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 expand, 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 SAR images as they look visually similar to other objects like rivers and railways. Deep convolutional neural networks have been very successful in object segmentation, yet no method was developed to extract entire road networks from SAR images. This letter proposes a method based on a Fully-Convolutional Neural Network (FCNN) adjusted for road segmentation in SAR images. We study two approaches, binary segmentation and regression, intolerant and tolerant to prediction errors, respectively. The segmentation consistency is improved by applying Fully-connected Conditional Random Fields (FCRFs). We also share insights on creating a suitable dataset to facilitate future research. Our FCNN model shows promising results, successfully extracting 57% of the roads in our test dataset. We find out that the erosion effect of the FCRFs can effectively remove incoherent predictions, but is detrimental to road interconnections. The predicted roads have smooth borders yet oscillating shapes, hence regularization would help improving their straightness and connectivity.**

***Index Terms*—Feature extraction, synthetic aperture radar, neural networks**

I. Introduction

THE overall urban growth in the past two decades has led to a considerable development of transportation networks. This constantly evolving infrastructure necessitates frequent updates of the 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 scan an area's topography. The resulting physical terrain information is more resistant than optical imagery to changes in exposition and color. Moreover, SAR sensors can operate independently from all weather conditions, a major advantage when surveying a region affected by a weather-related disaster. The present study focuses on the extraction of roads in SAR satellite images.

We build a solution based on one of the most successful deep learning techniques recently: Deep Convolutional Neural Networks (DCNNs), which first demonstrated unmatched effectiveness in image analysis in 2012 [1]. While many DCNN architectures specialize in image classification (predicting a single label from an image: plane, car, ship, etc.) [2] [3], others achieved state-of-the-art performance in remote sensing tasks like semantic labeling on aerial imagery [4] [5].

In order to maximize the spatial accuracy of the extraction, we perform pixel-wise segmentation with Fully-Convolutional Neural Networks (FCNNs). Introduced in 2015, FCN8s [6] established a new state-of-the-art on the PASCAL VOC 2012 challenge [7], one of the most popular benchmark datasets for semantic segmentation methods. It consists of medium-sized optical photographs annotated with the outlines and classes of all visible objects. However in our case, we analyze high resolution satellite images often covering hundreds of square kilometers. FCNNs possess a substantial advantage in this regard since they can take input images of any size, producing identically-sized prediction maps. They were already successfully applied to road segmentation on optical images [8] [4] and SAR images [5]. In the latter though, only the highly visible roads were labeled, thus leaving aside most of the smaller ones like country roads or dirt paths.

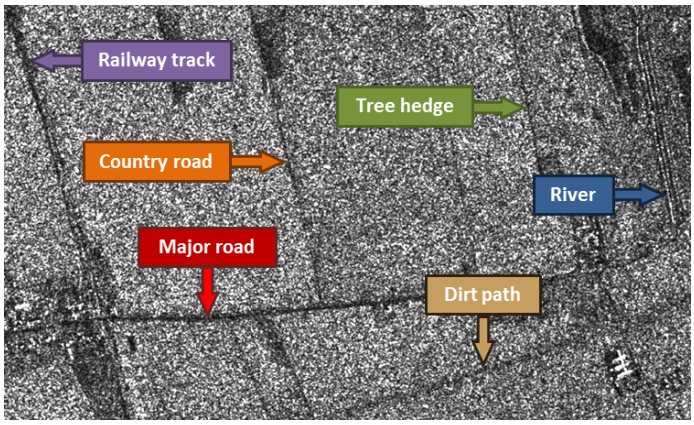

Fig. 1. SAR image sample showing that objects of different natures can look very similar. A segmentation model must learn to distinguish all kinds of roads from railway tracks, tree hedges and rivers.

Roads are difficult to identify in SAR images. They are generally characterized by thin dark lines, although significant feature disparities can be observed. Moreover, many other objects like railway tracks, rivers or even tree hedges look very similar to them, as illustrated in fig. 1. Labeling roads often involves the opinion of an expert and utmost precision when tracing their outlines. Likewise, the predicted outlines must meet the same accuracy requirements, but FCNNs are known to be approximate around object boundaries. This is especially problematic when extracting small objects like roads. To solve this issue, Krähenbühl and Koltun proposed using Fully-connected Conditional Random Fields (FCRFs) [9], a post-process tool leveraging the image wide context to improve the border smoothness and overall consistency of pixel-wise segmentations.

This letter presents a network based on FCN8s and specifically refitted for road segmentation on high resolution SAR satellite images. Two approaches are confronted: binary segmentation, with zero spatial tolerance for failure, and regres-

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

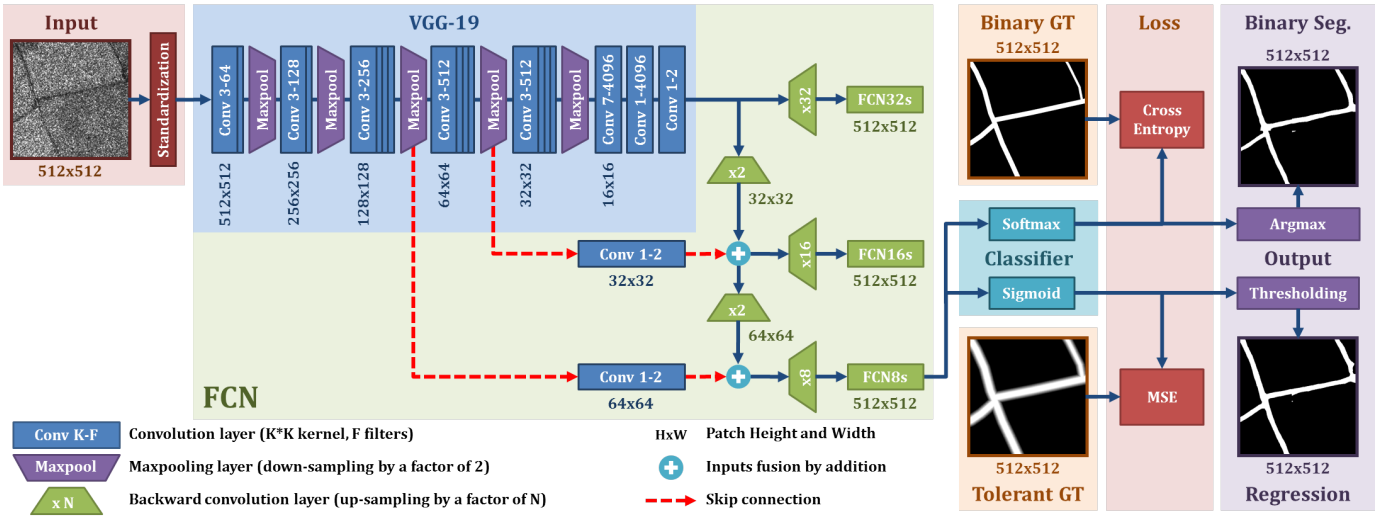

Fig. 2. An overview of the proposed method. All FCN versions are shown but only FCN8s is used.

sion, with an adjustable tolerance. FCRFs post-processing is applied to improve the visual quality of the segmentation. The performance of the different models is evaluated on a custom dataset. Its design is critical to the success of the method and is therefore detailed. We obtain satisfying results and demonstrate the effectiveness of FCNNs in road segmentation in SAR images. Most of the visible roads were accurately labeled. Suggestions for improvements are made as we discuss the achievements of this initial method proposal.

II. Method

A. Segmentation with Fully-Convolutional Neural Networks

We implement an FCNN architecture to segment roads at pixel level in SAR images. FCNNs can be given images of any size, thus can take into account a large 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, analyzes the images and outputs a cluster of predictions. The image data is gradually down-sampled by pooling layers, proportionately becoming more meaningful. The last pooling layer output is natively classified by fixed-size fully-connected layers, the network's bottleneck, imposing upstream a maximum input size. This size constraint is removed in FCNNs by replacing the fully-connected layers by convolutional layers. VGG-19 [2] is the backbone DCNN of our network. Since it down-samples an image by a factor of 32, its output cannot be visually interpreted. VGG-19's predictions are consequently processed by the second component of the FCNN, the up-sampling network. It restores the spatial properties of the predictions using backward convolution layers (commonly called deconvolution layers [10]) until the predictions share the same size as the input image. We use the FCN8s architecture proposed in [6]. This specific version of the FCN infuses the results from two intermediary layers of VGG-19 into the up-sampling process through skip-connections (see fig. 2). These layers have a finer prediction resolution than the DCNN output and help improving the segmentation accuracy over other FCN versions (FCN32s and FCN16s). FCN32s directly up-samples VGG-19's output 32 times, resulting in a coarse segmentation, while FCN16s fuses only one layer. By fusing an additional layer, FCN8s has a finer accuracy level (see prediction samples in fig. 3).

We propose two different setups to classify FCN8s' output. On the one hand, a binary segmentation with zero spatial tolerance. Each pixel in the ground truth must be classified either as road or as background. On the other hand, a regression with an adjustable spatial tolerance. A smooth target distribution is centered around the road labels, with maximal values of 1 on the road labels, linearly decreasing to 0 until a fixed distance, as proposed in [11]. The need for tolerance emerges from the fact that penalizing the network for predicting a road only a few pixels away from the label can be detrimental to the training effectiveness. Subsequently, the raw predictions are classified either by a softmax function, when doing binary segmentation, or by a sigmoid function, when doing regression (see fig. 2). Both produce a map where each data point carries a confidence score between 0 and 1. The closer to 1 it is, the more likely the pixel belongs to a road, and conversely for the background. The softmax function actually produces two confidence scores for each pixel such that $p_{i,road} + p_{i,background} = 1$ where $p_{i,c}$ is the probability that the class c is predicted for the i-th pixel. Finally, the confidence map is thresholded to obtain a binary segmentation map. For the softmax output, an argmax function selects the class with the highest confidence score for each pixel. For the sigmoid output we use a simple 0.5 threshold value.

B. Reducing the impacts of the class imbalance

Roads appear as thin objects in the SAR images and are likely outweighed by the background class, especially outside cities. We take necessary steps to limit the class imbalance during training. A similar problem in the case of sports field lines extraction is addressed in [12] 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. In the same way, increasing the spatial tolerance of the regression model helps

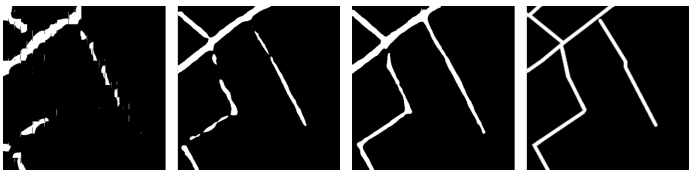

Fig. 3. Segmentation accuracy comparison between three FCN versions (left to right: FCN32s, FCN16s, FCN8s, ground truth)

reducing the imbalance. Eigen and Fergus [13] also proposed to remedy this issue by reweighing 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 weighting coefficient to 1 and experiment with several road weighting coefficients taken in the interval $W = [1, 1/f_{road}]$ where f_{road} is the ratio of road pixels over total pixels in the ground truth. We compute two different losses depending on the model type. On the one hand, the Cross-Entropy loss (CE):

$$Loss_{CE}(Y_{bin}, \hat{Y}_{bin}) = -\frac{1}{n}\sum_{i=1}^{n}\left(w_i \sum_{k} y_{i,k}\log(\hat{y}_{i,k})\right) \quad (1)$$

where $y_{i,k}$ and $\hat{y}_{i,k}$ are the label in the binary ground truth Y_{bin} and the softmax value in the binary segmentation predictions $\hat{Y}_{bin}$ for class k of pixel i, respectively. On the other hand, the Mean Squared Error loss (MSE):

$$Loss_{MSE}(Y_{tol}, \hat{Y}_{reg}) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 \quad (2)$$

where y_i and $\hat{y}_i$ are the label in the tolerant ground truth Y_{tol} and the sigmoid value in the regression predictions $\hat{Y}_{reg}$ for pixel i, respectively. The number of pixels in the image is given as n and the loss weighting coefficient w_i for pixel i is defined for both losses as:

$$w_i = \begin{cases} \lambda & \text{if pixel } i \text{ is labeled as a road in } Y_{bin} \\ 1 & \text{if pixel } i \text{ is labeled as background in } Y_{bin} \end{cases}$$

where λ is a fixed value taken from the interval W.

C. Post-processing

FCNNs lack precision around object boundaries, yet roads are thin and expected to be smooth and continuous: the predictions must be refined. We use Fully-connected Conditional Random Fields (FCRFs) [9] to solve this issue. FCRFs can enhance region boundaries on segmentation maps and were already successfully employed in combination with FCNNs [14]. Conditional Random Fields (CRFs) have been classically used to refine segmentation predictions [15]. FCRFs improve on CRFs by using the global context instead a local one, comparing two by two all pixels in an image. This post-processing tool must be trained to minimize an energy function defined in [9] and constituted of two potentials. First, the unary potential which penalizes any uncertainty in the prediction. Second, the bilateral pairwise potential which evaluates proximity in color and position for all pixel pairs in both the predictions and the input image. FCRFs have an erosion effect on the predictions. Since roads are already thin objects, they might be narrowed and even disconnected from each other in the process. Therefore, we apply the FCRFs on the background predictions. Grinding the background can fill the gaps between the roads and helps reconnecting them. For this purpose, we invert the input values given to the FCRFs: the softmax or sigmoid values are transformed into $(1.00 - \text{values})$. The resulting segmentation map is subsequently inverted to obtain the refined road predictions.

III. Experiments

A. Experimental procedure

Dataset: to the best of our knowledge, there is no publicly available dataset suitable for our study case. We created our own dataset using a TerraSAR-X image with a ground sampling distance of 1.25 m acquired in spotlight mode. To visualize the underlying difficulty of the annotation task, a SAR image sample is presented in fig. 1. We classified the roads as major roads, country roads or dirt paths, using optical images from Google Earth to identify their categories. Each road type was assigned a specific label thickness. However, labeling roads in urban areas was impractical: most objects were either difficult to distinguish in the SAR images, or very similar to roads but of a different nature. For this reason, we selected a region of 400 km^2 in the countryside of Lincoln in England with a fairly dense road network and very few cities, from which we removed all urban areas. We used a land segmentation map to delimit and mask most cities, then manually erased the remaining ones (source: http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012). This 20480*12288 px image was sliced into 512*512 px patches. Our dataset is therefore composed of 960 patches, with 768 patches in the training set (85% of the dataset) and 192 patches in the test set.

Pre-processing: to deal with the speckle inherent to SAR images, we generated a smoother input image by applying a Non-Local (NL) filter [16], which improved the overall feature homogeneity. We investigated the impact of different combinations of these input images on the segmentation performance, using the unprocessed SAR image only, the NL-filtered SAR image only and both images concatenated. These images were standardized (pixel values centered around zero and normalized). Data augmentation was performed on the training set, with patch rotations (0°, 90°, 180° and 270°), horizontal and vertical flips. The augmented training set is composed of 12288 patches, referred to as the epoch data. The test set was not augmented.

Network training: we implemented the network using Tensorflow 1.2 and trained it on a single NVIDIA Titan X Pascal with the hyper-parameters visible in table I. The convergence on the test set was reached after four to six hours of training (ten to fifteen epochs).

Loss weighting: around 5% of the pixels in the ground truth are roads (inverse frequency: $1/0.05 = 20$), therefore we experimented on the following loss weighting coefficients: 1, 2, 4, 8 and 16.

TABLE I
NETWORK TRAINING HYPERPARAMETERS

Convolution filters initialization: Glorot uniform distribution [17] Convolution biases initialization: zero
Learning rate initialization: 5.0e-4 Learning rate decay rate: 90% (after each epoch)
ADAM optimizer [18] with the following parameters: $\beta_1 = 0.9$ (gradient's 1st moment) $\beta_2 = 0.999$ (gradient's 2nd raw moment) $\hat{\epsilon} = 1e-8$ (cf. equation in section II of [18])

Post-processing: we used the efficient FCRFs implementation of Krähenbühl and Koltun [9], wrapped as Python library and publicly available at https://github.com/lucasb-eyer/pydensecrf. To find the best parameters for the energy function, we applied successive grid searches. We performed only one training iteration on the FCRFs, since no change was visible in the refined predictions with further training.

Evaluation metrics: we evaluated the performance of our models by computing the Intersection over Union (IoU) ($\frac{TP}{TP+FP+FN}$), the precision ($\frac{TP}{TP+FP}$) and the recall ($\frac{TP}{TP+FN}$), 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 cover the labels well and do not overflow, the IoU score will be high. Coupled with the precision (prediction correctness) and recall (prediction completeness), we can assess accurately the performance of a model. Although very common in computer vision, the accuracy metric ($\frac{TP+TN}{TP+FP+FN+TN}$) 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

Combining different input images: the IoU scores obtained using the unprocessed SAR image dataset constitute our baseline results. The binary segmentation model achieves 40.15% IoU, surpassed by the regression model reaching 40.53% IoU. Using NL-filtered SAR images in the dataset induces considerable drops in performance: –1.7% to –2.0% IoU in the case of binary segmentation, –5.5% to –8.3% IoU in the case of regression (using both images concatenated and the NL-filtered image only, respectively). The fact that FCNNs are progressively pooling the input images (thus applying a filtering effect) might explain why the information carried by the NL-filtered image is redundant and eventually detrimental to the performance. As a result, only the unprocessed SAR images are used in the following experiments.

Adapting the spatial tolerance: setting a tolerance of 4 pixels shows best results, as reported in table II. As anticipated, the greater the tolerance, the better the ground truth coverage (+14% recall between 2px and 16px of tolerance), at the cost of a tremendous loss in precision (–27%). The best model (4 px of tolerance) finds a compromise between these two quality metrics, losing less than 4% of precision when gaining 7% of recall compared to the base model (2 px of tolerance). We maintain a tolerance of 4 pixels for the rest of the regression experiments.

TABLE II
PERFORMANCE COMPARISON WITH REGRESSION TOLERANCE VARIATION

Tolerance	IoU	Precision	Recall
2 px	38.06	**62.79**	49.15
4 px	**40.53**	59.10	56.33
8 px	39.13	51.68	61.70
16 px	29.38	35.54	**62.91**

TABLE III
PERFORMANCE IMPACT OF FCRFS POST-PROCESSING ON REGRESSION (TOLERANCE OF 4 PX, LOSS WEIGHTING COEFFICIENT OF 2)

Applying FCRFs	IoU	Precision	Recall
No	41.85	58.04	**60.00**
Yes	**42.42**	**62.27**	57.10

Adjusting the loss weighting: loss weighting is proven efficient, inducing a notable gain in IoU for both models: +0.54% for the CE loss and +1.32% for the MSE loss, respectively using 8 and 2 as loss weighting coefficients. However, a higher coefficient implies a reduced precision. Indeed, predicted roads tend to be thicker, thus overflowing beyond the ground truth labels and diminishing the IoU score. The regression model now performs significantly better with a lower coefficient than the binary segmentation model, with 41.85% IoU against 40.69% IoU. Therefore, we perform the next experiments on the regression only.

Applying FCRFs: the post-processing yields mixed improvements in segmentation quality. The best model gains 0.57% IoU and maintains both precision and recall around 60% (see table III). However, the FCRFs deeply alter the visual quality of the predictions. On the one hand, the small isolated predictions are mostly eliminated. We notice an improvement in the prediction connectivity on the most visible roads, where accidental cuts were seen before. On the other hand, these benefits are counterbalanced by many drawbacks. A great number of successfully predicted roads were erased or disconnected at crossroads. In addition, the expected border smoothing is not visible at all. Instead, the roads are considerably ground at their extremities, explaining part of the aforementioned disconnections. These observations suggest that FCRFs are not adapted to road extraction in SAR images due to object visibility issues. Although they refine the predictions around the most visible roads, they considerably erode them where they are less visible.

Limits of the method: as mentioned in section III-A, annotating all roads is a difficult task and prone to misclassification. On a SAR image, a road can be characterized by another object, such as the trees bordering it, thus looking considerably different from completely exposed roads. Annotating such ambiguous objects induces confusion in the model. As anticipated, our best model had difficulties generalizing over this wide variety of patterns (cf. table III and fig. 4), predicting unexpected objects (62.27% of precision) and missing many roads (57.10% of recall). It appears necessary to distinguish the roads in the ground truth depending on their exposition and visibility. Beyond object identification difficulties, the results reveal that setting a road label thickness based on its importance is not accurate enough, as roads in the same category differ in width. We advise using individual label

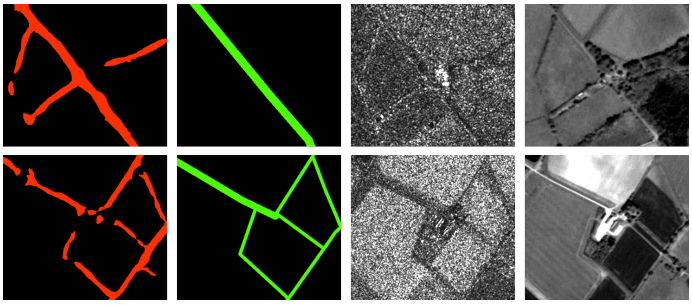

Fig. 4. Failure cases (left to right: FCRFs post-processed predictions, ground truth, SAR image, optical image). Row 1 shows wrong roads predictions over mounds and forest borders, well visible on the optical image. Row 2 reveals a striking case of predictions disconnected at a crossroad.

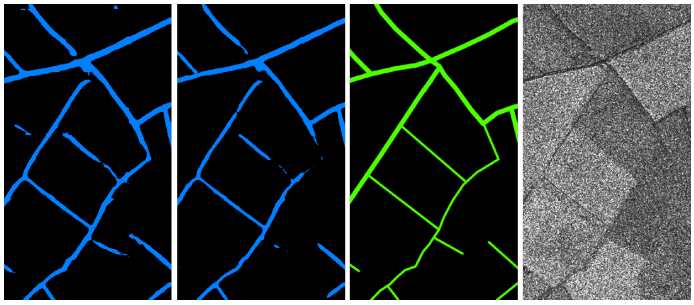

Fig. 5. Successful road network segmentation (left to right: FCNN predictions, FCRFs post-processed predictions, ground truth, SAR image)

thickness in the future. Unlike in optical images, roads can look irregular in SAR images. Thus, the straight labels in the ground truth do not perfectly fit the road outlines, more closely matched by the oscillating shape of the predictions. 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 FCRFs provide a solution for connecting close-by objects, however it does not guarantee that all gaps will be bridged. Existing connections might also be severed due to the erosion effect. Transforming the coarse predictions into regular shapes would also make it easier to detect intersections and reconnect the roads together.

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. In comparison to previous methods, FCNNs extract significantly more roads, in particular the country roads. The predictions are smoother, although undulating, and for the most part continuous, showing that FCNNs successfully leverage the image wide context to improve the consistency of local predictions. FCRFs showed their potential on top of FCNNs, thoroughly erasing abnormal predictions. A sample is shown in fig. 5 where successful predictions can be seen before and after the application of FCRFs post-processing. The predictions seem to correspond very well to the ground truth, especially before the post-processing. However, as shown by our results, slight offsets and shape irregularities cause considerable drops in IoU.

IV. Conclusion

Fully Convolutional Neural Networks prove to be an effective solution to perform road segmentation in SAR images. Our method reaches a baseline performance of 42.42% intersection over union on our test dataset with a model based on FCN8s. Almost all objects which were undoubtedly annotated as roads were extracted, while the uncertain ones were often partially predicted. We find out that the tolerance of small spatial mistakes is beneficial, thus showing the advantages of the regression over the binary classification. Fully-connected conditional random fields provide an efficient way of removing most aberrant predictions, but frequently disconnect the roads from each other. Future works are encouraged to convert the pixel-wise predictions into regular lines to improve the roads' shapes and interconnections. This study underlines the difficulty of road annotation in SAR images. Labeling roads covered by trees or characterized by other objects is necessary to obtain a complete segmentation but leads to confused models.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. NIPS*, Lake Tahoe, 2012, pp. 1097–1105.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, Las Vegas, 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. CVPR*, Honolulu, 2017, pp. 2261–2269.

[4] J. Sherrah, "Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery," *CoRR*, vol. abs/1606.0, 2016.

[5] 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, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[8] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense Semantic Labeling of Very-High-Resolution Aerial Imagery and LiDAR with Fully-Convolutional Neural Networks and Higher-Order CRFs," in *CVPR Workshops*, 2017, pp. 1561–1570.

[9] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *NIPS*, Granada, 2011, pp. 109–117.

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

[11] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient Deep Learning for Stereo Matching," in *CVPR*, 2016, pp. 5695–5703.

[12] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports Field Localization via Deep Structured Models," in *CVPR*, Honolulu, 2017, pp. 4012–4020.

[13] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture," in *ICCV*, 2015, pp. 2650–2658.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation With Deep Convolutional Nets And Fully Connected CRFs," *CoRR*, vol. abs/1412.7, 2014.

[15] N. Plath, M. Toussaint, and S. Nakajima, "Multi-class image segmentation using Conditional Random Fields and Global Classification," in *Proc. ICML*, Montreal, 2009, pp. 817–824.

[16] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, vol. 2, San Diego, 2005, pp. 60–65.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. AISTATS*, 2010, pp. 249–256.

[18] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.