MULTI-TASK LEARNING FOR SEGMENTATION OF BUILDING FOOTPRINTS WITH NEURAL NETWORKS

Benjamin Bischke, Patrick Helber, Joachim Folz & Andreas Dengel
German Research Center for Artificial Intelligence (DFKI) and TU Kaiserslautern
Kaiserslautern, Germany
{firstname.lastname}@dfki.de

Damian Borth
University of St. Gallen (HSG)
St. Gallen, Swiss
{damian.borth}@unisg.ch

ABSTRACT

The increased availability of high-resolution satellite imagery allows to sense very detailed structures on the surface of our planet and opens up new directions in the analysis of remotely sensed imagery. While deep neural networks have achieved significant advances in semantic segmentation of high-resolution images, 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 novel multi-task loss. The loss leverages multiple output representations of the segmentation mask and biases the network to focus more on pixels near boundaries. We evaluate our approach on the large-scale Inria Aerial Image Labeling Dataset. Our results outperform existing methods with the same architecture by about 3% on the Intersection over Union (IoU) metric without additional post-processing steps. Source code and all models are available under https://github.com/bbischke/MultiTaskBuildingSegmentation.

1 INTRODUCTION

In this paper, we focus on the segmentation of building footprints from high resolution satellite imagery. Building footprints are of vital importance in the context of urban planning and emergency response since they can be used to derive population estimates and household densities. While such information is important for rescuers during a disaster event to better disptach the existing resources, this information can be also used beforehand to mitigate and reduce risk with sustainable planning. Building footprint segmentation at a global scale has therefore direct impact on the UN Sustainable Development Goal 11 (*Sustainable cities and communities*) and Goal 15 (*Life on Land*).

The task of automatically segmenting building footprints at a global scale is challenging since satellite images often contain deviations depending on the geographic location. To address this problem of global variation, Maggiori et al. (2017) created a benchmark database of labeled imagery. The authors observed that the shape of the building predictions on high resolution images is often rounded and does not reveal straight boundaries which buildings usually have. In this paper, we propose a novel multi-task loss 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 Maggiori et al. (2017) or more recent segmentation models such as DeepLabv3 and U-Net that keep feature maps at the full original image resultion in memory to preserve edge information. Since we want to apply our model on a global scale, we use the lightweight SegNet model. In this regard, the contributions of this paper can be summarized as follows:

- We introduce an uncertainty weighted multi-task loss based on the distance transform to improve semantic segmentation predictions of deep neural networks. We show that the combination of related prediction tasks via the proposed loss yields to a higher accuracy compared to networks trained on single tasks only.
- Experiments show that without any major changes in the architecture, same data and same complexity, we are able to improve the IoU metric of building footprints in remote sensed imagery by about 2% compared to networks trained on single task losses.

2 RELATED WORK

Semantic Segmentation is one of the core challenges in computer vision and fully convolutional neural networks or encoder-decoder based architectures have been successfully applied to this problem. One of the main problems when applying CNNs on semantic segmentation tasks is the downsampling 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 Shelhamer et al. (2017); Ronneberger et al. (2015); Badrinarayanan et al. (2015), dilated convolutions Yu & Koltun (2016); Zhao et al. (2017); Chen et al. (2017) and expensive post-processing with conditional random fields Yu & Koltun (2016); Chen et al. (2018); Lin et al. (2017). More recent approaches focus on the incorporation of boundary information in the models. This is often achieved on the architectural level by introducing special boundary refinement modules Lin et al. (2017); Peng et al. (2017), on the fusion level by combining feature maps with boundary predictions Marmanis et al. (2018) or by using a different output representation in the training Hayder et al. (2017); Bai & Urtasun (2017). Closest to our work are the approaches of Hayder et al. (2017) and Yuan (2016) which train the network to predict distance classes to object boundaries instead of a segmentation mask. Our work differs from these approaches, that we additionally predict semantic labels through a multi-task loss and further improve the segmentation results.

In the remote sensing domain, the extraction of building footprints has been extensively studied in the past decade. Yuan (2016) used a FCN to predict the pixel distance to boundaries and thresholded the predictions to get the final segmentation mask. Similar to our approach is the work of Marmanis et al. (2018) which tries to preserve boundary information on segmentation classes. This was achieved by first using SegNet as feature extractor and applying additionally an edge detection network to extract edges. 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.

3 MULTI-TASK LEARNING

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 on an additional term which incorporates the boundary information of the segmentation mask into a single loss function. We achieve this shared representation of semantic and boundary 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 $(x(n), S(n), D(n)); n = 1, 2, ..., N$.

3.1 OUTPUT REPRESENTATION

The goal of our multi-task approach is to incorporate besides semantic information about class labels also geometric 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 this representation, we bias the network to focus more on pixels close to boundaries of building footprints. Compared to other representations such as the Watershed Transform used in Bai & Urtasun (2017), our representation does not only capture boundary information inside the object class but also considers boundary pixels outside the object. 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:

$$D(p) = \delta_p min(\min_{\forall q \in Q} d(p, q), R),$$
$$\delta_p = \begin{cases} +1 & \text{if} \quad p \in C_{building} \\ -1 & \text{if} \quad p \notin C_{building} \end{cases} \tag{1}$$

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. (2017) we one-hot encode the distance map into a binary vector representation $b(p)$ as:

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

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 kth border distance.

3.2 ENODER-DECODER NETWORK ARCHITECTURE

The network in this work is based on the fully convolutional network SegNet. SegNet has an encoder-decoder architecture which is commonly used for semantic segmentation. The encoder has the same architecture as VGG16 Simonyan & Zisserman (2014), consists of 13 convolutional layers of 3x3 convolutions and five layers of 2x2 max pooling. We add one convolutional layer H_{dist} to the last layer of the decoder to predict the distance to the border of buildings and one convolutional layer H_{seg} to predict the segmentation mask. We squash the outputs of H_{seg} and H_{dist} through a softmax layer to get the probabilities for the class labels. Please note, that our approach could also rely on other segmentation models, we used SegNet due to its memory efficiency.

3.3 UNCERTAINTY BASED MULTI-TASK LOSS

We define the multi-task loss as follows:

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

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. (2018), 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 by the network. Within this 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:

$$\begin{aligned} L_{total}(x; \theta, \sigma_{dist}, \sigma_{seg}) = L_{dist}(x; \theta, \sigma_{dist}) \\ + L_{seg}(x; \theta, \sigma_{seg}) \end{aligned} \tag{4}$$

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

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

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

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

Applying the same assumption as in Kendall et al. (2018):

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

allows to simplify Eq. 6 to:

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

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 worth mentioning, that for numerical stability, we trained the network to predict $log(\sigma_i^2)$ instead of σ_i^2.

Table 1: Results for the same network trained on multiple losses: networks using the multi-task loss outperform the ones relying on a single task, uncertainty task-weights achieve overall best results.

		Austin	Chicago	Kitsap Co.	West Tyr.	Vienna	Overall
FCN + MLP	IoU	61.20	61.30	51.50	57.95	72.13	64.67
(Baseline Maggiori et al. (2017))	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
SegNet (Single-Loss)	IoU	69.37	74.08	68.36	67.36	72.69	71.27
NLL-Loss for Seg. Classes	Acc.	95.02	94.80	95.11	98.79	93.55	95.45
SegNet (Single-Loss)	IoU	69.29	75.39	69.56	65.74	73.74	72.05
NLL-Loss for Dist. Classes	Acc.	94.92	95.04	95.21	98.77	93.75	95.53
SegNet (MultiTask-Loss)	IoU	71.62	77.13	71.32	64.24	75.33	73.78
(Equally Weighted)	Acc.	95.63	95.47	95.69	**98.80**	94.28	95.97
SegNet (MultiTask-Loss)	IoU	**72.43**	**77.68**	**72.28**	**64.34**	**76.15**	**74.49**
(Uncertainty Weighted)	Acc.	**95.71**	**95.60**	**95.81**	98.76	**94.48**	**96.07**

4 Experimental Results

4.1 Dataset and Metrics

Our Experiments are based on the large-scale Inria Aerial Image Labeling Dataset Maggiori et al. (2017), which 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. Ground-truth is only provided for the training set which covers five cities in from of segmentation maseks for the two classes *building* and *non-building*. For comparability, we split the dataset as described by Maggiori et al. (2017) (image 1 to 5 of each location for the validation set, 6 to 36 for the training set). We evaluate our approach with the Intersection over Union (IoU) metric 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.

4.2 Importance of uncertainty multi-task learning

In this section, we show the advantage of combining boundary and semantic information via the proposed multi-task loss to improve the segmentation results. We train all networks end-to-end with SGD, the details can be found in the Supplemental Material i. To evaluate the influence on the learned uncertainty weights we additionally train the same network with the multi-task loss but fix the importance factors λ_i of both tasks with one. The results in Table 4 illustrate that (1) the network trained on the uncertainty task loss achieves overall on both evaluation metrics the best results. Location-wise the network achieves also the highest IoU scores and accuracies with the exception of Kitsap Co. and West Tyrol. (2) When training the network with a multi-task loss (equally and uncertainty weighted) the overall accuracy is better compared to both single loss tasks. (3) The equally weighted multi-task loss is slightly worse than the uncertainty one.

5 Conclusion

In this paper, we focused on semantic segmentation of building footprints from high resolution satellite imagery and showed how boundary information of segmentation masks can be leveraged in a multi-task loss in order to overcome the problem of "blobby" predcitions. Without additional parameters or higher network complexity, our model could still achieve an improvement of 3% on the IoU metric against the baseline with resulting segmentation masks that better reflect straight boundaries of building footprints. Building upon this work, we want to incorporate a deeper analysis for the influence of the number of classes and size of the bins of our proposed multi-task approach. We want to make the step from semantic segmentation towards instance segmentation by adding an additional term to the multi-task loss.

Acknowledgments

The authors would like to thank NVIDIA for support within the NVAIL program. Additionally, this work was supported BMBF project DeFuseNN (01IW17002).

REFERENCES

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 2858–2866. IEEE, 2017.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2018.

Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-227439, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 1, pp. 3, 2017.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *arXiv preprint arXiv:1409.1556*, 2017.

Dimitrios Marmanis, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018.

Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters—improve semantic segmentation by global convolutional network. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1743–1751. IEEE, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

Jiangye Yuan. Automatic building extraction in aerial scenes using convolutional networks. *arXiv preprint arXiv:1602.06564*, 2016.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

I SUPPLEMENTAL MATERIAL

In the following, we provide more details on reproducibility and further experiments.

I.1 MOTIVATION

These images show that the segmentation masks overlap quite with the ground truth masks. The boundaries of the predicted building footprints do not reflect straight boundaries.

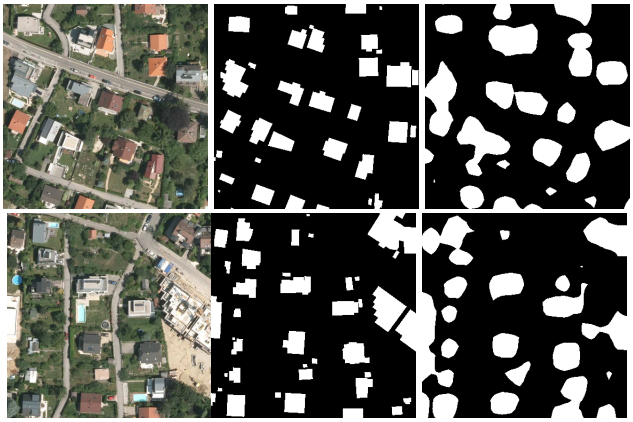


Figure 1: The predictions for building footprints (right) overlap well with the ground truth masks (middle) but often fail to reflect straight boundaries.

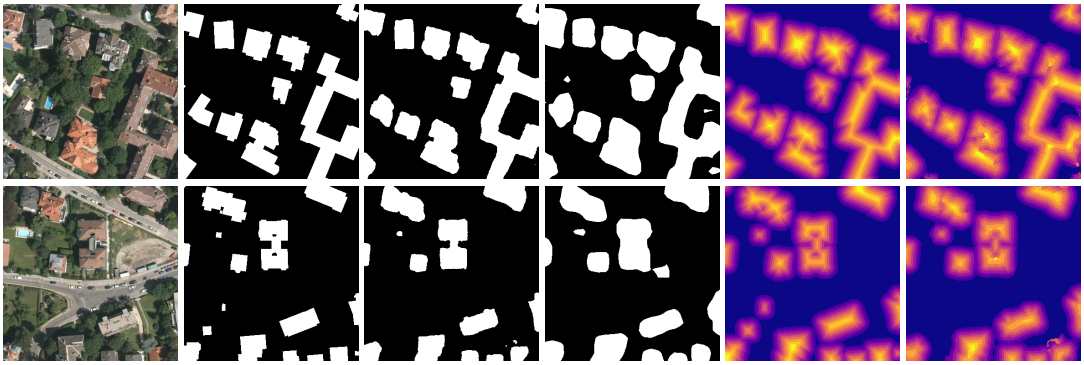
I.2 TRAINING DETAILS

We initialize the weights of the encoder with the weights of a VGG16 Simonyan & Zisserman (2014) model pre-trained via ImageNet Russakovsky et al. (2015). All networks are trained with SGD using a learning rate of 0.01, weight decay of 0.0005 and momentum of 0.9. We use the cross entropy loss on the segmentation class labels reduce the learning rate according to the poly policy Zhao et al. (2017) (p=*0.5*) and stop the training after 30 epochs. We extract 7 batches from each satellite image and randomly crop 24 patches of size 384 x 384 pixels for each batch from the images. We apply randomly flipping in vertical and horizontal directions as data augmentation.

I.3 IMPORTANCE OF THE DISTANCE PREDICTION TASK

We evaluate the advantage of predicting distance classes to boundaries using a single loss function. As baseline we train SegNet with the NLL loss on the semantic segmentation classes and achieve an IoU of 71.27% for the building class. We modify this network such that we remove the H_{seg} and attach H_{dist} as shown in i.5. 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 setup but 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 result of this experiment is illustrated in Table 4. It shows that by training the network on this output representation with border classes the overall IoU is increased by about 1% compared to the one containing only segmentation classes.

I.4 VISUALIZATIONS OF THE NETWORK PREDICTIONS



(a) (b) (c) (d) (e) (f)

Figure 2: 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 a SegNet Badrinarayanan et al. (2015), (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)*

I.5 MULTI-TASK NETWORK ARCHITECTURE

An illustration of the proposed multi-task network 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 a convolutional layer H_{dist} after the last decoder layer to predict distance classes and one layer H_{seg} to predict the segmentation masks.

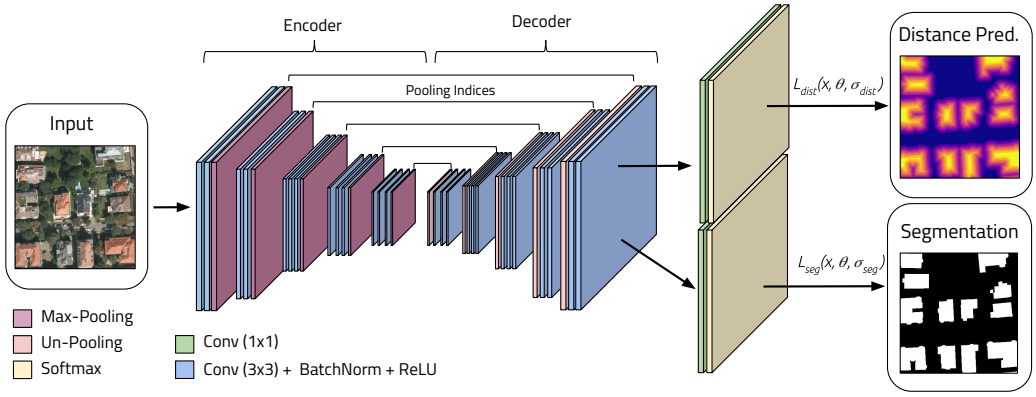


Figure 3: An illustration of the proposed multi-task network architecture for semantic segmentation.