CleanRoad: Road Segmentation of Aerial Images by Fusing Modern and Traditional Methods

Kenan Bešić, Niklaus Houska, Steven Stalder, Sebastian Winberg
Group: CILlitbang
Department of Computer Science, ETH Zurich, Switzerland

***Abstract*—Automatic road segmentation brings great value for traditional and modern uses of satellite road data. In this paper we propose an approach for the semantic segmentation of roads from aerial images, in which we combine neural networks with more traditional, well-established computer vision algorithms such as graph cut, Hough transform and region removal. We modify the well-known U-Net architecture by adding dilated and transposed convolutional layers and also introduce a dedicated post-processing pipeline using the aforementioned traditional algorithms. By doing so, we produce clean segmentation masks and outperform existing solutions, achieving an F1-score of around 92.3% on a separate test set.**

I. Introduction

Many applications benefit from the automatic detection of roads. The creation of more efficient traffic or GPS systems can be facilitated through more accurate live satellite data of the road network. Additionally, more modern applications such as autonomous driving may get an even larger benefit from this information. Consequently, road segmentation – or semantic segmentation in general – has been an intensively researched topic.

With the rise of deep learning, traditional computer vision methods have been replaced by sophisticated convolutional neural network architectures. In this paper, we apply the popular, fully convolutional U-Net design [1] to road extraction and present a novel architecture that benefits from dilated convolutions which have been shown to be more accurate in segmentation tasks [2].

Moreover, we present a dedicated post-processing pipeline designed for accurate road segmentation. We apply traditional computer vision methods such as the Hough transform [3], graph cut [4] and connected region detection [5] to predictions of the neural network to get smoother and better connected results. Furthermore, we believe our way of utilizing the Hough transform has not been described in any other publication.

This paper is structured in the following way. In Chapter II, we present our proposed models and methods. In Chapter III, we evaluate the performance gains of our approaches and compare the results to multiple baselines. In Chapter IV, we discuss our results and show the strengths and weaknesses of our proposal.

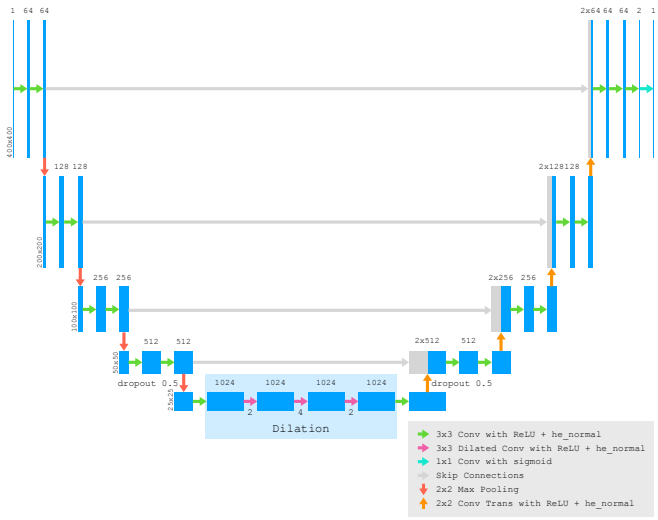

Figure 1. U-Net with dilated convolution layers in the middle and transposed convolutions for upsampling.

II. Models and Methods

A. Models

For our underlying neural network model, we implemented and evaluated several variants of the U-Net architecture [1], which has established itself as an excellent method, especially for biomedical image segmentation, but also for other segmentation tasks. Taking an existing implementation from [6], we modified the architecture to boost its performance for the road segmentation task at hand.

1) Patch-based U-Net: To mitigate the effect of the noticeable differences between our training and test sets, our first modified model architecture is trained on patches rather than entire images. More precisely, we first randomly sample 128×128 patches from the training images and predict only the center 32×32 patch. We have adapted the U-Net to allow varying input and output sizes by removing the two last upsampling steps and adapting the skip connections to only merge the 32×32 center neurons of the previous layers. This network design allows for natural data augmentation, as patches can be sampled from random locations in the training images, however, it also requires more training iterations to produce good results.

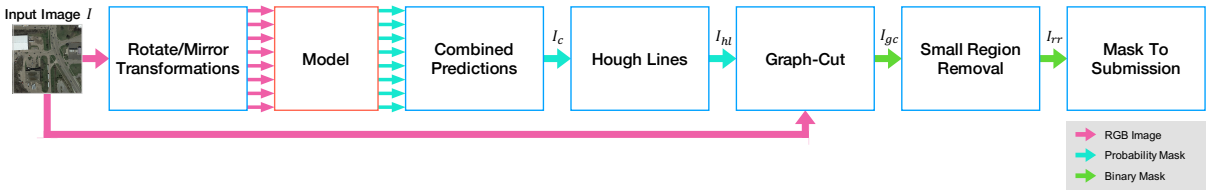

Figure 2. Dedicated post-processing pipeline.

2) Dilated U-Net: In contrast to the previous approach, we trained our second architecture on whole images and also experimented with dilated convolutions [2]. By increasing the receptive field of some of our convolutional layers without losing any coverage or resolution, we aimed at assisting the neural network with difficult areas in the images which require more context information. Similar to [7], we placed a dilated convolution module between the contracting and the expansive path of the U-Net. The reasoning behind this approach is that during the contraction, the U-Net loses spatial information, which can be preserved to some extent by utilizing dilated convolutions in the deepest part of the network.

We have evaluated many different ways of introducing dilated convolutions into the U-Net structure and selected the two most promising configurations. In the first dilated model *U-Net Dilated v1*, we added two layers with dilation rates of 2 and 4 respectively to the bottom part of the U-Net. The second dilated model *U-Net Dilated v2* uses five layers in the bottom part of the architecture compared to the two of the base implementation of the U-Net. These come with dilation rates of 1, 2, 4, 2, and 1 respectively. This idea of first increasing and then again decreasing the dilation rates has been loosely adapted from [8].

3) Dilated U-Net with Transposed Convolutions: Additionally, we implemented the above-mentioned dilated models with transposed convolutions [9], which have been used as a replacement for the upsampling layers in many state of the art architectures [10][11][12]. In our previously described networks, every upsampling phase consists of an upsampling layer followed by a convolutional layer. Replacing those two layers with one transposed convolutional layer with stride 2, we achieve that the neural network learns a custom upsampling method. Applying transposed convolutions to the previously mentioned *U-Net Dilated v2* results in the architecture illustrated in Figure 1.

B. Data Augmentation

An important part of [1] is data augmentation, which is also crucial in our setting, as we only have a very small dataset of 100 training images. These augmentations include mirroring, rotation, small horizontal and vertical shifts, brightness changes, shearing, as well as zooming in and out. It has been shown that data augmentation can reduce overfitting to the training set significantly while making the neural networks more rotation, brightness, and scale invariant [13].

C. Additional Data

In order to improve our results and further prevent early overfitting, we selected another 121 images from the same dataset [14] as our initial training images. We put our emphasis on parking lots, light streets, shadows, and industrial sights, as these types of images were very present in the test set, however heavily underrepresented in the training set.

Another challenge was the size difference between training and test images with 400×400 and 608×608 pixels respectively. While we investigated a sliding window approach, where multiple 400×400 sub-images are predicted and the overlapping parts averaged, we eventually decided to downsample the test images to a target size of 400×400 and resize the output again. This approach especially helped since our additional training data was of a similar scale as the test images, allowing the neural network to be more scale invariant.

D. Dedicated Post-Processing Pipeline

1) Combined Predictions: To stabilize the prediction of our model, we rotate each test image by 90, 180, and 270 degrees. Additionally, we mirror all four images and then predict a mask for the resulting eight inputs. By reversing the transformations and combining the results, we get a final prediction by either taking the average over the probability masks or by applying a voting mechanism on the binary segmentations.

2) Hough Lines: The Hough transform [3] enables easy detection of lines in binary images. As it has been used often for classical image segmentation, we propose a post-processing step to improve the connectivity of segmented roads. Given a pixel-wise probability map as input, we output an updated probability map using the Hough transform. First, we map the continuous probability mask to a binary mask by thresholding. Next, we apply the Hough transform on the binary mask to detect straight lines. These mostly hit pixels already classified as road segments. However, some lines connect separate regions. For pixels hit by such lines, we raise the probability in the probability map by a constant factor ϵ. For an appropriately chosen value of ϵ, this step can yield better connectivity for road segments that are disconnected by lower confidence regions.

3) Graph Cut: Inspired by [4], we represent the image as a graph and find a graph cut to derive the binary masks. The graph $G = (V, E)$ is constructed as follows: V contains a vertex for every pixel as well as two sink vertices S and B representing the labels for street and background. E is the set of edges connecting each pixel vertex to S with a weight corresponding to its probability in the probability map I_{hl} and to B with $1 - I_{hl}$ respectively. Additionally, E contains edges for neighboring pixel vertices with an edge weight derived from the similarity of the pixels in the Lab color space. The graph cut minimizes the following energy function:

$$E(L) = \sum_i P(l_i) + \lambda \sum_i \sum_{j \in \mathcal{N}(l_i)} R(l_i, l_j),$$

where L are the binary labels and $P(l_i)$ is the penalty to assign label l_i based on the probability values of I_{hl}:

$$P(l_i) = \begin{cases} I_{hl,i} & l_i = 1 \\ 1 - I_{hl,i} & l_i = 0 \end{cases}.$$

Further, $R(l_i, l_j)$ is the penalty to assign different labels to pixels with similar color values I_i and I_j:

$$R(l_i, l_j) = \begin{cases} \exp\left(-\frac{(I_i - I_j)^2}{2\sigma^2}\right) & l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}.$$

Lastly, $\lambda > 0$ is the parameter to control the penalty induced by the second term. Finding a minimum cut in G and labeling the pixels based on their partition yields the desired binary mask I_{gc} (see Figure 2).

4) Region Removal: The classification from a probability map to a binary mask can result in images with many small unconnected parts. To clean up the binary mask we detect small regions classified as roads and remove them. Even though we might remove correctly classified parts of the image, we decide to discard such artifacts since they are not useful.

E. Threshold for Submission

Finally, for the Kaggle competition we perform a per-patch voting to decide the overall value of 16×16 patches. We set the voting threshold to 50% since the neural network and the post-processing pipeline perform a good job on a pixel-level basis. Thus, a patch gets classified as a road if more than half of the pixels have been classified as a road.

III. Results

In this section, we present the evaluations of the aforementioned models and methods. For the optimizer, we chose Adam [15] with a learning rate of $1e-4$ for all experiments of our own models. Moreover, we always utilized the He normal initializer [16], as well as ReLU and sigmoid activation functions as described in Figure 1. Furthermore, we have experimented with different loss functions such as dice loss, weighted binary cross-entropy, and binary cross-entropy, with the latter being our final choice, as it yielded the best results.

On a GeForce GTX 1080 GPU, our longest training runs over 500 epochs and 100 steps per epoch took 7-8 hours.

Table I
Model Architecture Experiments

Model	Kaggle Score
U-Net	0.87722
Keras Segmentation (Resnet50 U-Net)	0.89589
U-Net Patch-Based	0.86083
U-Net Dilated v1	0.89898
U-Net Dilated v2	0.88979
U-Net Dilated v1 Transposed	0.88886
U-Net Dilated v2 Transposed	**0.90268**

Figure 3. Result comparison between standard U-Net and *U-Net Dilated v2 Transposed* for test images 107, 151 and 201. Both models ran for 50 epochs with 100 steps per epoch on the original 100 training images.

A. Model Architectures

Table I lists our first set of experiments, in which we compare various different neural network architectures. With the exception of *U-Net Patch-Based*, all of the models have been trained for 50 epochs and 100 steps per epoch, using the original training set containing 100 training images. We have trained the patch-based model for 250 epochs and 200 steps to account for the lower amount of information that is passed to the neural network in each training iteration. Also, every network utilizes the same data augmentation on the training images, as it is crucial for getting reasonably good results (see Section II-B).

We compare our own, modified U-Net models to two baselines: the basic, unmodified U-Net architecture from [6] and the *keras-segmentation* library [17], using their *resnet50_unet* model. With the exception of our patch-based model, all of our modified architectures were able to significantly outperform the basic U-Net. Especially the effect of the increased receptive field of the dilated models on the results is apparent in Figure 3. Moreover, two of

Table II
REFINEMENT EXPERIMENTS

Improvements	Kaggle Score
U-Net Dilated v2 Transposed	0.90268
U-Net Dilated v2 Transposed + Additional Data (50E)	0.90589
U-Net Dilated v2 Transposed + Additional Data (500E)	0.91796
U-Net Dilated v2 Transposed + Average Combination	0.92231
U-Net Dilated v2 Transposed + Hough + Region Removal	0.92331
U-Net Dilated v2 Transposed + Graph Cut	**0.92355**

Note that each step going downwards is including all previous improvements from above. Hough and region removal steps were combined for the experiments to account for the clutter introduced by the Hough transform.

our models even achieved a higher score on Kaggle than *resnet50_unet*. Since our *U-Net Dilated v2 Transposed* is our overall best-performing network architecture, we continue further experiments on this model.

B. Methods and Measures

1) Additional data: For the evaluation of our proposed methods, we conducted a second set of experiments to illustrate the benefits of each approach. To start, we enhanced our training and validation sets with additional data (see Section II-C), which already yielded an improvement of around 0.3% over 50 epochs on our best-performing model (see Table II). However, the overall larger and improved dataset also enabled us to let the model run for more iterations. Due to the under-representation of certain kinds of images in the original dataset (see Section II-C), our validation set did not serve as a reliable indicator for the performance of our model on the test set. For a larger amount of training iterations, this led to overfitting problems which could be avoided with the improved dataset.

2) Post-Processing: Next, we evaluated how combining predictions of the original image with its transformed versions improves the final mask (see Section II-D1). With that, we achieved an improvement of around 0.4% on Kaggle. Note that we averaged the continuous masks instead of applying the voting mechanism to combine the predictions.

The purpose of the remaining post-processing steps is to clean up and further improve the already good results. On Kaggle these traditional computer vision algorithms increase the score by another 0.124%. This improvement can be greater for less sophisticated neural networks. We also evaluated the benefit of these steps on the predictions of the standard U-Net and achieved an improvement of over 0.3%. In Figure 4 it becomes apparent how the Hough transform is able to connect previously unconnected roads. Furthermore, the region removal and graph cut steps remove unnecessary clutter from the images. Overall, these final post-processing measures undoubtedly improve the general quality of the resulting segmentation masks.

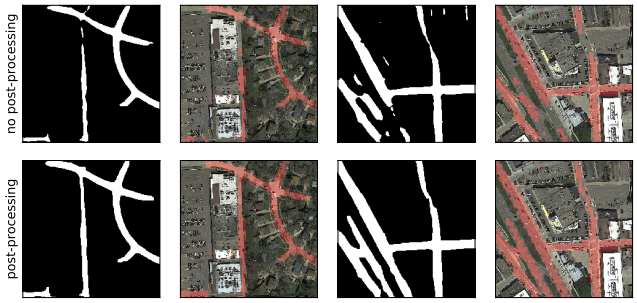

Figure 4. Comparison between unprocessed and post-processed results (Hough transform, graph cut, region removal) for test images 155 and 169. The *U-Net Dilated v2 Transposed* model was trained with additional data for 500 epochs, 100 steps per epoch and used average combined outputs.

IV. DISCUSSION

With the inclusion of all proposed methods, our own, modified U-Net architecture can correctly classify most of the roads in the provided satellite images whilst maintaining a low number of false positives. The application of dilated and transposed convolutions clearly enhances the network's ability to connect previously disconnected road segments. Also, most of the roads are very clearly bordered and the resulting images appear very clean and smooth as a result of our post-processing steps. However, one can also identify cases in which our network still fails to detect roads. It appears to struggle with two-lane roads as well as roads that have a lighter surface color. These problems might be alleviated when the neural network is provided with more such types of roads, e.g. by utilizing another dataset from a different city. Unfortunately, manually modifying the color of streets in the existing dataset or converting the input into different color spaces did not show clear improvements. Furthermore, our post-processing pipeline is sensitive to changes in hyper-parameters. Selecting improper values might lead to a deterioration of the resulting images.

V. SUMMARY

We have successfully tackled the problem of extracting roads from satellite imagery with high accuracy, despite our final training and validation datasets only containing a total of 221 images. Our modified U-Net architecture is able to reliably detect most roads in a separate test set after a training time of only a few hours and has been evaluated to significantly outperform the original U-Net model in this setting. The proposed post-processing methods introduce traditional and well-established algorithms from the computer vision domain into a machine learning environment. We have utilized these post-processing steps to visibly clean up and enhance the outputs of the neural network. Especially our way of applying of the Hough transform is – to the best of our knowledge – a novel approach in this field.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[2] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.07122

[3] P. V. Hough, "Method and means for recognizing complex patterns," Dec. 18 1962, U.S. Patent 3,069,654.

[4] F. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," in *2012 International Conference on Systems and Informatics (ICSAI2012)*, 2012, pp. 1936–1941.

[5] K. Wu, E. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," vol. 5747, 04 2005.

[6] X. Zhi. unet for image segmentation. Accessed: 02.06.2020. [Online]. Available: https://github.com/zhixuhao/unet

[7] S. Piao and J. Liu, "Accuracy improvement of UNet based on dilated convolution," *Journal of Physics: Conference Series*, vol. 1345, p. 052066, nov 2019. [Online]. Available: https://doi.org/10.1088%2F1742-6596%2F1345%2F5%2F052066

[8] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 2018, pp. 1442–1450. [Online]. Available: https://doi.org/10.1109/WACV.2018.00162

[9] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.

[10] C. C. Chatterjee, M. Mulimani, and S. G. Koolagudi, "Polyphonic sound event detection using transposed convolutional recurrent neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 661–665.

[11] S. Piao and J. Liu, "Accuracy improvement of UNet based on dilated convolution," *Journal of Physics: Conference Series*, vol. 1345, p. 052066, nov 2019. [Online]. Available: https://doi.org/10.1088%2F1742-6596%2F1345%2F5%2F052066

[12] Y. Chen, K. Wang, X. Liao, Y. Qian, Q. Wang, Z. Yuan, and P.-A. Heng, "Channel-unet: A spatial channel-wise convolutional neural network for liver and tumors segmentation," *Frontiers in Genetics*, vol. 10, p. 1110, 11 2019.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[14] K. Pascal, W. J. Dirk, L. Aurelien, J. Martin, H. Thomas, and S. Konrad, "Learning Aerial Image Segmentation From Online Maps," Jul. 2017. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2719738

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015. [Online]. Available: http://arxiv.org/abs/1502.01852

[17] D. Gupta. keras-segmentation. Accessed: 18.07.2020. [Online]. Available: https://github.com/divamgupta/image-segmentation-keras

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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):

CleanRoad: Road Segmentation of Aerial Images by Fusing Modern and Traditional Methods

Authored by (in block letters):

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

Name(s):	**First name(s):**
Stalder	Steven
Bešić	Kenan
Houska	Niklaus
Winberg	Sebastian

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

Zurich, 31.07.2020

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.