INCORPORATING INTERFEROMETRIC COHERENCE INTO LULC CLASSIFICATION OF AIRBORNE POLSAR-IMAGES USING FULLY CONVOLUTIONAL NETWORKS

S. Schmitz[1, 2], M. Weinmann[1], A. Thiele[1, 2]

[1] Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT),
Englerstraße 7, 76131 Karlsruhe - (Sylvia.Schmitz, Martin.Weinmann, Antje.Thiele)@kit.edu
[2] Fraunhofer IOSB, Gutleuthausstraße 1, 76275 Ettlingen - (Sylvia.Schmitz, Antje.Thiele)@iosb.fraunhofer.de

Commission I, WG I/3

KEY WORDS: LULC classification, Airborne PolSAR, Interferometric Coherence, Fully Convolutional Network

ABSTRACT:

Inspired by the application of state-of-the-art Fully Convolutional Networks (FCNs) for the semantic segmentation of high-resolution optical imagery, recent works transfer this methodology successfully to pixel-wise land use and land cover (LULC) classification of PolSAR data. So far, mainly single PolSAR images are included in the FCN-based classification processes. To further increase classification accuracy, this paper presents an approach for integrating interferometric coherence derived from co-registered image pairs into a FCN-based classification framework. A network based on an encoder-decoder structure with two separated encoder branches is presented for this task. It extracts features from polarimetric backscattering intensities on the one hand and interferometric coherence on the other hand. Based on a joint representation of the complementary features pixel-wise classification is performed. To overcome the scarcity of labelled SAR data for training and testing, annotations are generated automatically by fusing available LULC products. Experimental evaluation is performed on high-resolution airborne SAR data, captured over the German Wadden Sea. The results demonstrate that the proposed model produces smooth and accurate classification maps. A comparison with a single-branch FCN model indicates that the appropriate integration of interferometric coherence enables the improvement of classification performance.

1. INTRODUCTION

The automatic analysis of Synthetic Aperture Radar (SAR) images, which can be acquired independently of cloud cover, weather conditions and daylight, allows the generation of up-to-date land use and land cover (LULC) maps. These maps provide an essential prerequisite for the efficient planning and management of urban and agricultural land use as well as for environmental monitoring. The task that underlies the generation of LULC maps is to semantically segment captured SAR images. In the course of strongly increasing data availability, particularly methods from the field of machine learning have proven to be suitable for this purpose. For example, the use of a Random Forest (RF) (van Beijma et al., 2014) or Support Vector Machines (SVMs) (Huang et al., 2002) achieve good results for pixel-based classification tasks.

However, the success of machine learning approaches strongly depend on the design and composition of suitable features. Typically, handcrafted low-level features are used, which often have the disadvantage of being location- and data-specific. Furthermore, such features are usually engineered for a particular task, which limits the ability to generalise to other requirements. These challenges are countered by methods from the field of Deep Learning (DL). DL methods have the ability to learn abstract, hierarchical features from raw data, thus eliminating the need for heuristic feature engineering and increasing generalisation performance. In addition, end-to-end training schemes allow task-specific outputs to be provided without expensive pre- and post-processing of data. For the task of semantic segmentation, particularly Fully Convolutional Networks (FCNs) have become established.

While FCNs have been used very successfully for LULC classification of optical imagery (Kampffmeyer et al., 2016; Fu et al., 2017; Mboga et al., 2019), the potential of this method for application to SAR data has not yet been fully exploited. Due to the intrinsic differences between the imaging mechanisms of SAR and optical images, FCNs that have been pre-trained on optical data do not achieve satisfactory results (Yao et al., 2017). In contrast, a complete training of FCNs from scratch with annotated SAR data is promising, because domain-specific low-level and high-level features can be learned. Therefore, this method can be successfully used for LULC classification of polarimetric SAR (PolSAR) images. For instance, Cao et al. (2019) introduced a complex-valued FCN designed for PolSAR image classification that outperforms conventional machine learning tools (e.g. RF and SVM). To distinguish several LULC classes, Li et al. (2018) proposed a sliding window FCN and reduced time and memory consumption by using sparse coding. Despite these encouraging results, further investigations are necessary to exploit the full potential of FCNs for pixel-wise LULC classification. Most previous work only employs information contained in single PolSAR images. In contrast, this work includes interferometric SAR measurements to further enhance classification performance. SAR interferometry has proven to be a valuable technique that allows the measurement of geophysical parameters such as surface topography or ground deformation. The central idea of InSAR is to gain information by comparing the phase of two radar images, which capture the same scene from slightly different positions at the same time (single-pass) or with a time offset (repeat-pass). An important measure relevant to LULC classification is the interferometric coherence, which quantifies the local phase correlation between the two complex images. As discussed in (Wegmüller, Werner, 1995), interferometric coherence provides complementary information to that

contained in backscattered intensities. For example, considering water surfaces, the backscattering coefficient can vary due to water movements caused by wind, while the interferometric coherence (in case of repeat-pass measurement) is consistently low because of temporal change of water surfaces. It is shown in various studies (Wegmüller, Werner, 1997; Abdelfattah, Nicolas, 2006; Mohammadimanesh et al., 2018) that combining backscattered intensities with interferometric coherence has the potential to significantly improve LULC classification. Hence, this paper addresses the questions of how to incorporate the complementary information contained by coherence images into FCN segmentation and to what extent the additional information improves the classification performance.

An existing challenge that still prevents the successful and widespread use of FCNs for SAR data analysis is the limited availability of densely labeled data. In most cases, data is manually labelled by experts in time-consuming processes as described for instance in (Mohammadimanesh et al., 2019). In contrast, this work investigates how automatically generated sparse and potentially noisy annotations based on the fusion of different available LULC products can be used for training. Particular attention is paid to a method which mitigates the negative influence of incorrectly labelled data.

This paper is organised as follows: in Section 2, an FCN architecture is introduced that combines backscattering coefficients and interferometric coherence to classify PolSAR images. Subsequently, details concerning the training of this network are described. In Section 3, experiments to evaluate the performance of the proposed FCN are outlined and the outcomes are presented in Section 4. Finally, in Section 5, results are summarised, conclusions are drawn and suggestions for future work are given.

2. METHODOLOGY

In the following, the FCN-based method for LULC classification using polarimetric SAR images is explained. First the generation of the input data is described followed by the architecture of the network and its training.

2.1 Input Image Generation

2.1.1 Pauli decomposition: To encode measurements of a polarimetric SAR system, the complex polarimetric scattering matrix

$$\mathbf{S} = \begin{bmatrix} s_{hh} & s_{hv} \\ s_{vh} & s_{vv} \end{bmatrix} \tag{1}$$

is used that describes the transformation between transmitted and received wave vectors caused by a scatterer (Lee, Pottier, 2017). The matrix $\mathbf{S}$ provides information about scattering processes of an observed object and thus about the object itself. However, it turns out to be difficult to derive physical properties of a scatterer directly from the matrix $\mathbf{S}$. In contrast, the Pauli decomposition of the scattering matrix allows the representation of polarimetric information that corresponds directly to physical scattering mechanisms of coherent targets. Assuming a monostatic system configuration that results in $s_{hv} = s_{vh}$, the decomposition of the scattering matrix based on the Pauli-basis

$$\mathbf{S}_a = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

$$\mathbf{S}_b = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{3}$$

$$\mathbf{S}_c = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{4}$$

is given by:

$$\mathbf{S} = a\mathbf{S}_a + b\mathbf{S}_b + c\mathbf{S}_c, \tag{5}$$

with

$$a = \frac{s_{hh} + s_{vv}}{\sqrt{2}} \tag{6}$$

$$b = \frac{s_{hh} - s_{vv}}{\sqrt{2}} \tag{7}$$

$$c = \sqrt{2}\, s_{hv}. \tag{8}$$

This decomposition subdivides the scattering matrix $\mathbf{S}$ into three components that refer to specific scattering mechanisms. The matrix $\mathbf{S}_a$ corresponds to single- or odd-bounce scattering, $\mathbf{S}_b$ represents double- or even-bounce scattering and $\mathbf{S}_c$ indicates a scattering mechanism characterised by volume scattering. The related complex coefficients a, b and c indicate the contribution of the corresponding matrices to the scattering matrix $\mathbf{S}$, whereas $|a|^2$, $|b|^2$ and $|c|^2$ express the scattered power by the associated types of target. To represent this information in a single three-channel image, denoted as Pauli-RGB image, the following codification is used:

$$|a|^2 \rightarrow Red \qquad |b|^2 \rightarrow Green \qquad |c|^2 \rightarrow Blue$$

In this work, the Pauli-RGB image is used as one input image of an FCN. Its rich texture and color features match the visual perception of the captured scene. Thus, spatial features can be extracted by the network that give a good indication of requested LULC classes.

2.1.2 Interferometric Coherence: Based on two co-registered complex SAR image values s_1 and s_2 the interferometric coherence is calculated by:

$$\gamma = \frac{\langle s_1 s_2^* \rangle}{\sqrt{\langle s_1 s_1^* \rangle \langle s_2 s_2^* \rangle}} \tag{9}$$

where * indicates complex conjugation and $\langle x \rangle$ denotes the expected value, which is commonly approximated by averaging adjacent pixels. The resulting correlation coefficient $|\gamma|$ has a value range from 0 indicating total decorrelation to 1 denoting complete conformity and depends on system and acquisition parameters as well as on structural parameters of the scatterer and temporal scene coherence. In this work, a coherence image is formed using two SAR images that are captured within repeat-pass acquisition, thus it is predominantly related to random changes of scatterers. The coherence image is used as second input of an FCN to include complementary information to that contained in the Pauli-RGB image.

2.2 Network Architecture

In order to realise pixel-wise LULC classification, we propose an FCN, denoted as Fused U-Net, shown in Figure 1, that relies

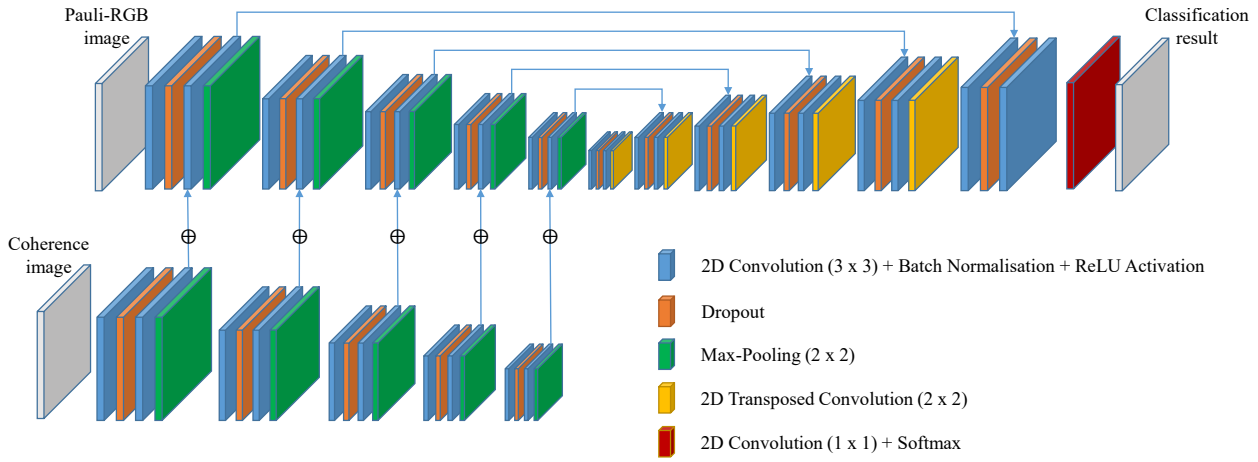

Figure 1. Architecture of the proposed Fused U-Net with Pauli-RGB image and coherence image as inputs.

on the generic encoder-decoder paradigm. Within the encoder stage, high-level features are extracted from the input layer, while the decoder stage is designed to consecutively up-sample the feature maps to the original input resolution.

To effectively combine the information contained in the Pauli-RGB image on the one hand and in the interferometric coherence image on the other hand, a network architecture inspired by the FuseNet structure (Hazirbas et al., 2016) is used. The fusion-based Convolutional Neural Network (CNN) architecture was originally developed to incorporate depth information into the semantic segmentation of RGB images. Following the same fusion approach, the proposed network comprises two encoder branches that are trained to extract and combine features from Pauli-RGB and coherence images. The three-channel Pauli-RGB image is taken as input for the main branch, while the single-channel coherence image is the input of the additional branch. Within both encoder branches, hierarchical feature maps are computed by a stack of convolution, batch normalisation and activation layers. Aggregation of feature maps is performed by max-pooling on five levels. The convolution operators act as image filters with trainable kernel weights and the Rectified Linear activation Unit (ReLU) function enables the learning of non-linear mappings. During batch normalisation, which is applied before each activation, feature maps are first normalised over a mini batch to have zero mean and unit variance. In order to maintain the expressivity of the model, the normalised values are scaled and shifted by two additional parameters that are learned along the training of the network. Batch normalisation enables a faster and more stable training process due to reduced internal covariate shift (Ioffe, Szegedy, 2015) and a smoother optimisation landscape (Santurkar et al., 2018). Furthermore, it provides regularisation effects and thus strengthens the model to better generalise to unseen examples. The trainable parameters of batch normalisation allow the network to learn internal representations of the Pauli-RGB and coherence images, which complement each other optimally in the following fusion steps.

The step-wise fusion is accomplished by adding feature maps extracted from the coherence image to feature maps derived from the Pauli-RGB image using element-wise summation before each max-pooling layer. In this way, feature maps in the main branch are enriched by features from the additional branch and a joined complementary representation is learned. The use of this fusion design, instead of simply stacking Pauli-RGB and coherence images within one input layer, is based on the assumption that the two distinct modalities require different sets of filters for the extraction of significant features. The separated encoder branches of Fused U-Net enable independent learning of features, specialised in the discriminant representation of information from the different data sources.

Within the common decoder part of the network, resulting fused feature maps of low spatial resolution are consecutively up-sampled by transposed convolution operations. In order to reduce the loss of information due to down-sampling in the encoder, five skip connections are used, which incorporate high-resolution feature maps from the encoder to the decoder stage. This concept was introduced in (Ronneberger et al., 2015) and is widely applied in many FCN approaches for semantic segmentation. By concatenating deep coarse features with shallow fine features, accurate detail information can be preserved. Five up-sampling blocks are followed by a 1×1 convolution layer that reduces the depth dimension of feature vectors to the desired number of output classes. To transform feature vectors, that each describe one pixel, into probabilities, the softmax function is applied. Final class labels that build up the intended segmented map are determined based on the highest probability values.

2.3 Network Training

The described network can be modelled by a chain of functions:

$$f(\mathrm{x};\mathbf{W}) = f^{(L)}(f^{(L-1)}(...f^{(2)}(f^{(1)}(\mathrm{x};\mathrm{w}^{(1)});\mathrm{w}^{(2)}... \mathrm{w}^{(L-1)});\mathrm{w}^{(L)}). \quad (10)$$

Here $f(\mathrm{x};\mathbf{W})$ denotes a feature vector of length K (number of classes) for a sample pixel x. The functions $f^{(1)},...,f^{(L)}$ describe the filtering and processing operations of L layers used in the network, which are parameterised by $\mathbf{W} = [\mathrm{w}^{(1)},...,\mathrm{w}^{(L)}]$. Given a K-class training set $\mathcal{D} = \{(\mathbf{X_1}_i, \mathbf{X_2}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, an optimal set of parameters $\mathbf{W}^*$ is determined during network training. Here $\mathbf{X_1}$ and $\mathbf{X_2}$, with $\mathbf{X_1} \in \mathbb{R}^{W\times H\times 3}$ and $\mathbf{X_2} \in \mathbb{R}^{W\times H}$, denote two input images (i.e. the Pauli-RGB and the corresponding coherence image); $\mathbf{Y}_i \in \mathcal{K}^{W\times H}$, with $\mathcal{K} = \{1,...,K\}$, denotes the associated ground-truth labeling. To find a parameter set $\mathbf{W}^*$, a suitable loss function is minimised that compares predicted class distributions resulting from

softmax mapping

$$p(k|\mathrm{x}) = \frac{\exp(f_k(\mathrm{x};\mathbf{W}))}{\sum_{i=1}^{K}\exp(f_i(\mathrm{x};\mathbf{W}))} \tag{11}$$

to corresponding one-hot encoded ground-truth distributions $q(k|\mathrm{x})$. A loss function that is commonly used in conjunction with neural networks, which apply softmax activations in the output layer, is the categorical cross entropy defined by:

$$\mathcal{L}_{cce} = -\sum_{k=1}^{K} q(k|\mathrm{x})\log p(k|\mathrm{x}). \tag{12}$$

However, in this work, the categorical cross entropy in its original form is not suitable, due to characteristics of employed training data that are described in the following.

To derive required ground-truth LULC class images $\mathbf{Y}_i$ and form a diverse and sufficiently large training data set $\mathcal{D}$ without the necessity of time-consuming manual labeling, PolSAR images are annotated automatically using the approach described in (Schmitz et al., 2020): Information from publicly available LULC products, namely OpenStreetMap, CORINE Land Cover 2018, and Global Water Surface, is extracted and fused with information that can be derived from the PolSAR image itself based on interferometric coherence and polarimetric features. As part of the automatic annotation process, class assignments are excluded that are not sufficiently reliable due to conflicting information of the various sources of input data. Thus, the resulting training data is not densely but only sparsely labeled. Despite filtering uncertain labels, the training data may contain incorrectly assigned class labels that have a negative impact on the network training. As empirically evaluated in (Wang et al., 2019), the cross entropy loss (Equation (12)) reveals weaknesses in the context of learning on erroneous training data. It is stated that, within the learning process, the network tends to overfit to noisy labels on "easy" classes, while the effect of under-learning occurs for "hard" classes. To overcome these limitations, a suitable loss function proposed in (Wang et al., 2019), based on symmetric learning, is implemented and used for the training of the Fused U-Net. The chosen symmetric entropy loss $\mathcal{L}_{sce}$, inspired by the symmetric KL-divergence, is defined as the weighted sum of cross entropy and reverse cross entropy loss:

$$\mathcal{L}_{sce} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{rce} \tag{13}$$

$$\text{with} \quad \mathcal{L}_{rce} = -\sum_{k=1}^{K} p(k|\mathrm{x})\log q(k|\mathrm{x}). \tag{14}$$

While the cross entropy term leads to good convergence, the additional reverse cross entropy term is noise tolerant. The hyperparameters α and β can be tuned to find a balance between the reduction of overfitting and speed of convergence.

During the training of the Fused U-Net, symmetric entropy loss values are calculated on each training pixel $\mathrm{x} \in \Omega$. Here, Ω denotes the set of pixels of the training image tuples $(\mathbf{X}_{1,i}, \mathbf{X}_{2,i})$ that have a valid label in $\mathbf{Y}_i$. Based on the loss values, the network parameters $\mathbf{W}$ are updated iteratively using the gradient-based Adam optimisation algorithm (Kingma, Ba, 2014). To stabilise the training, the update frequency is reduced by using mini batch gradient descent. Weight updates are performed based on averaged sample losses over a subset (called mini batch) of Ω.

3. EXPERIMENTS

3.1 Study Area

Within the framework of the GeoWAM project, which aims to generate high-resolution geodata for coastal monitoring, fully polarimetric SAR data are captured over the tide-influenced German Wadden Sea. The geographic location of the study area considered in this paper, namely Otzumer Balje, is illustrated in Figure 2. With the objective of creating an accurate model of the watercourse for the study area, the main focus of the classification is on the distinction between water and dry fallen mudflats. The foreshore and land area, which is less focused in this work, is merely divided into two classes, soil and non-soil. The soil class includes areas with low vegetation, crop land, meadows and roads, while the non-soil class includes human-made objects, built-up areas and forestation areas. For data acquisition, the F-SAR system developed at the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt; DLR) (Horn et al., 2009) was used. F-SAR is an airborne SAR system, equipped with multiple antennas that enable capturing fully polarimetric SAR data at different wavelengths. In this work, image data are employed that were recorded by the S-band antenna during a measurement campaign in July 2019. Interferometric measurements, needed for the calculation of coherence images, were performed with repeat-pass baselines in the order of 40 metres. At the time of acquisition, the tidal range was low, thus large areas of dry fallen mudflats are depicted in the SAR images.

3.2 Experimental Setup

Three co-registered complex PolSAR image pairs were selected for the training and testing of the Fused U-Net model. The respective Pauli-RGB images were calculated according to Equations (6) to (8) and subsequently projected from slant-range to ground-range geometry. To minimise the influence of varying incidence angles, a gamma-naught calibration were performed. The coherence images were derived from VV-polarised complex image pairs by applying Equation (9) using a Gaussian filter of size 11×11 with standard deviation $\sigma = 5.0$ to approximate the expected values. The resulting image was geocoded as well. Before the input images were fed into the network, the values within an image were normalised to [0, 1]. Since we expected the network to learn basic SAR-specific image filters, no additional filtering of the input images was carried out. As

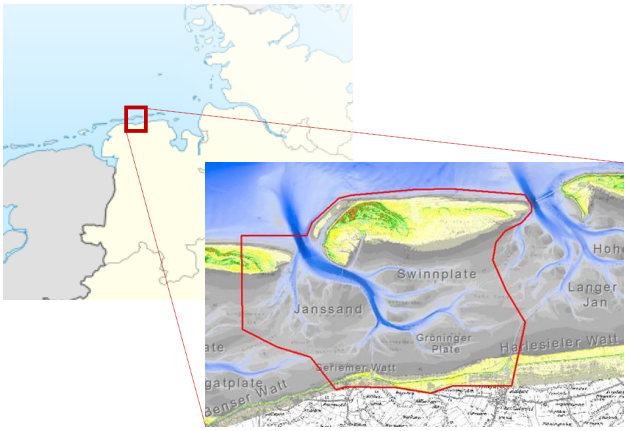

Figure 2. Geolocation of the study area Otzumer Balje, a tidal basin in the German Wadden Sea.

described in Section 2.3, reference images that contain the class labels including *water*, *mudflats*, *soil* and *non-soil* on pixel level were generated automatically. A quantitative evaluation of the accuracy of the labels that were generated in this way for the study area is given in (Schmitz et al., 2020). The data were divided into training and test data in a ratio of 70% to 30%, taking care to use geographically separate areas. While the reference images used for testing were manually post-processed to reduce faulty labels, there was no correction of the reference images used for training. In this way, it can be examined whether the model is robust against faulty training labels. Figure 3 illustrates exemplary sections of the resulting training and testing image triplets (Pauli-RGB, coherence and reference image).

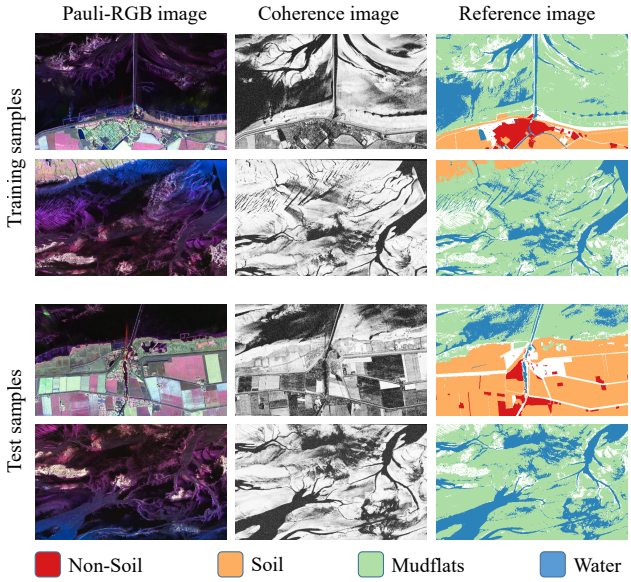

Figure 3. Exemplary image sections of training and testing data, including the Pauli-RGB image, the coherence image and the corresponding reference image containing class labels.

For the training of the Fused U-Net, images were divided into patches of size 512×512 pixels with an overlap of $25\,\%$. Image patches that contained less than $5\,\%$ annotated pixels were excluded from training. In order to counteract the imbalanced class distribution within the training data, random undersampling was performed to reduce the number of samples from dominant classes. Based on selected training image patches, an optimal set of model parameters was determined by minimising the loss (Equation (14)) using the keras implementation of the Adam optimiser. The optimisation started with a learning rate of 0.01 that was reduced by a factor of 0.5 every 10 epochs. The model parameters were updated iteratively after the evaluation of a mini batch of 8 patches.

To examine the extent to which the proposed way of inclusion of coherence images affects classification performance, we compared the proposed Fused U-Net model to a single-branch FCN with U-Net architecture. It follows the same basic structure as the presented Fused U-Net, but the additional encoder branch is omitted. This model was trained with two different configurations of the input layer. In the first approach, the input layer contained only the Pauli-RGB image, for the second approach, Pauli-RGB and coherence images were stacked to a 4-channel image. In the following, the two methods are referred to as Pauli U-Net and Pauli-Coh U-Net. The training was performed using the same strategy and data as that used for the Fused U-Net.

After training each network for 100 epochs, the resulting models were applied to the classification of the remaining unseen test data. Therefore, the corresponding image data was divided into overlapping patches of size 1024×1024 pixels that were fed into the neural networks. For each network, the predicted output maps were combined to form one classification image, whereby the overlap was used to eliminate artefacts occurring at the borders of single output patches. The resulting classification images provide the basis for the following performance evaluation.

4. RESULTS

For the evaluation and comparison of the classification performance of the different models, several metrics are considered. On pixel-level, the precision and recall rate as well as the F_1-Score that indicates the harmonic mean between precision and recall rate are determined for each class. These metrics are defined as follows:

$$\mathrm{F}_1 = 2\frac{\mathrm{r}_{\mathrm{precision}} \cdot \mathrm{r}_{\mathrm{recall}}}{\mathrm{r}_{\mathrm{precision}} + \mathrm{r}_{\mathrm{recall}}} \tag{15}$$

with

$$\mathrm{r}_{\mathrm{precision}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{16}$$

$$\mathrm{r}_{\mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{17}$$

where TP denotes the number of true positives; FP the number of false positives and FN the number of false negatives. The macro-average F_1-Score is determined for each model by equally weighting all class-specific scores. To assess the performance on region-level, the IoU is used, a similarity measure, which determines the degree of overlap between predicted classification and ground-truth masks defined as:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{18}$$

The mean IoU for each model is calculated by averaging the IoU over all classes. The achieved performance scores, based on the classification results for the test area, are summarised in Tables 1 and 2 for each model. The highest average F_1-Score as well as the best average IoU is provided by the Fused U-Net model with 0.9 and 0.82, respectively. The average performance of the Pauli-Coh U-Net is only slightly below, with an average F_1-Score of 0.88 and a mean IoU of 0.81. Comparing the class-specific performance, it can be seen that the achieved results for classification of water and mudflat are similarly good for both models. The poorer average performance of the Pauli-Coh U-Net is mainly due to a lower ability to recognize non-soil areas accurately, which is reflected in lower IoU, precision and recall rates. Apparently, the Pauli-Coh U-Net model performs worse compared to the Fused U-Net model in areas where the Pauli-RGB image provides better distinguishing features compared to the coherence image. This suggests that the simple stacking of features causes the learned filters to be mainly suitable for feature extraction from the coherence image. The prediction of the Pauli U-Net model that, in contrast to the Fused and Pauli-Coh U-Net, does not include interferometric coherence, shows the lowest F_1-Scores and average IoU. While the

Model	Precision				Recall			
	non-soil	soil	mudflat	water	non-soil	soil	mudflat	water
Pauli U-Net	0.46	**0.94**	0.82	0.61	0.81	0.93	0.62	0.82
Pauli-Coh U-Net	0.58	0.92	**0.95**	0.98	0.85	0.91	**0.96**	**0.95**
Fused U-Net	**0.72**	0.88	0.92	**0.99**	**0.91**	**0.94**	**0.96**	0.90

Table 1. Class-wise recall and precision rates for classification of test data achieved by the Pauli U-Net, Pauli-Coh U-Net and Fused U-Net. The best results are marked in bold.

Model	F_1-Score					Intersection over Union				
	non-soil	soil	mudflat	water	∅	non-soil	soil	mudflat	water	∅
Pauli U-Net	0.59	**0.94**	0.71	0.70	0.73	0.42	0.88	0.55	0.54	0.60
Pauli-Coh U-Net	0.69	0.91	**0.96**	**0.97**	0.88	0.53	**0.84**	**0.92**	**0.93**	0.81
Fused U-Net	**0.80**	0.91	0.94	0.95	**0.90**	**0.67**	**0.84**	0.89	0.90	**0.82**

Table 2. Class-wise and average F_1-Scores and Intersection over Union (IoU) for classification of test data achieved by the Pauli U-Net, Pauli-Coh U-Net and Fused U-Net. The best results are marked in bold.

classification of soil regions succeeds, the Pauli U-Net model fails to reliably recognize water and mudflat areas. As expected, this suggests that the inclusion of the coherence image has a clear benefit for classification performance, especially in the investigation of tidal-influenced areas, where the distinction between mudflats and water plays a crucial role.

In order to better interpret the quantitative results, achieved classification results are presented in Figure 4 for a few example regions of the test area. The visual comparison of the predictions leads to the following observations: As already indicated by the performance scores, the Pauli U-Net model, that uses only the Pauli-RGB image for classification, fails to accurately distinguish between water and mudflats. In the tidal basin, a large part of the mudflats are falsely classified as water. Consequently, the watercourse, which is mainly characterised by the course of tidal creeks, seaweeds and narrow water channels, cannot be accurately captured by the Pauli U-Net model. In contrast, the water and mudflat separation which is obtained by the Fused U-Net model and the Pauli-Coh U-Net model, matches the reference image very precisely. This can be explained by the fact that the watercourse is clearly visible in the coherence image. Coherence is low in water-covered areas, while the Wadden areas lead to high coherence values resulting in a high contrast that can be easily detected by convolution filters and is apparently learned by both models. Considering the classification results of the mainland area, the different smoothnesses of the three result images are clearly visible. Large continuous areas such as meadows and salt marshes are well captured by all models without contamination of speckle noise that is present in input images. Greater differences between the predictions are evident for urban areas, isolated farmyards and coastal structures. In the reference image used for training and testing, not every single building and object is labelled. Instead, urban regions are combined into one labelled region, which encloses buildings, streets, trees, etc.; farmyards include surrounding meadows and tiny coastal buildings are not labeled at all. The Fused U-Net and Pauli-Coh U-Net model are able to learn this kind of annotation as illustrated in the middle example of Figure 4. Thus, smooth and homogeneous predictions are generated that resemble the reference image. However, the effect of over-smoothing occurs. Individual buildings and small artificial objects are filtered out and the boundaries of urban areas and farmyards are not accurately segmented. On the contrary, the Pauli-Net generates less smoothed results, which on the one hand leads to misclassification of single pixels in actually connected areas, but on the other hand allows the segmentation of fine structures.

5. CONCLUSION

In this paper, Fused U-Net, a two-branch encoder-decoder network, was presented that combines polarimetric backscattering intensity and interferometric coherence and is trained to perform LULC classification based on SAR data. The model was developed and applied to classify water, mudflats, soil and non-soil areas in airborne S-band SAR data acquired over the German Wadden Sea. Instead of manually labelled accurate training data, training labels were automatically generated. In order to deal with few faulty labels that accompany the automatic label generation, the error-robust symmetrical cross entropy loss function was used for the training of the Fused U-Net. The experimental results demonstrate that the model trained in this manner achieves a fine-grained segmentation of the watercourse in the tidal area and smooth predictions in the mainland area. By comparison with similar network architectures, it is shown that the integration of interferometric coherence significantly increases the classification performance, especially for the separation of water and mudflats. This work serves as first proof of concept for the presented Fused U-Net model and its training on partly incorrect training data.

Future work will focus on the application of this approach for the classification of additional, more resembling classes. We assume that, for this case, the superiority of the presented method for the fusion of features extracted from different types of input data, as opposed to a stacking of different input data, will be even more evident. Furthermore, the results of this work suggest that a fine-tuning of the trained Fused U-Net model with a

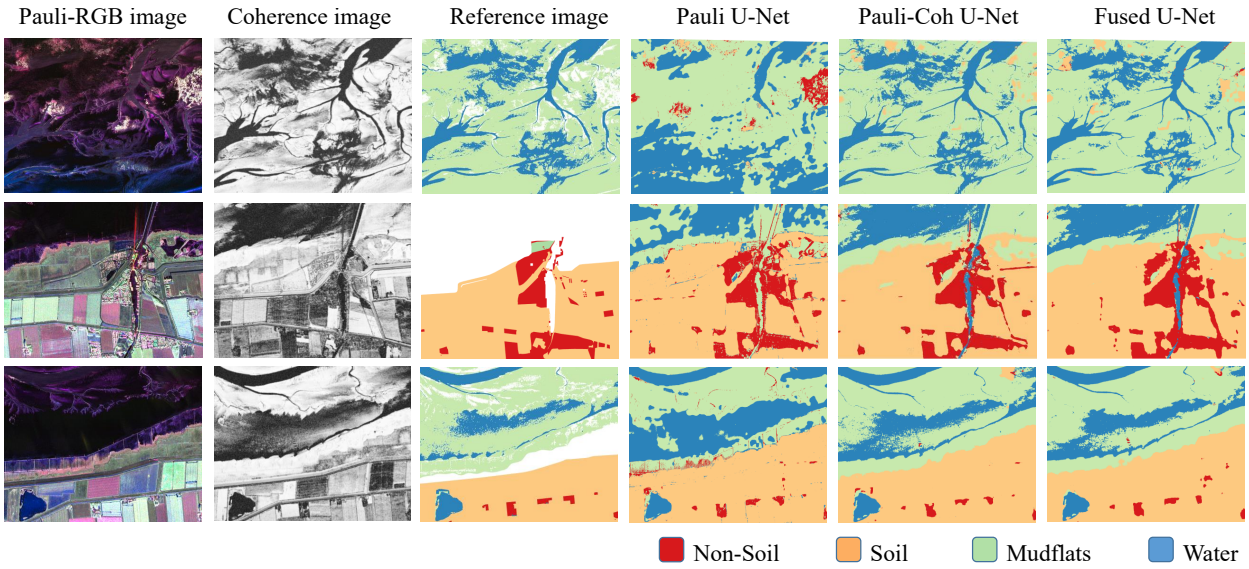

Figure 4. Predictions of the Pauli U-Net, Pauli-Coh U-Net and Fused U-Net model for exemplary sections of the test region.

few additional fine-grained manually labeled training data may allow the segmentation of single objects and fine-structured regions.

ACKNOWLEDGEMENTS

This study is part of the GeoWAM project that is funded by the German Federal Ministry of Transport and Digital Infrastructure within the framework of the Modernity Fund ("mFUND").

REFERENCES

Abdelfattah, R., Nicolas, J., 2006. Interferometric synthetic aperture radar coherence histogram analysis for land cover classification. *2006 2nd International Conference on Information & Communication Technologies*, 1, 343–348.

Cao, Y., Wu, Y., Zhang, P., Liang, W., Li, M., 2019. Pixel-wise PolSAR image classification via a novel complex-valued deep fully convolutional network. *Remote Sensing*, 11(22), 2653:1–2653:29.

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sensing*, 9(5), 498:1–498:21.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. *Asian Conference on Computer Vision*, 213–228.

Horn, R., Nottensteiner, A., Reigber, A., Fischer, J., Scheiber, R., 2009. F-SAR—DLR's new multifrequency polarimetric airborne SAR. *2009 IEEE International Geoscience and Remote Sensing Symposium*, 2, II:902–II:905.

Huang, C., Davis, L., Townshend, J., 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4), 725–749.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 680–688.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J.-S., Pottier, E., 2017. *Polarimetric radar imaging: from basics to applications*. CRC press.

Li, Y., Chen, Y., Liu, G., Jiao, L., 2018. A novel deep fully convolutional network for PolSAR image classification. *Remote Sensing*, 10(12), 1984:1–1984:17.

Mboga, N., Georganos, S., Grippa, T., Lennert, M., Vanhuysse, S., Wolff, E., 2019. Fully convolutional networks and geographic object-based image analysis for the classification of VHR imagery. *Remote Sensing*, 11(5), 597:1 – 597:17.

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Brisco, B., Motagh, M., 2018. Multi-temporal, multi-frequency, and multi-polarization coherence and SAR backscatter analysis of wetlands. *ISPRS Journal of Photogrammetry and Remote Sensing*, 142, 78–93.

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E., Molinier, M., 2019. A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS Journal of Photogrammetry and Remote Sensing*, 151, 223–236.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Santurkar, S., Tsipras, D., Ilyas, A., Madry, A., 2018. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 2483–2493.

Schmitz, S., Weinmann, M., Weidner, U., Hammer, H., Thiele, A., 2020. Automatic generation of training data for land use and land cover classification by fusing heterogeneous data sets.

Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V, 29, 73–86.

van Beijma, S., Comber, A., Lamb, A., 2014. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. *Remote Sensing of Environment*, 149, 118–129.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J., 2019. Symmetric cross entropy for robust learning with noisy labels. *IEEE International Conference on Computer Vision*, 322–330.

Wegmüller, U., Werner, C., 1995. SAR interferometric signatures of forest. *IEEE Transactions on Geoscience and Remote Sensing*, 33(5), 1153-1161.

Wegmüller, U., Werner, C., 1997. Retrieval of vegetation parameters with SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 35(1), 18–24.

Yao, W., Marmanis, D., Datcu, M., 2017. Semantic segmentation using deep neural networks for SAR and optical image pairs. *2017 Conference on Big Data From Space*, 289–292.