# Road Segmentation in SAR Satellite Images With Deep Fully Convolutional Neural Networks

Corentin Henry, Seyed Majid Azimi, and Nina Merkle

***Abstract*— Remote sensing is extensively used in cartography. As transportation networks grow and change, extracting roads automatically from satellite images is crucial to keep maps up-to-date. Synthetic aperture radar (SAR) satellites can provide high-resolution topographical maps. However, roads are difficult to identify in these data as they look visually similar to targets, such as rivers and railways. Most road extraction methods on SAR images still rely on a prior segmentation performed by the classical computer vision algorithms. Few works study the potential of deep learning techniques, despite their successful applications to optical imagery. This letter presents an evaluation of fully convolutional neural networks (FCNNs) for road segmentation in SAR images. We study the relative performance of early and state-of-the-art networks after carefully enhancing their sensitivity toward thin objects by adding the spatial tolerance rules. Our models show promising results, successfully extracting most of the roads in our test data set. This shows that although FCNNs natively lack efficiency for road segmentation, they are capable of good results if properly tuned. As the segmentation quality does not scale well with the increasing depth of the networks, the design of specialized architectures for roads extraction should yield better performances.**

***Index Terms*— Deep learning, high-resolution synthetic aperture radar (SAR) data, road extraction, SAR, semantic segmentation, TerraSAR-X.**

## I. Introduction

THE overall urban growth in the past two decades has led to a considerable development of transportation networks. Such constantly evolving infrastructure necessitates frequent updates of existing road maps. A wide range of applications are depending on this information, such as city development monitoring, automated data update for geolocalization systems, or support to disaster relief missions. A satellite equipped with a synthetic aperture radar (SAR) can get information on an area's topography. The resulting information is more robust to changes in illumination conditions and color fluctuation with respect to the optical imagery. Moreover, SAR sensors can operate independently of weather conditions and are therefore the sensor of choice to survey regions affected by weather-related disasters.

The extraction of roads in SAR satellite images has been researched for several decades [1] and is generally addressed in the following manner: road candidates are extracted from SAR images using a feature detector. This initial segmentation is then transformed into a topological graph, where each segment represents a road section. The graph is finally optimized to form a coherent road network, often by applying a Markov random field [1], [2] using contextual information from the SAR image to reconnect loose segments and correct the overall network structure. Recently, Xu *et al.* [3] proposed a conditional random field (CRF) model capable of jointly extracting road candidates and applying topological constraints. This end-to-end scheme reduced the inevitable performance loss occurring when separately extracting road priors and constructing a road network graph. These methods all rely on an efficient road candidate extraction algorithm, and most of them entrust this task to traditional computer vision algorithms. To date, few works study the potential of the recent advances in deep learning in the context of road segmentation.

Deep convolutional neural networks (DCNNs) first demonstrated unmatched effectiveness in 2012 on the ImageNet classification challenge, and their performance has been improving at a fast pace ever since, receiving a lot of attention from the computer vision community. However, unlike the medium-sized images used in classification competitions, the aerial images used in remote sensing often cover hundreds of square kilometers. Today, fully convolutional neural networks (FCNNs) are the most successful method to perform pixelwise segmentation on large-scale images. Given an input image, they produce an identically sized prediction map. Introduced in 2015 with FCN-8s [4], FCNNs allowed the establishment of new states of the art in semantic segmentation of aerial optical images [5] and were successfully applied to satellite SAR images [6].

Yao *et al.* [6] use off-the-shelf pretrained FCNNs on SAR images to classify buildings, land use, bodies of water, and other natural areas. They report good segmentation results for the land use and natural classes but unsatisfactory results for buildings, showing a striking performance contrast between larger and smaller objects. As roads are thin objects by nature, it becomes evident that the FCNN models must be specifically adjusted for our task. Starting from another perspective, Geng *et al.* successively proposed two methods for land cover classification, including roads. In [7], they emphasize low-level features in SAR images using traditional computer vision techniques on top of which they train a stack of autoencoders. In [8], they further improve their results by using long-short-term memory units to transform the 2-D information contained in the image into 1-D information fed to autoencoders. They report around 95% overall accuracy across several areas totaling 21 km$^2$; however, the road class accuracy

Manuscript received January 30, 2018; revised May 7, 2018 and June 27, 2018; accepted July 18, 2018. Date of publication August 27, 2018; date of current version December 5, 2018. *(Corresponding author: Corentin Henry.)*

The authors are with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: corentin.henry@dlr.de; seyedmajid.azimi@dlr.de; nina.merkle@dlr.de).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2864342

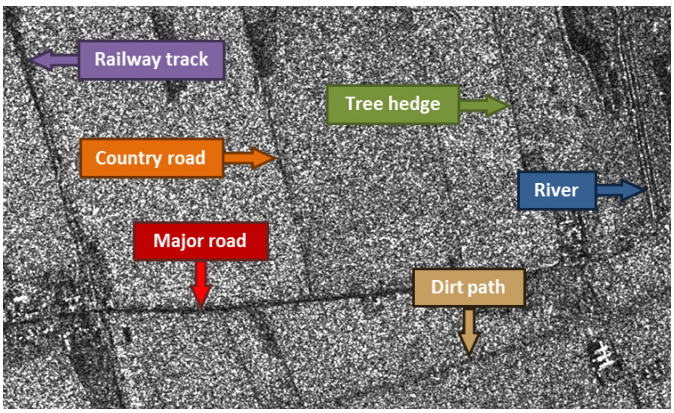


Fig. 1. SAR image sample showing that objects of different natures can look very similar. A segmentation model must learn to distinguish all kinds of roads from railway tracks, tree hedges, and rivers.

remains invariably behind the accuracy of all other classes by 10%–20%.

Roads are difficult to identify even in high-resolution SAR images. They can often be confused with other targets, such as railway tracks, rivers, or even tree hedges, as illustrated in Fig. 1. Identifying roads often involves the opinion of an expert, but deep learning proved it could deal with such delicate study cases, motivating the thorough assessment of the potential of some powerful FCNNs on the task at hand. The success of this initial experiment would open new prospects in the future, such as semiautomated annotation of SAR images, which could prove much faster and more reliable than fully manual annotation. Time-critical missions, such as disaster relief, would particularly benefit from a steep increase in the speed of satellite data analysis.

This letter presents the evaluation of three FCNNs for road segmentation in high-resolution SAR satellite images: FCN-8s [4], Deep Residual U-Net [5], and DeepLabv3+ [9]. Crucial adjustments are made in the training procedure to improve the base performance of the FCNNs with a class-weighted mean squared error (mse) loss and a control parameter over the spatial tolerance of the models. The evaluation is performed on several custom data sets whose design is critical to the success of the method and is therefore detailed. Unlike Yao *et al.* [6], we set aside the OpenStreetMap data due to the lower geolocalization accuracy compared to SAR data. Unlike previous works, we manually label every single road from the most visible highways to the less distinguishable dirt paths. We obtain good qualitative results and satisfying quantitative results, thus demonstrating the effectiveness of well-fit FCNNs as road candidate extractors in SAR images.

## II. Method

### A. Segmentation With Fully Convolutional Neural Networks

FCNNs are currently the most successful methods for pixelwise segmentation and are especially convenient for large-scale image processing. As they can deal with images of any size, they can take into account a wider context when trying to identify objects. They owe this flexibility property to their adaptive bottleneck layers, connecting the two key components of the network. The first element, a DCNN encoder, analyzes the images and outputs a cluster of predictions. The image data are gradually downsampled, proportionately becoming more meaningful. The second element, a decoder, applies upsampling operations to restore the spatial properties of the predictions until the predictions share the same size as the input image. It is often done using bilinear interpolation or fractionally strided convolutions, also called deconvolutions [10]. For classification tasks, the DCNN output is classified by fixed-size fully connected layers, the network's bottleneck, imposing a maximum input size upstream. For segmentation tasks, FCNNs remove this input size constraint by replacing the fully connected layers by convolutional layers.

We implement three substantially different FCNNs. The first one is FCN-8s [4] with a VGG-19 backbone [11], the first of all FCNNs that was successfully applied to a wide variety of computer vision tasks. In FCN-8s, two skip connections fuse the high-resolution information from early VGG19 layers into the upsampling process, thus improving the spatial accuracy of the resulting segmentation. To increase its training speed, we add a batch normalization step between each convolutional layer and ReLU activation, as well as after each deconvolutional layer. We use it to set the baseline performance for comparison with more recent architectures. The second one is Deep Residual U-Net [5] that demonstrated a great segmentation performance on the Massachusetts roads data set [12]. Its overall architecture is similar to FCN-8s although entirely symmetrical with a skip connection fusing each block of the encoder into the corresponding block of the decoder. Its backbone uses residual units [13] that let the input image data flowthrough the whole network. Propagating this information helps the network learn complex patterns more efficiently, and its application on SAR imagery could help reduce the impact of the speckle. The third one is DeepLabv3+ [9], one of the most recent architecture for semantic segmentation. Its Xception backbone also uses residual connections, but the network is much deeper with 65 nonresidual layers compared to Deep Residual U-Net's 15. Using dilated convolutions, DeepLab can leverage a larger context and better recognize targets from cluster, which should prove valuable for applications to SAR imagery.

### B. Adjusting the FCNNs for Road Segmentation

Roads appear as thin objects in SAR images and are likely outweighed by clutter, especially outside cities. We take some necessary steps to limit the class imbalance during training. A similar problem in the case of sports field lines extraction is addressed in [14] by tracing thick labels in the ground truth. In our case, it means that the labels must exactly cover the road outlines and embankments, insofar as they are visible in the SAR images. Pixels labeled as roads are set to a value of 1 and background pixels to 0 in the ground truth. In addition, we introduce a spatial tolerance parameter $t_{\max}$ operating as follows. The value of background pixels located at a distance $t \leq t_{\max}$ to the nearest pixel labeled as road is redefined as $1-(t/t_{\max}+1)$. The resulting ground truth is a smooth target distribution centered around the road labels, similar to what Luo *et al.* [15] proposed. Varying $t_{\max}$ allows controlling the tolerance of the training toward spatially small mistakes. Note that when referring to a binary ground truth (two classes), we assume $t_{\max}=0$.

As a consequence, the task changes from a binary classification to a binary regression: instead of predicting each pixel as either road or background, the network weighs how much each

pixel is likely to be a road. We make the following changes to adapt the FCNNs: the final activation on the logits is changed from a softmax to a sigmoid function, and the cross-entropy loss is replaced by an mse loss.

Eigen and Fergus [16] also tackle the class imbalance issue by reweighting each class upon the loss calculation. The loss for each pixel prediction is multiplied by a coefficient inversely proportional to the frequency of its true class in the ground truth. However, the median class frequency is used to compute these coefficients, which is irrelevant in our case since we only have two classes. Therefore, we set the background class weighting coefficient to 1 and test several road class weighting coefficients taken in the interval $W = [1, 1/f_{\text{road}}]$, where $f_{\text{road}}$ is the ratio of road pixels over total pixels in the entire ground truth. The MSE loss thus becomes

$$\text{Loss}_{\text{MSE}}(Y_{\text{tol}}, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2 \tag{1}$$

where $y_i$ is the value in the tolerant ground truth, $Y_{\text{tol}}$, and $\hat{y}_i$ is the sigmoid value in the predictions, $\hat{Y}$, for pixel $i$. The number of pixels in the image is given as $N$ and the loss weighting coefficient $w_i$ for pixel $i$ is defined as

$$w_i = \begin{cases} \lambda & \text{if pixel } i \text{ is 1 (road) in } Y_{\text{bin}} \\ 1 & \text{if pixel } i \text{ is 0 (background) in } Y_{\text{bin}} \end{cases}$$

where $\lambda$ is a fixed value taken from the interval $W$ and $Y_{\text{bin}}$ is the binary ground truth.

### C. Applying Preprocessing and Postprocessing

We study the effect of two operations commonly applied to similar tasks: nonlocal (NL) filtering of SAR images [17] and segmentation postprocessing with fully connected CRFs (FCRFs) [18]. NL filtering improves the overall feature homogeneity in SAR images, often mitigating the negative effects of speckle noise. FCRFs have been very successful in improving the consistency of FCNN segmentation maps, especially by refining the borders between object regions. They optimize an energy function combining two spatial- and color-based correlation potentials in order to remove the inconsistent predictions and refine the correct ones. They can be extremely valuable since road segmentation is very sensitive to object smoothness.

## III. Experiments

### A. Experimental Procedure

*1) Data Set:* To the best of our knowledge, there is no publicly available data set suitable for our study case. We created our own data set using high-resolution TerraSAR-X images acquired in spotlight mode (see Table I). We identified the roads as either major roads, country roads, or dirt paths with the help of Google Earth optical images. Each road type was assigned a specific label thickness, best matching their respective outline thickness overall. The masks for all road types were merged into a binary ground truth that was then smoothed, as explained in Section II-B. However, manually labeling roads in urban areas was impractical: most objects were either difficult to distinguish or very similar to roads but of a different nature, such as building edges. For this reason, we selected regions with fairly dense road networks and very few cities from which we removed all urban areas. We used a land segmentation map[1] to delimit and mask out most cities, then manually removed the remaining ones.

TABLE I
METADATA OF THE TERRASAR-X IMAGES USED IN OUR DATA SET

| | |
|---|---|
| **Lincoln, England** | |
| Size, Ground Sample Distance | 20480*12288 px, 1.25 m/px |
| Projection Coordinate System | WGS 84 / UTM zone 30N |
| Coordinates Top-Left | [683056.875, 5931158.125] |
| Coordinates Bottom-Right | [698416.875, 5905558.125] |
| Reference Time UTC | 2009-12-27T06:25:21.938000Z |
| **Kalisz, Poland** | |
| Size, Ground Sample Distance | 4000*4000 px, 1.25 m/px |
| Projection Coordinate System | WGS 84 / UTM zone 34N |
| Coordinates Top-Left | [316607.000, 5720181.000] |
| Coordinates Bottom-Right | [321607.000, 5715181.000] |
| Reference Time UTC | 2009-04-12T04:59:32.920000Z |
| **Bonn, Germany** | |
| Size, Ground Sample Distance | 3600*4080 px, 1.25 m/px |
| Projection Coordinate System | WGS 84 / UTM zone 32N |
| Coordinates Top-Left | [356400.000, 5630000.000] |
| Coordinates Bottom-Right | [361500.000, 5625500.000] |
| Reference Time UTC | 2009-01-22T05:51:25.023344Z |

*2) Training:* We implemented the networks using Tensorflow 1.4 and trained them on a single NVIDIA Titan X Pascal. All networks were trained from scratch, as we noticed a considerable performance drop when using weights pretrained on ImageNet, certainly due to the different nature of SAR images compared to optical ones. The convolutional weights were initialized with He uniform distributions,[2] introduced in [19] the deconvolutional weights with bilinear filters and the biases with zeros. We used an ADAM optimizer with a learning rate of $5e-4$ and an exponential learning rate decay of 0.90 applied after each epoch. When not using weights pretrained on three-channel RGB images, one-channel SAR images could be used as input since the weights were initialized accordingly. The area of Lincoln was split into a training and test set as follows: the upper 80% of the image ($16\,384 \times 12\,288$ pixels) was used for training and the lower 20% ($4096 \times 12\,288$) for testing. The images from Kalisz and Bonn were used as additional test sets. The input data were normalized and data augmentation was performed on the training set with patch rotations (0°, 90°, 180°, and 270°) and horizontal and vertical flips. The augmented training set is composed of 12 288 patches, referred to as the epoch data.

*3) Evaluation Metrics:* For the evaluation, the predictions were thresholded at 0.5 to obtain a binary mask, which was then compared to the binary ground truth. We evaluated the performance of our models by computing the Intersection over Union (IoU) ($\text{TP}/\text{TP}+\text{FP}+\text{FN}$), the precision ($\text{TP}/\text{TP}+\text{FP}$), and the recall ($\text{TP}/\text{TP}+\text{FN}$), where TP, FP, TN, and FN denote the total number of true positives, false positives, true negatives, and false negatives for the road predictions, respectively. The IoU is a robust metric for segmentation quality assessment since it yields the overlapping ratio between predictions and labels (intersection) over their total surface (union). If the predictions match the labels well and do not extend outside of them, the IoU score will be high. Coupled with the precision (prediction correctness) and

[1] http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012
[2] www.tensorflow.org/api_docs/python/tf/keras/initializers/he_uniform

TABLE II
PERFORMANCE OF FCN-8s OVER THE TEST AREA IN LINCOLN

| $t_{max}$ | Loss weight | IoU | Precision | Recall |
|---|---|---|---|---|
| 0 px | 1 | 43.79% | **71.69**% | 52.94% |
| 1 px | 1 | 44.44% | 70.68% | 54.48% |
| 2 px | 1 | 44.93% | 69.45% | 56.00% |
| 4 px | 1 | 44.98% | 62.96% | 61.16% |
| 8 px | 1 | 42.92% | 54.72% | 66.56% |
| 4 px | 2 | **45.46**% | 65.34% | 59.91% |
| 4 px | 4 | 45.21% | 57.96% | 67.27% |
| 4 px | 8 | 43.73% | 51.13% | **75.17**% |

TABLE III
IOU SCORES OF THE BEST MODELS OVER THREE TEST AREAS

| Area | FCN-8s | Deep Res. U-Net | DeepLabv3+ |
|---|---|---|---|
| Lincoln | 45.46% | 40.18% | **45.64**% |
| Kalisz | 43.85% | 27.31% | **44.66**% |
| Bonn | **42.57**% | 35.90% | 40.91% |

recall (prediction completeness), we can accurately assess the performance of a model. Although very common in computer vision, the accuracy metric (TP + TN/TP + FP + FN + TN) is unsuitable for our study case. Since roads make up for around 5% of the pixels in our ground truth, 95% of accuracy could mean that only background was predicted.

### *B. Discussion*

To setup a baseline performance, we optimize our hyperparameters $t_{\max}$ and the loss weighting coefficient on FCN-8s and present the results on the test area from Lincoln (see Table II). We then train a Deep Residual U-Net model and a DeepLabv3+ model using the best parameters found for $t_{\max}$ and the weighted loss function and compare their results with the corresponding FCN-8s model across all our test images.

*1) Adapting the Spatial Tolerance $t_{max}$:* We test the following tolerance values: 0, 1, 2, 4, and 8 pixels. As anticipated, the greater the tolerance, the better the ground truth coverage (+13% recall between 0 and 8 pixels of tolerance) at the cost of a larger loss in precision (−17%). The best model reaches 44.98% IoU for $t_{\max} = 4$ pixels. There is a compromise between precision and recall with a loss below 10% in precision for a gain of 8% in recall compared to the model with $t_{\max} = 0$ pixels. We maintain $t_{\max} = 4$ pixels for the rest of the experiments.

*2) Adjusting the Loss Weighting:* Around 5% of the pixels in the ground truth are roads (inverse frequency: $1/0.05 = 20$); therefore, we experiment on the following loss weighting coefficients: 1, 2, 4, and 8. Loss weighting induces a maximum gain of 0.48% IoU with a coefficient of 2, reaching a value of 45.46%.

*3) Applying NL-Filtering:* The results become worse when using NL-filtered SAR images with a considerable decrease of 8% in IoU. A plausible explanation is that FCNNs natively apply spatial filtering through downsampling and convolutional filtering, so NL-filtering discards meaningful information.

*4) Applying FCRFs Postprocessing:* Contrary to our expectations, FCRFs fail to improve the connectivity of severed predicted road sections. In our case, the overall result is close to that of an erosion operation, removing not only spurious predictions but also valid ones. Moreover, the segmentation is already smooth and regular, limiting the benefits of FCRFs.

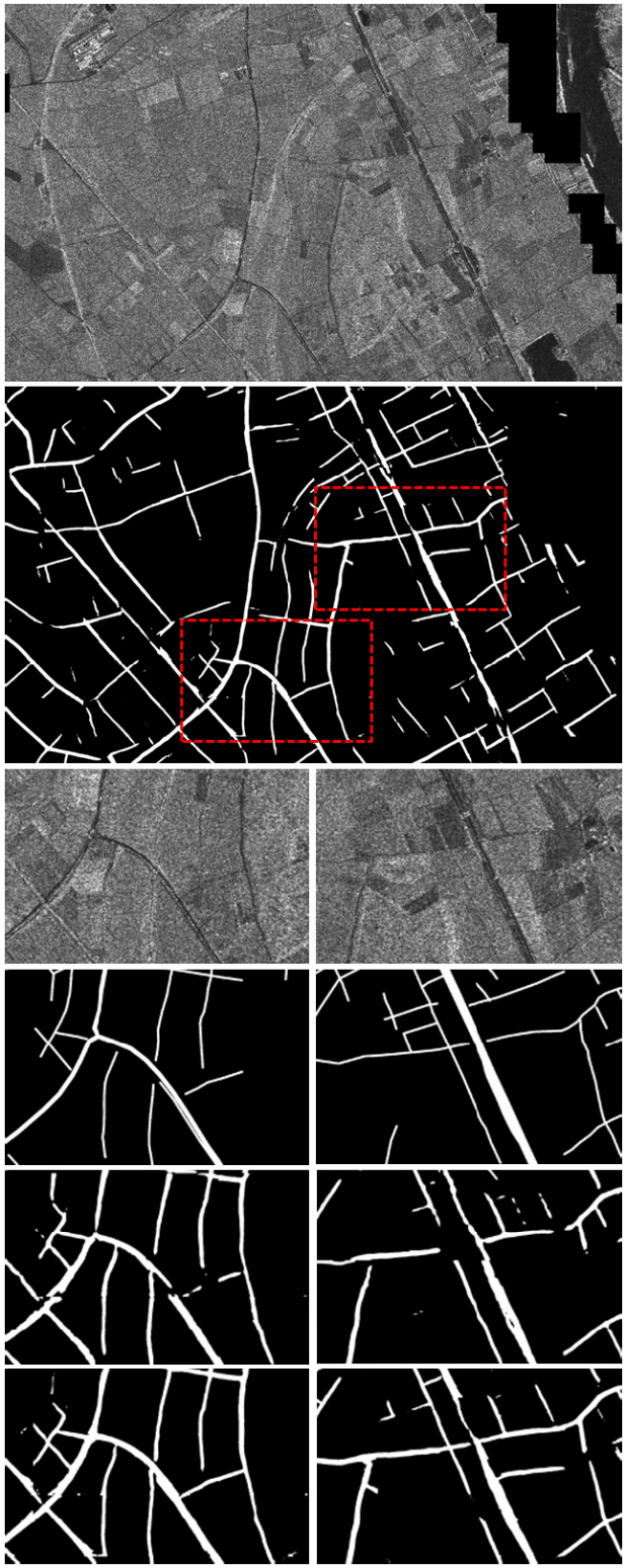

Fig. 2. Segmentation results of FCN-8s and DeepLabv3+ over the area of Bonn. (Top to Bottom) SAR image with masked urban areas, DeepLabv3+ predictions, and zoomed-in samples; in the samples, SAR image, ground truth, FCN-8s predictions, and DeepLabv3+ predictions.

*5) Additional FCNNs and Test Images:* We train a Deep Residual U-Net model and a DeepLabv3+ model with $t_{\max} = 4$ pixel and a loss weighting coefficient of 2. We report

the results over the full test set in Table III to compare their performance with FCN-8s and assess the generalization capacity of each architecture. On the one hand, Deep Residual U-Net shows surprisingly low performance compared to the other models. Like DeepLab, this network propagates the noisy SAR data down to deep layers but, unlike DeepLab, does not have a sufficient depth to abstract from it. On the other hand, DeepLabv3+ and FCN-8s achieve very close scores for all test images. Their performance is moderately reduced (maximum of $-4.73\%$ IoU) when applied to images from another region with respect to the training area. However, we find out that DeepLabv3+ converges 2.4 times faster than FCN-8s and produces far smoother and less noisy road predictions. A visualization of the segmentation of DeepLabv3+ over the area of Bonn is shown in Fig. 2.

*6) Limits of the Method:* Our models achieving the best results had difficulties generalizing over a wide variety of patterns, predicting unexpected objects, such as mounds and forest borders and missing many roads, mostly the less visible ones. A visual inspection of the results shows the limits of the proposed annotation scheme, as the label thickness based on road types does not reflect the actual thickness of many roads. Many prediction failures are due to this shortcoming. To improve the ground truth, a specific label thickness must be set for each individual road object. Besides, polygonal chain labels do not capture perfectly irregular road borders, which the predictions match more closely. In this specific regard, the models outperform the ground truth in terms of pixelwise correspondence to the roads but are yet penalized in the metrics. Consequently, straightening the predicted roads would make them coincide better with the labels. In parallel, because of the absence of object awareness in FCNNs, predicted roads are sometimes disconnected at intersections. The next step after road candidate extraction is the construction of a road graph that can be optimized to reconnect loose segments to each other. This is, however, outside the scope of this letter.

*7) Strengths of the Method:* The proposed method overcomes the major difficulty of isolating thin objects in a speckled environment and detecting many road patterns despite significant visual differences. The FCNNs were trained using a small data set, relatively to other data sets used for deep learning. They succeeded nonetheless, not only for the area they were fine-tuned on (Lincoln) but also for other completely unrelated areas (Kalisz and Bonn). FCNNs show an encouraging potential for adaptation given the complexity of the task at hand and would undoubtedly benefit from further training on additional images. Moreover, the predictions are smooth for the most part continuous and almost entirely free of noise, showing that the FCNNs successfully leverage the image wide context to improve the consistency of local predictions. The construction of a road graph can therefore be applied without any preprocessing on the road candidates, as they already constitute a solid baseline segmentation.

## IV. Conclusion

FCNNs prove to be an effective solution to perform road extraction from SAR images. We establish that off-the-shelf FCNNs can be substantially enhanced specifically for road segmentation by adding a tolerance rule toward spatially small mistakes. Our version of Deeplabv3+ modified with an mse regression loss, rebalanced toward the road class, achieves, in average, 44% IoU across our test sets. We also show that FCN-8s, no longer the state of the art, reaches scores very close to those of DeepLabv3+ while being far shallower. However, FCN-8s' predictions are more noisy and less smooth, making DeepLabv3+ a more robust road candidate extractor. This narrow performance gap points out the need to design new FCNN architectures specialized for road segmentation. The use of FCNNs as highly adaptable road candidate extractors should provide future works with a reliable means to obtain prior segmentations, on which graph reconstruction can be applied to map entire road networks.

## References


[1] F. Tupin, H. Maitre, J.-F. Mangin, J.-M. Nicolas, and E. Pechersky, "Detection of linear features in SAR images: Application to road network extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 2, pp. 434–453, Mar. 1998.

[2] T. Perciano, F. Tupin, R. Hirata, and R. M. Cesar, "A hierarchical Markov random field for road network extraction and its application with optical and SAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 1159–1162.

[3] R. Xu, C. He, X. Liu, D. Chen, and Q. Qin, "Bayesian fusion of multi-scale detectors for road extraction from SAR images," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 1, Jan. 2017, Art. no. 26. [Online]. Available: http://www.mdpi.com/2220-9964/6/1/26

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[5] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[6] W. Yao, D. Marmanis, and M. Datcu, "Semantic segmentation using the fully convolutional networks for SAR and optical image pairs," in *Proc. Conf. Big Data Space*, Toulouse, France, 2017, pp. 289–292.

[7] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[8] J. Geng, H. Wang, J. Fan, and X. Ma, "SAR image classification via deep recurrent encoding neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2255–2269, Apr. 2018.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (Feb. 2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1802.02611

[10] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. CVPR*, San Francisco, CA, USA, Jun. 2010, pp. 2528–2535.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 730–734. [Online]. Available: https://ieeexplore.ieee.org/document/7486599/

[12] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[14] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports field localization via deep structured models," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 4012–4020. [Online]. Available: https://ieeexplore.ieee.org/document/8099910/

[15] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 5695–5703.

[16] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. ICCV*, Santiago, Chile, Dec. 2015, pp. 2650–2658.

[17] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, San Diego, CA, USA, Jun. 2005, pp. 60–65.

[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. NIPS*, Granada, Spain, 2011, pp. 109–117. [Online]. Available: https://papers.nips.cc/paper/4296-efficient-inference-in-fully-connected-crfs-with-gaussian-edge-potentials

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