

SDC-SDL-UNet3++: Highly Accurate Full-Scale Spatially Dependent Dilated U-Net Architecture for Road Segmentation

Group: Cillers Mike Boss Jonas Passweg Jonathan Ehrat
Department of Computer Science, ETH Zurich, Switzerland

Abstract

Since the original U-Net paper exploring its use in biomedical image segmentation many extensions have been developed to improve the performance in medical image segmentation. In this paper, we explore multiple such extensions and the combination thereof to solve segmentation of streets in road aerial satellite images. We show that the U-Net architecture as well as its extensions perform well on such an out-of-domain task to their original goal. [12]

1 Introduction

Semantic segmentation describes the task of partitioning a picture into different segments with dissimilar characteristics. For image segmentation, each pixel is assigned to a different segment class. Non-connecting segments belong to the same class in contrast to instance segmentation which differentiates between different instances of the same segment class. In this paper, we perform binary road segmentation assigning each pixel-patch to the road or background segment respectively.

Semantic image segmentation has many applications, most notably in medicine [8] or agriculture [15]. For road segmentation in particular, applications include automatic road detection [10], automatic map generation [7, 18], road damage detection [9], as well as "urban planning, telecommunication, disaster monitoring, navigation, updating geographic databases, and urban dynamic monitoring" in the geospatial field [3]. Most of these examples rely on having up-to-date maps to route personal or emergency vehicles. With our transportation system constantly growing [10], global positioning systems not being able to function everywhere [18] and natural disaster destroying roads [9], such automatic detection solutions have become an important human helper to keep maps and map systems up-to-date.

Especially in the medical field, the U-Net architecture has made a great impact with many extensions and adaptations [17, 6, 19, 1, 4] to the original architecture for different use cases and improvements. U-Nets and its extensions have been used for tasks in different domains such as aerial image segmentation [14, 13, 3, 5, 16] closely related to the road segmentation explored in this paper.

We explore the use and combination of four such U-Net [6, 17, 2, 19] extensions, originally developed for medical image segmentation, for the task of road

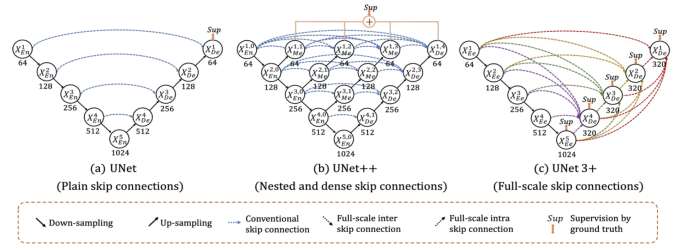


Figure 1: U-Net architecture comparison [6]

segmentation. The extensions were chosen because of their properties relating to road segmentation and their interactions.

Two more powerful U-Net architectures were explored, namely UNet++[19] and its extension UNet3+[6], connecting the encoder and decoder through nested, dense skip pathways. UNet3+ extends UNet++ using full-scale skip connections to explore more information from full scale.

The three U-Net architectures were enhanced by adapting them with stacked dilated convolutions[17] and Spatial Dependency Layers[2] as well as a combination of both adaptations.

2 Method

In this section we shortly describe the base architectures as well as the layers used in our model. For a detailed explanation we refer to the respective papers. Afterwards, we describe the reason each extension translates from the medical domain to road segmentation as well as how they were combined.

2.1 Architectures

2.1.1 U-Net

A U-Net architecture consists of a contracting encoder and expansive decoder connected at each layer as seen in figure 1. The layers are connected through a max-pool for the encoder and up-convolution for the

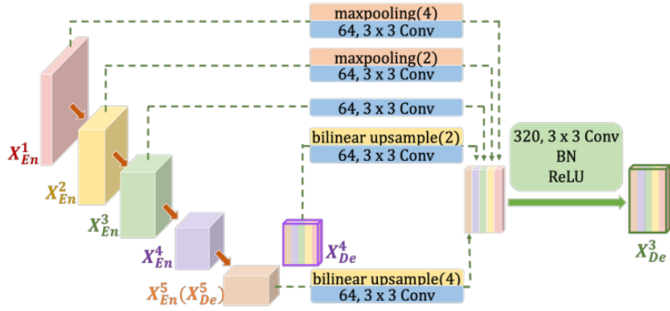


Figure 2: Illustration of how to construct the full-scale aggregated feature map of third decoder layer X_{De}^3 . [6]

decoder. Each layer step in the encoder downsamples the input while doubling the number of feature channels which is reversed at each layer of the decoder.

2.1.2 U-Net++

The U-Net++ architecture re-designs the skip pathways between the encoder and decoder. The pathways are enhanced by a nested pyramid structure of convolutional blocks as well as the original skip connections replaced with dense convolutional blocks as seen in figure 1. The architecture supports deep-supervision which is not used in this paper.

2.1.3 U-Net3+

The U-Net3+ architecture replaces the nested pyramid structure introduced in U-Net++ with full-scale skip connections as seen in figure 1. In contrast to the original U-Net each decoder layer is connected to all encoder layers. The full-scale connections at each decoder level are concatenated after which convolutions are performed as seen in figure 2. The architecture allows for full-scale deep supervision in accordance with U-Net++. The architecture adds a classification-guided module (CGM) which separately classifies if segments are prevent over-segmentation in the main model.

2.2 Layers

2.2.1 Stacked Dilated Convolutions (SDC)

At each decoder/encoder layer the convolutions are replaced by multiple stacked convolutions increasing in dilation rate while decreasing in channel numbers. The dilation rate is decreased such that the concatenated convolutions are the same as the output size of the layer as seen in figure 3.

2.2.2 Spatial Dependency Layers (SDL)

In each layer the convolutions are replaced with a SDL after the convolutional layer. Only the last three layers are replaced, seen in figure 4, as SDLs are performance intensive and results are similar. SDLs consist of a project-in, a correction, and a project-out stage which in contrast to convolutions are applied on non-symmetric receptive field as seen in figure 4. The field is extended much larger than normal convolutions in both directions.

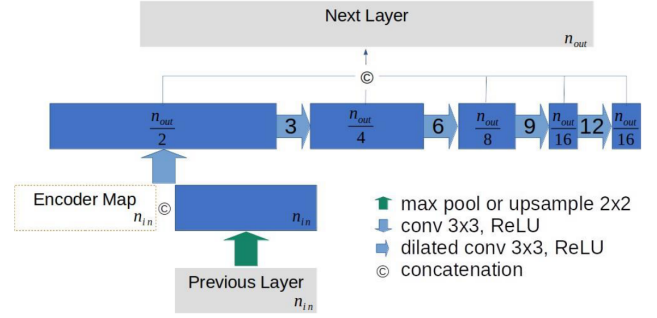


Figure 3: Illustration of one layer of the dilated UNet variant. The encoder (max pooling the preceding layer and without dashed-line box) or decoder (upsampling the preceding layer and concatenating dashed-line box) operation. The boxes indicate the feature maps, with the channel number denoted by equations of n_{in} and n_{out} in the boxes. The number within the arrow indicates the dilation rate. [17]

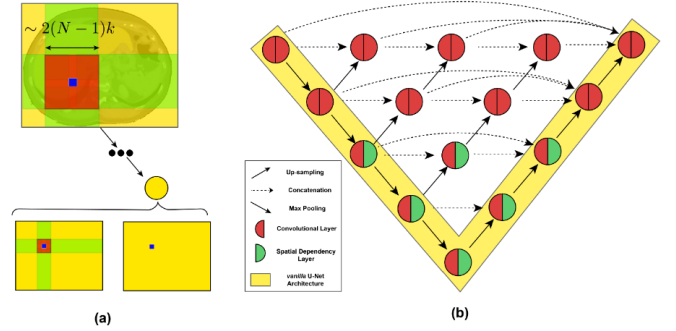


Figure 4: Illustration of SDNU-Net. Convolutional layers are switched with spatial dependency layers, which extend the receptive field bidirectionally [2].

2.3 Domain Relation

Road segmentation contains unique properties different to certain tasks in the medial image domain. Roads mostly present themselves as a single connected segment spanning across the whole image while separating the other segments in a highly structured fashion. As such more information from higher scales greatly increase the spatial information the model obtains from the road structure. Therefore, the U-Net++ and U-Net3+ architectures provide a natural fit for the task as both reduce the semantic map between the decoder/encoder while increasing the spatial information from higher layers. Of additional interest is the classification-guided module of the U-Net3+ to detect images without roads present. This module fits nicely with automatically generated data sets that may contain pictures without roads, as well as cropped training, that crops into a non-road part.

The receptive fields of standard convolutions is symmetrical and quite shallow for long structured segments such as roads spanning the image. SDLs extend the receptive fields in a bi-directional fashion fitting straight road structures. The larger symmetric

receptive-field of the SDCs fit better to crossing road sections as more spatial information about the nature of the in-leading roads is captured in comparison to the standard convolution.

For these reasons we believe these U-Net extensions are a natural fit for road segmentation irrespective of their original problem domain. Additionally, as the layers can organically be introduced in the extended architectures as well as combined blending these extensions fits with the unique properties of the task.

2.4 Extension Combinations

To introduce the SDCs to the architectures, the convolutions in each layer are replaced with their dilated versions and concatenated together. For the U-Net and U-Net++ architectures, the SDLs replace the second convolution in a convolutional block, as described in the paper. In the U-Net3+ architecture the SDL replaces the convolution on the concatenated skip-connections. Both added to an architecture first introduce the SDCs replacing the normal convolutions after which the SDL is added.

2.5 Models

All possible combinations of architectures with layers were tested and the results are presented in the evaluation section. Therefore, for every architecture (U-Net, U-Net++, U-Net3+) a model was created for the base model, the model with SDCs, a model with SDLs, as well as a model with both SDCs and SDLs resulting in a total of 12 models. The UNet++ and UNet3+ models were used without deep supervision as the model size was quite significant for the data. The UNet3+ model is without the CGM as in the dataset there are no images without roads.

2.6 Dataset

The dataset consists of aerial satellite RGB training images of size (400, 400) and their corresponding gray-scale masks of size (400, 400) and test images of size (608, 608). The original dataset contains 100 training image/mask pairs as well as 94 test images. We used a larger dataset[11] containing about 2000 image/mask pairs from the same dataset for training. The larger dataset is slightly pre-processed with random brightness and contrast adjustments. We use a 90% train/validation split with about 1600 training and 400 validation images. The images are batched into batches of size 8 by the data-loader.

2.7 Loss function

We use a combination of a dice score with binary cross entropy loss on the predicted masks with the ground-truth.

2.8 Preprocessing

To reduce image size with consistent sizes we split the images into overlapping images of size (320, 320). For the training dataset the stride is 40 generating 9 overlapping images. As the images are larger in the

test dataset we use a stride of 70 generating 25 overlapping images. For training the dataloader returns a single image part whereas for predicting the test set all parts are returned. The masks are split in the same fashion. The image parts are down-sampled by a patching function to size (160, 160). For the training dataset the image parts is randomly rotated by any angle and sometimes flipped horizontally and vertically.

Other preprocessing steps were tried such as filtering road color and super-pixels. The superpixels were further processed by hierarchical merging on region boundary Region Adjacency Graphs. The super-pixel approach seemed to work quite well by visual inspection but the filling of the image with the mean super-pixel color was computationally too expensive. These additional preprocessing steps are present in the implementation for possible future use.

2.9 Postprocessing

For predicting the test images, as each image is returned split into 25, each part is predicted after which all parts are recombined into a final image. As the model sometimes misses road parts surrounded by other road we dilate then erode the predicted mask. Every image is predicted 4 times by rotating the image by 90° each time. The predicted mask is patched to the output format of patches of 16 by 16 pixels. Majority voting on the patched predicted masks is performed to obtain a single mask. The patched predicted mask is again dilated and eroded.

Additionally, we introduced fixing the predicted mask of each part by another U-Net model trained on the predicted mask and ground-truth. This approach interprets fixing the mask as another segmentation problem with one channel instead of three. The approach seems quite natural as it is difficult for the U-Net to predict all roads as some are obstructed but easier for the U-Net to fix such clearly visible obstructions on the predicted mask. The same approach was taken to fix the final predicted mask as well. For training a single image part is up-scaled after which it is patched and fixed by the U-Net. For prediction the U-Net is used on the final predicted mask.

Both of these additional U-Net lead to slight performance increases but because of time constraints and pytorch-lightning limitations we did not thoroughly test these approaches therefore decided against using it in our final model. The approaches are in the system in the implementation and can easily be trained by freezing the model and

3 Evaluation

Our model approaches were first severely limited by the available data. As such we present two tables one with the models fitted on the original data and one on the larger dataset. As these results were created during creation of the models the first table is based

Network	Pixel-Wise Accuracy
UNet	0.941
UNetDilated	0.941
UNetSpatial	0.944
UNetSpatialDilated	0.938
NestedUNet	0.943
NestedUNetDilated	0.953
NestedUNetSpatial	0.940
NestedUNetSpatialDilated	0.950
UNet3Plus	0.941
UNet3PlusDilated	0.949
UNet3PlusSpatial	0.946
UNet3PlusSpatialDilated	0.953

Table 1: original data, *patience* = 60

on pixel-wise accuracy scores. For the second table we switched to pixel-wise f1 score more closely resembling scores achieved in the competition. The accuracy scores corresponding to the f1 scores in the second table are considerably higher than for the models trained on the original data.

From the evaluation of the tables we can see that the UNet3PlusSpatialDilated model twice outperformed the other models. However, it is not clear which parts improved which model as the introduced layers in the architectures vary in score. As training these models is quite time intensive doing so in a clear way to differentiate between improvements was not possible. It is possible that patience was set to low for some of the larger models to fully converge.

3.1 Competition

The data bottleneck is clearly visible in the public Kaggle scores. Our model trained on the original data received a score of just 0.86682 which improvement greatly to 0.91665 simply by training on more data.

The best score in the Kaggle competition is achieved by NestedUNet with a score of 0.92686. In contrast, the best local model UNet3PlusSpatialDilated achieved a score of 0.91409. We deemed this discrepancy to be because of the test set and not because of the model as it gained a better local score and looked better on visual inspection. It will be of interest to see the private score on each model in the competition as it could be a lucky score instead of a consistent one.

To further improve our score we could have changed the validation set to selected images more resembling the final test set. However, this would not lead to a great road segmentation model instead overfitting on the competition pictures. Therefore, we did not pursue such paths further.

4 Conclusions

In this work, we propose the usage of U-Net extensions and their combinations for the task of road seg-

Network	Pixel-Wise F1
UNet	0.770
UNetDilated	0.752
UNetSpatial	0.771
UNetSpatialDilated	0.787
NestedUNet	0.772
NestedUNetDilated	0.763
NestedUNetSpatial	0.737
NestedUNetSpatialDilated	0.757
UNet3Plus	0.769
UNet3PlusDilated	0.766
UNet3PlusSpatial	0.778
UNet3PlusSpatialDilated	0.789

Table 2: all data, *patience* = 40

mentation. We show that with the right extensions, that fit to the task domain, we can achieve good performance. The performance was evaluated in a Kaggle competition. We show how such extensions can be combined to perform better. In general, our results indicate that the U-Net architecture and its extensions, even though developed originally for the medical domain, generalize over other problem domains for semantic segmentation.

References

- [1] João Carvalho, João Santinha, Dorde Miladinović, and Joachim Buhmann. Spatially dependent u-nets: Highly accurate architectures for medical imaging segmentation, 03 2021.
- [2] João B. S. Carvalho, João A. Santinha, Dorde Miladinović, and Joachim M. Buhmann. Spatially dependent u-nets: Highly accurate architectures for medical imaging segmentation, 2021.
- [3] Ugur Avdan Firat Erdem. Comparison of different u-net models for building extraction from highresolution aerial imagery. *International Journal of Environment and Geoinformatics (IJE-GEO)*, 7(3):221–227, 2020.
- [4] Pius Gadosey, Yujian Li, Enock Adjei Agyekum, Ting Zhang, Joying Liu, Peter Yamak, and Firdaus Essaf. Sd-unet: Stripping down u-net for segmentation of biomedical images on platforms with low computational budgets. *Diagnostics*, 10:110, 02 2020.
- [5] Yuwu Hou, Zhaoying Liu, Ting Zhang, and Yujian Li. C-unet: Complement unet for remote sensing road extraction. *Sensors*, 21(6):2153, 2021.
- [6] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet

- 3+: A full-scale connected unet for medical image segmentation, 2020.
- [7] Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler. Learning aerial image segmentation from online maps. *CoRR*, abs/1707.06879, 2017.
- [8] Tao Lei, Risheng Wang, Yong Wan, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey, 2020.
- [9] Haijian Ma, Nan Lu, Linlin Ge, Qiang Li, Xinzhaoyou, and Xiaoxuan Li. Automatic road damage detection using high-resolution satellite images and road maps. In *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, pages 3718–3721, 2013.
- [10] Volodymyr Mnih and Geoffrey Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th Annual International Conference on Machine Learning (ICML 2012)*, June 2012.
- [11] Additional road data. https://github.com/matejsladek/CIL_street/tree/cluster_new/data/maps1800/all. Accessed: 2021-07-31.
- [12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
- [13] Reza Akbari Dotappeh Sofla, Tayeb Alipour-Fard, and Hossein Arefi. Road extraction from satellite and aerial image using SE-Unet. *Journal of Applied Remote Sensing*, 15(1):1 – 14, 2021.
- [14] Ashish Soni, Radhakanta Koner, and Vasanta Govind Kumar Villuri. M-unet: Modified u-net segmentation framework with satellite imagery. In Jyotsna Kumar Mandal and Somnath Mukhopadhyay, editors, *Proceedings of the Global AI Congress 2019*, pages 47–59, Singapore, 2020. Springer Singapore.
- [15] Priit Ulmas and Innar Liiv. Segmentation of satellite imagery using u-net models for land cover classification. *CoRR*, abs/2003.02899, 2020.
- [16] Leipeng Wang, Yinzhong Ye, et al. Computer vision-based road crack detection using an improved i-unet convolutional networks. In *2020 Chinese Control And Decision Conference (CCDC)*, pages 539–543. IEEE, 2020.
- [17] Shuhang Wang, Szu-Yeu Hu, Eugene Cheah, Xiaohong Wang, Jingchao Wang, Lei Chen, Masoud Baikpour, Arinc Ozturk, Qian Li, Shinn-Huey Chou, Constance D. Lehman, Viksit Kumar, and Anthony Samir. U-net using stacked dilated convolutions for medical image segmentation, 2020.
- [18] Terence J. Yi. Semantic segmentation of aerial imagery using u-nets. Master’s thesis, Air Force Institute Of Technology, 3 2020. Theses and Dissertations. 3593.
- [19] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

SDC-SDL-UNet3++: Highly Accurate Full-Scale Spatially
Dependent Dilated U-Net Architecture for
Road Segmentation

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):

Ehrat
Passweg
Boss

First name(s):

Jonathan
Jonas
Mike

With my signature I confirm that

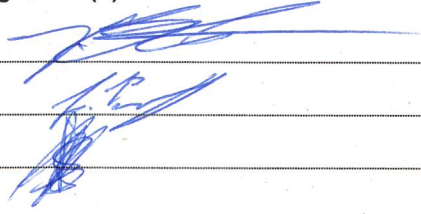
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Zürich, 31.07 2021

Signature(s)



For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.