LOCATION-SPECIFIC EMBEDDING LEARNING FOR THE SEMANTIC SEGMENTATION OF BUILDING FOOTPRINTS ON A GLOBAL SCALE

Benjamin Bischke [1, 2] *Patrick Helber* [1, 2] *Jörn Hees* [2] *Andreas Dengel* [1, 2]

[1] TU Kaiserslautern, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Germany

ABSTRACT

In this paper, we analyze the feasability of learning a latent embedding space from aerial and satellite imagery in order to capture semantic properties of geographical locations. We show that deep neural network, trained with a triplet loss function, can be effectively used to obtain a location-specific embedding. Considering the problem of building footprint segmentation from aerial imagery of varying cities, we leverage these embeddings together with a clustering for the training of location-specific segmentation networks and the selection of the corresponding segmentation network during inference time. We evaluate our approach on the large-scale Inria Aerial Image Labeling Dataset which contains aerial images of globally distributed cities. Our approach achieves an outperformance against state-of-the-art approaches on the Intersection over Union metric for the building class over all cities and by more than 2% for specific cities.

Index Terms— Building Segmentation, Embedding Learning, Semantic Segmentation, Deep Neural Networks

1. INTRODUCTION

Satellite data and high resolution aerial imagery is becoming more and more accessible in the recent years and changing the understanding of our planet. One key challenge in this context is the segmentation of aerial imagery into different land-use and land-cover classes in order to extract a meaningful information layer from the raw imagery and support high-level decision making. For example, segmentation of building footprints and the derived layers such as population estimation is an important aspect in urban development, disaster management and for governments. In the recent years, several datasets such as the *Inria Aerial Image Labeling Dataset* [1], *SpaceNet Datasets*[1] and *Kaggle Competitions*[2] have been released to advance research on this particular topic. While there have been a lot of advances in building footprint segmentation with deep convolutional neural networks (CNNs) [2, 3], one challenging problem when applying these approaches on a global scale, is a large number of variations in the images that can be observed at different locations around the globe. Fig. 1 visualizes image patches from ten different cities. These aerial images show that the underlying distributions of building-densities, -types and -sizes can be very diverse. In order to cope with these variations in the images, most state-of-the-art approaches increase the number of layers in deep convolutional networks to have a higher capacity for learning these specific properties. While these approaches are only able to learn the joint distributions over all locations, it is also slowing down the inference time. Additionally, the extension of the network capacity to numerous different locations is limited in practice due to the high number of parameters and size of the model.

Fig. 1. Image patches sampled from different locations of the Inria Aerial Image Labeling Dataset [1]. Images convey location-specific properties such as different rooftop materials, building footprint sizes and building densities (urban vs. rural).

In this paper, we propose a different approach to address the problem of building footprint segmentation from images of very different locations. We rely on deep convolutional neural networks to learn a location-specific embedding of images. We study how these image embeddings can be used to train a set of expert segmentation network. Thereby one particular network is trained on images that depict similar properties of other locations. During inference, we leverage the embedding to select the model that was trained on a most similar data dis-

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the DGX-1 used for this research. This work was partially funded by the BMBF Project DeFuseNN (01IW17002).

[1] https://spacenetchallenge.github.io/

[2] https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection

tribution (e.g. for a new image from the city of New York use a model that was trained on Washington rather than Tyrol). The contributions of this work can be summarized as follows:

- To the best of our knowledge, we show for the first time that aerial images can be embedded into a location-specific latent space using deep neural networks.
- We present an approach that uses these embeddings to train and select location-specific expert segmentation networks in order to cope with high variance in images from different locations.
- We evaluate our approach on the Inria Aerial Image Labeling Dataset and obtain better segmentation results for all cities compared to training with the complete dataset.

2. APPROACH

In this section, we describe our approach for segmenting building footprints from aerial images, that have been collected from different locations with varying data distributions. We first train an embedding network F_e over all images X to learn an location-specific embedding $e_i \in E$ for each image $x_i \in X$. The distance of image embeddings belonging to the same location should be smaller compared to the ones that belong to images of very different locations. We leverage the learned embedding space, to perform clustering and identify the most distinctive locations based on the cluster centroids. We then train for each cluster c a dedicated expert segmentation network S_c with a set of training images x_i and their ground truth segmentation masks $s_i \in S$. More formally, the embedding network F_e learns the mapping from image to embedding space with $F_e : X \rightarrow E$ and the segmentation network is trained to learn the mapping from aerial image x to the segmentation mask s with $S_c : X \rightarrow S$. During inference, we obtain the embedding from F_e, determine the nearest cluster c and select the corresponding expert segmentation network S_c. In the following, we describe the details of the network architectures of F_e and S_c along with the corresponding loss functions to train these networks.

2.1. Location Embedding Network

The objective of the embedding network F_e is to learn a mapping from the input image space to a lower dimensional embedding space. In order to achieve such an image dimensionality reduction, we use an autoencoder network with an encoder-decoder architecture. Through the bottleneck structure in the middle of the architecture, the network is forced to compress the image into a lower dimensional representation without losing important image features. The autoencoder network in this work has a similar architecture than the segmentation network described in the next section, with the exception that we use three output layers and that a ResNet34 [4] is used as the convolutional part of the encoder. We optimize the network with SmoothL1Loss [5] that is less sensitive to outliers than the often used MSELoss.

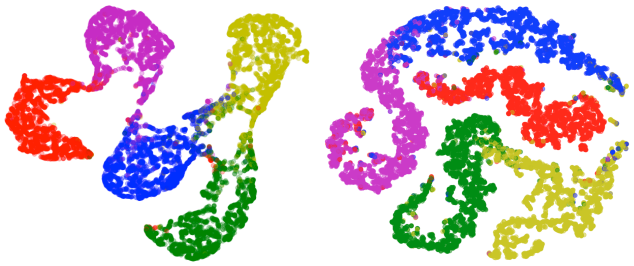

Fig. 2. Visualization of the learned location-specific embedding space using UMAP [6] and T-SNE [7]. The embeddings are color-coded for the corresponding locations *(Austin (red), Chicago (magenta), Kitsap Co.(blue), Vienna (green), West-Tyrol (yellow)*).

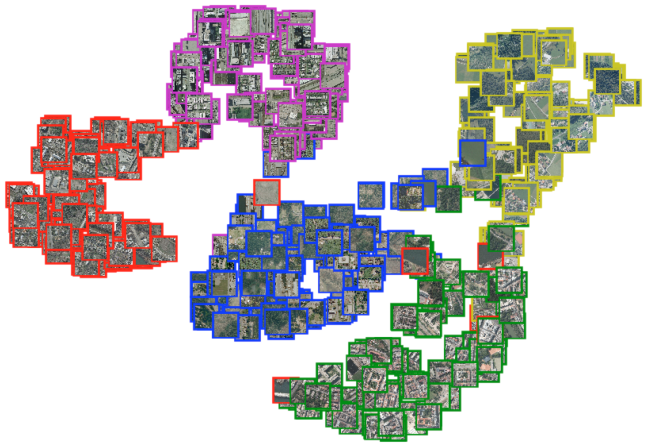

Fig. 3. Visualization of the learned embedding space using UMAP [6] for randomly sampled image patches of different locations.

Since we are not only interested into finding a low-level image representation but we also want to learn an embedding in which images of the same location are closer to each other compared to images from very different locations, we additionally apply an average pooling layer with a 5x5 kernel on the bottleneck layer of the autoencoder and add to two fully connected layers. This results in an 2048 dimensional network output which represents our location-specific image embedding. We use the triplet-marging loss [8] to measure a relative similarity between samples. A triplet is composed by an anchor sample x_i, a positive sample x_p of the same location as x_i and a negative sample x_n from a differnt location than x_i. The triplet-margin loss is defined as follows:

$$L_{triplet}(X) = max(d(y_i, y_p) - d(y_i, y_n) + margin, 0) \quad (1)$$

where we compute the distance between two embedding vectors with the cosine similarity $d(y, y_i) = 1 - cosine(y, y_i)$ The network is trained end-to-end by combining the loss of the autoencoder and the triplet-marging loss on the second output of the network.

Table 1. Evaluation results of our approach on the validation set against two baselines. Please note, that the architecture of the segmentation network is always the same. Our approach was trained on **three** subsets of the dataset, determined using the embedding.

		Austin	Chicago	Kitsap Co.	W. Tyrol	Vienna	Overall
Baseline 1	IoU	72.59	66.83	62.58	70.21	76.07	71.56
No embedding	Acc.	95.90	91.61	99.18	97.55	92.86	95.42
Baseline 2	IoU	71.87	66.74	62.40	70.00	76.38	71.50
Rnd. embedding	Acc.	95.92	91.82	99.19	97.55	93.05	95.51
Our Proposed	IoU	**74.11**	**66.97**	**64.14**	**72.25**	**76.65**	**72.35**
Approach	Acc.	**96.23**	**91.92**	**99.22**	**97.79**	**93.24**	**95.68**

Table 2. Evaluation results of our approach on the validation set against two baselines. Please note, that the architecture of the segmentation network is always the same. Our approach was trained on **five** subsets of the dataset, determined using the embedding.

		Austin	Chicago	Kitsap Co.	W. Tyrol	Vienna	Overall
Baseline 1	IoU	72.59	66.83	62.58	70.21	76.07	71.56
No embedding	Acc.	95.90	91.61	99.18	97.55	92.86	95.42
Baseline 2	IoU	68.07	64.52	51.38	62.10	73.49	67.95
Rnd. embedding	Acc.	95.31	91.09	98.98	96.95	92.06	94.88
Our Proposed	IoU	**75.26**	**66.95**	**63.97**	**74.24**	**76.72**	**72.81**
Approach	Acc.	**96.28**	**91.95**	**99.21**	**97.96**	**93.24**	**95.73**

2.2. Semantic Segmentation Network

The segmentation network S_c that we use in this work is based on the fully convolutional network U-Net [9]. U-Net has an encoder-decoder architecture which is commonly used in semantic segmentation problems. U-Net uses skip connections between blocks of the same spatial size in the encoder and decoder parts to enable a precise localization. Skip connections allow information to flow directly from the low level to high-level feature maps without alternations that even further improve localization accuracy and speed up convergence [9]. As an improvement over the originally proposed VGG16 [10] based U-Net architecture, we replace the encoder with a convolutional part of the Residual network ResNet101[4] that was pre-trained on ImageNet [11]. In the last network layer, we use a 1x1 convolution with one output channel and squash the network output through a *Sigmoid* activation layer.

3. EXPERIMENTS AND RESULTS

3.1. Aerial Image Dataset and Evaluation Metrics

Our approach is evaluated on the Inria Aerial Image Labeling Dataset [1]. This dataset 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 very 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 only provided for the training set which covers five cities is and the two semantic classes *building* and *non-building*. For comparability, we split the dataset in training and validation set as described in [1].

We use the following two metrics to evaluate our approach. The first one is the Intersection over Union (IoU) for the positive building class. This metric 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 ground truth. As the second metric, we report accuracy, the percentage of correctly classified pixels.

3.2. Network Training

We initialize the segmentation networks in this paper with a ResNet101 model pre-trained on ImageNet [11]. All segmentation networks in this paper are trained for 30 epochs with a batch size of 8. We extract image patches of size 512 x 512 pixels and apply random flipping in horizontal and vertical dimension as data augmentation. We use the Adam optimizer with a learning rate of 0.0001 weight decay of 0.0005 to optimize the network parameters. We optimize the network with a loss function that combines the binary cross entropy loss L_{BCE} and the dice loss L_{Dice}:

$$L_{combined} = \alpha L_{BCE} - (1-\alpha) L_{Dice} \tag{2}$$

where L_{Dice} is a version of the Jaccard Index for non-discrete objects that is defined as follows [12]:

$$L_{Dice} = \frac{1}{n} \sum_{c=1}^{2} \sum_{i=1}^{n} \left(\frac{y_i, y_i'}{y_i + y_i' - y_i y_i'} \right) \tag{3}$$

Inspired by [13], we use the "poly" learning rate policy, in which the learning rate determined by $(1 - \frac{iter}{max\_iter})^{power}$. We set the hyperparameter *power* to 0.9. The embedding network is trained according to the same procedure except that ResNet34 pre-trained on ImageNet [11] is used as encoder.

3.3. Experimental Results

3.3.1. Learning and Visualizing Embeddings

In order to evaluate our approach, we train the proposed embedding network F_e over all images in the training set. We then extract location embeddings of all images in the training set with F_e, apply k-means clustering with *k=3* and *k=5*. For each cluster c determine the centroid, construct k training sets for which we train one dedicated expert segmentation network S_k. During test time, we also extract the image embeddings, determine the distance to the nearest cluster centroid and select the corresponding segmentation network from which we obtain the segmentation mask. Visualizations of the embeddings in Fig. 2 and Fig. 3 show that image patches of the same location are closer in the embedding space compared to the ones from different locations. Please note, that

we visualized the embedding space with two different dimensionality reduction techniques (UMAP[6] and T-SNE[7]) and independently of the technique, we obtain similar results.

3.3.2. Comparison against Baselines

In order to evaluate the effectiveness of our approach, we compare our method against two baseline approaches:.

- We define the first baseline as segmentation network that was trained on images over all locations in the dataset. (no embedding)
- As the second baseline, we randomly select one of the *k=3,5* trained expert segmentation network during test-time and compute the segmentation mask. If the embedding would not cluster images of the same location together and be rather randomly distributed, our approach would behave similarly to this baseline.

The quantitative results of the three approaches can be seen in Tab. 1 and Tab. 1. Both tables show that our approach performs best against both baselines yielding to an improvement of the IoU by more than 2% for specific cities. It can be also observed that the first baseline performs better than the second one, which shows the large variety of image features across locations. When we compare results across both tables, for a different number of expert segmentation networks, we can see that the approach with a higher number of expert segmentation networks performed better. This also indicates the large variety of images from different locations.

4. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the challenge of learning a location-specific image embedding for aerial and satellite imagery using deep neural networks. We focused on semantic segmentation of building footprints from high resolution satellite imagery and showed how such learned embeddings can be effectively used to select a model from a set of expert networks that is suited best to segment the correspondig image. On the Inria Aerial Image Labeling Dataset, our approach outperformed recent methods that were trained on the whole dataset, by more than 2% for specific cities.

Building upon this work, we plan to extend the learned embedding space to other location-specific features that can be derived from the ground truth. It would be for instance possible to learn an image embedding that takes the density and size of buildings into account. In this context, we also plan to make the step from segmenting building footprints from a set of few and diverse locations to the footprint segmentation at a global scale using satellite imagery and by relying on a small set of expert networks only. This would allow mapping human settlements with publically available satellite imagery, as introduced by Helber et al. [14] on a global scale with a set of location-specific segmentation networks.

5. REFERENCES

[1] 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.

[2] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," *arXiv preprint arXiv:1709.05932*, 2017.

[3] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[6] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018.

[7] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[8] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks.," in *BMVC*, 2016, vol. 1, p. 3.

[9] 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*. Springer, 2015, pp. 234–241.

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

[11] 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*, vol. 115, no. 3, pp. 211–252, 2015.

[12] Vladimir Iglovikov, Selim Seferbekov, Alexander Buslaev, and Alexey Shvets, "Ternausnetv2: Fully convolutional network for instance segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[13] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[14] Patrick Helber, Benjamin Bischke, Jörn Hees, and Andreas Dengel, "Towards a sentinel-2 based human settlement layer," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019.