respectively. It is important to note, that for numerical stability, we trained the network to predict $\log(\sigma_i^2)$ instead of σ_i^2 . All network parameters and the uncertainty tasks weights are optimized with stochastic gradient descent (SGD).

IV. EXPERIMENTAL RESULTS

A. Inria Aerial Image Labeling Dataset

The Inria Aerial Image Labeling Dataset [2] is comprised of 360 ortho-rectified aerial RGB images at 0.3m spatial

resolution. The satellite scenes have tiles of size 5000 x 5000 px, thus covering a surface of 1500 x 1500m per tile. The images comprise ten cities and an overall area of 810 sq. km. The images convey dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco's financial district) to alpine towns (e.g., Linz in Austrian Tyrol). Ground-truth data is provided for the two semantic classes *building* and *non-building*. The ground truth is only provided for the training set with covers five cities. For comparability, we split the dataset as described by Maggiori et al. [2] (image 1 to 5 of each location for validation, 6 to 36 for training).

B. Evaluation Metrics

We evaluate our approach in the following experiments with two metrics. The first one is the Intersection over Union (IoU) for the positive building class. This is the number of pixels labeled as building in the prediction and the ground truth, divided by the number of pixels labeled as pixel in the prediction or the ground truth. As second metric, we report accuracy, the percentage of correctly classified pixels.

C. Importance of a Deeper Encoder-Decoder-Architecture

In the first experiment, we analyze the importance of the encoder and decoder architecture. Unlike the past work [3], [2], we used the deeper network based on VGG16 [24] as encoder and evaluate different decoder architectures. In our comparison we train the following networks:

- 1) a FCN [5] which uses an up-sampling layer and convolutional layer as decoder
- 2) a SegNet [7] which attaches a reversed VGG16 as decoder to the encoder
- 3) the combination of FCN + MLP as introduced in [2] which up-samples and concatenates all feature maps of the FCN encoder and uses a MLP to reduce the

TABLE I

THIS TABLE SHOWS THE EVALUATION RESULTS OF THE SAME NETWORK USING SINGLE TASK VS. MULTI-TASK. BOTH NETWORK USING THE MULTI-TASK LOSS OUTPERFORM THE SINGLE TASK PREDICTIONS. THE UNCERTAINTY BASED TASK WEIGHTS LEAD TO AN FURTHER IMPROVEMENT AND ACHIEVE OVERALL THE BEST RESULTS.

		Austin	Chicago	Kitsap Co.	West Tyrol	Vienna	Overall
FCN + MLP (Baseline)	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
SegNet (Single-Loss) NLL-Loss for Seg. Classes	IoU	74.81	52.83	68.06	65.68	72.90	70.14
	Acc.	92.52	98.65	97.28	91.36	96.04	95.17
SegNet (Single-Loss) NLL-Loss for Dist. Classes	IoU	76.49	66.77	72.69	66.35	76.25	72.57
	Acc.	93.12	99.24	97.79	91.58	96.55	95.66
SegNet + MultiTask-Loss (Equally Weighted)	IoU	76.22	66.64	71.70	67.03	76.68	72.65
	Acc.	93.03	99.24	97.71	91.66	96.60	95.65
SegNet + MultiTask-Loss (Uncertainty Weighted)	IoU	76.76	67.06	73.30	66.91	76.68	73.00
	Acc.	93.21	99.25	97.84	91.71	96.61	95.73

TABLE II

THIS TABLE LISTS THE PREDICTION ACCURACIES FOR THE SEGMENTATION MASKS IN IV-C ON THE VALIDATION SET. ALL NETWORKS USE THE SAME ENCODER BUT DIFFERENT DECODER TYPES. ALL VGG16-BASED MODELS OUTPERFORM STATE-OF-THE-ART, WHILE SEGNET IMPROVES THE IOU BY MORE THAN 5%.

	mean IoU	Acc. (Pixel)
Baseline FCN [2]	53.82%	92.79 %
Baseline FCN + MLP[2]	64.67%	94.42 %
FCN (VGG16 encoder)	66.21%	94.54 %
FCN + MLP (VGG16 encoder)	68.17%	94.95 %
SegNet (VGG16 encoder)	70.14%	95.17 %

feature maps to class predictions. This approach achieves currently the highest accuracy on the dataset.

Where applicable, we initialize the weights of the encoder with the weights of a VGG16 [24] model pre-trained via ImageNet [26]. All networks are then trained with SGD using a learning rate of 0.01, weight decay of 0.0005 and momentum of 0.9. We use the negative log likelihood loss on the segmentation class labels, reduce the learning rate every 25,000 iterations by the factor 0.1 and stop the training after 200,000 iterations. We extract 10 mini-batches from each satellite image and randomly crop four patches of size 384 x 384 pixels for each mini-batch from the satellite scenes. We apply randomly flipping in vertical and horizontal directions as data augmentation. Table II shows the results of the different architectures on the validation set. The following observations can be made from the table: (1) Due to the deeper architecture of the encoder all architectures outperform previous state-of-the-art approaches on the dataset. This indicates that features learned by the networks proposed in the past were not expressive enough for the segmentation task. (2) When comparing SegNet against the proposed method [2] we do not observe an improvement. This indicates that the SegNet decoder produces better results as compared to the SegNet + MLP combination. (3) The different architectures show that the decoder plays a crucial role for the semantic segmentation task. While the FCN achieves an IoU of 66.21% for the building class, we see an improvement of 3.9% (against the FCN) when using the same encoder but the more complex decoder as in SegNet.

D. Importance of Distance Prediction

In this section, we evaluate the advantage of predicting distance classes to boundaries using a single loss function. As baseline we take SegNet from the previous experiment which was trained with the NLL loss on the semantic segmentation classes and achieved the best results of 70.14% IoU for the building class. We modify this network such that we remove the H_{seg} and attach H_{dist} as shown in III-B. As output representation we use the truncated and quantized distance mask, setting the truncation threshold $R=20$ and the number of bins $K=10$. We train the network as in the previous experiment with SGD and let the network predict distance classes to boundaries. To get the final segmentation mask from the distance predictions, we threshold all distances above five to only get the pixels inside the buildings. The results for this approach are illustrated in Table I and show that by relying on boundary information, we can improve the overall IoU for SegNet by about 2.4%.

E. Importance of Uncertainty Based Multi-Task Learning

In the last experiment, we show the advantage the multi-task loss combining boundary and semantic information to improve the segmentation result. We initialized the network shown in Fig. III-B with the network-weights from the previous experiment and retrain it with the uncertainty based multi-task loss using Eq. 3. The network is trained with SGD using an initial learning rate of 0.001, weight decay of 0.0005 and momentum of 0.9. To evaluate the influence on the uncertainty weights we additionally train the same network using the multi-task loss but set the importance factors λ_i of both tasks to one. The results on Table I illustrate that the uncertainty task loss achieves per location and overall on both evaluation metrics the best results. When using the equally weighted multi-task loss, the overall accuracy is better compared to both single loss tasks but worse than the uncertainty weighted multi task loss. The final prediction results for segmentation masks and distance classes are illustrated in Fig. 4.

V. CONCLUSION

In this paper, we addressed the problem of incorporating geometric information into the internal representation of deep

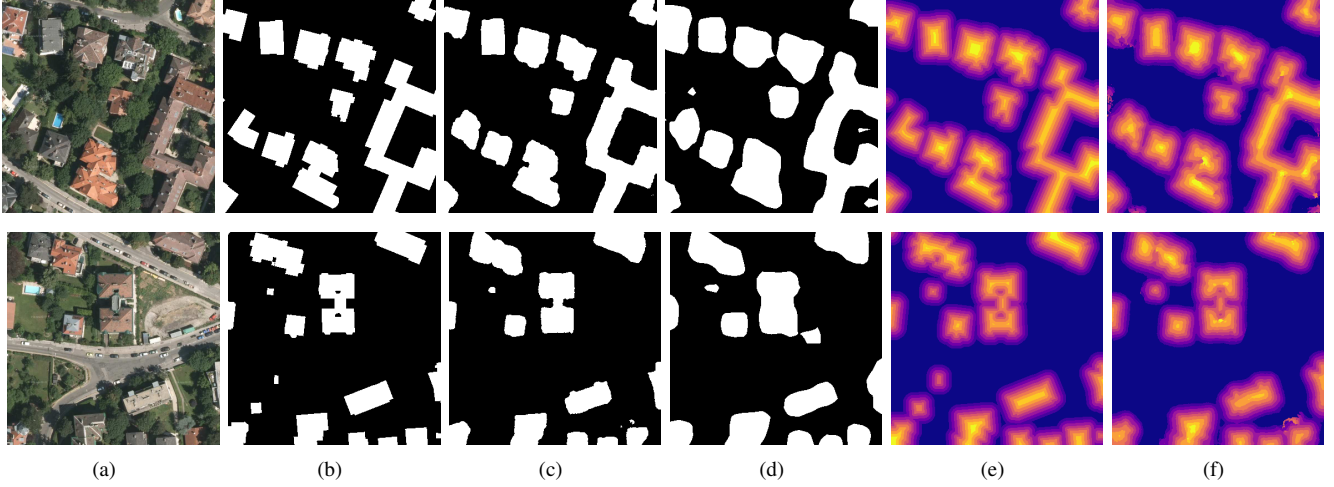


Fig. 4. Two different locations, column-wise: (a) satellite images in RGB, (b) ground truth masks for the building footprints, (c) segmentation predictions by our proposed multi-task network, (d) segmentation predictions by an FCN [5], (e) ground truth masks for distance classes and (f) predicted distance classes. It can be seen that our approach produces less “blobby” predictions with sharper edges compared to the FCN. (Best viewed in electronic version)

neural networks. We therefore focused on semantic segmentation of building footprints from high resolution satellite imagery and showed how boundary information of segmentation masks can be leveraged using a multi-task loss. Our proposed approach outperforms recent methods on the Inria Aerial Image Labeling Dataset significantly by 8.3% which shows the effectiveness of our work. Building upon this work, we plan to extend our multi-task network with further geometric cues and to multiple classes to preserve semantic boundaries. In this context we also plan to make the step from semantic segmentation towards instance segmentation.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA for support within the NVAIL program. Additionally, this work was supported BMBF project MOM (Grant 01IW15002).

REFERENCES

- [1] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” *arXiv preprint arXiv:1709.00029*, 2017.
- [2] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark,” *arXiv preprint arXiv:1409.1556*, 2017.
- [3] Maggiori, Emmanuel and Tarabalka, Yuliya and Charpiat, Guillaume and Alliez, Pierre, “High-Resolution Semantic Labeling with Convolutional Neural Networks,” *arXiv preprint arXiv:1409.1556*, 2016.
- [4] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, “Classification with an edge: improving semantic image segmentation with boundary detection,” *arXiv preprint arXiv:1612.01337*, 2016.
- [5] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.
- [6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” *arXiv preprint arXiv:1704.06857*, 2017.
- [7] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *arXiv preprint arXiv:1511.00561*, 2015.
- [8] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [9] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
- [10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” *arXiv preprint arXiv:1612.01105*, 2016.
- [11] 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,” *arXiv preprint arXiv:1606.00915*, 2016.
- [12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [13] G. Lin, A. Milan, C. Shen, and I. Reid, “Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation,” *arXiv preprint arXiv:1611.06612*, 2016.
- [14] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [15] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters—improve semantic segmentation by global convolutional network,” *arXiv preprint arXiv:1703.02719*, 2017.
- [16] Z. Hayder, X. He, and M. Salzmann, “Boundary-aware instance segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, no. EPFL-CONF-227439, 2017.
- [17] J. Yuan, “Automatic building extraction in aerial scenes using convolutional networks,” *arXiv preprint arXiv:1602.06564*, 2016.
- [18] X. Sun, X. Lin, S. Shen, and Z. Hu, “High-resolution remote sensing data classification over urban areas using random forest ensemble and fully connected conditional random field,” *ISPRS International Journal of Geo-Information*, vol. 6, no. 8, p. 245, 2017.
- [19] Y. Wei, W. Yao, J. Wu, M. Schmitt, and U. Stilla, “Adaboost-based feature relevance assessment in fusing lidar and image data for classification of trees and vehicles in urban scenes,” *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, 7, pp. 323–328, 2012.
- [20] S. Jabari, Y. Zhang, and A. Suliman, “Stereo-based building detection in very high resolution satellite imagery using ihs color system,” in *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*. IEEE, 2014, pp. 2301–2304.
- [21] J. Niemeyer, F. Rottensteiner, and U. Soergel, “Classification of urban lidar data using conditional random field and random forests,” in *Urban Remote Sensing Event (JURSE), 2013 Joint*. IEEE, 2013, pp. 139–142.
- [22] Q. Zhang, Y. Wang, Q. Liu, X. Liu, and W. Wang, “Cnn based suburban building detection using monocular high resolution google earth images,” in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016, pp. 661–664.
- [23] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, “Building extraction from multi-source remote sensing images via deep deconvolutional neural networks,” in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016, pp. 1835–1838.
- [24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

- [25] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” *arXiv preprint arXiv:1705.07115*, 2017.
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.