Deep Relearning in the Geospatial Domain for Semantic Remote Sensing Image Segmentation

Christian Geiß, *Member, IEEE*, Yue Zhu, Chunping Qiu, Lichao Mou, *Student Member, IEEE*, Xiao Xiang Zhu, *Senior Member, IEEE*, and Hannes Taubenböck

***Abstract*— We present a classification post-processing (CPP) technique based on Fully Convolutional Neural Networks (CNN) for semantic remote sensing image segmentation. Conventional CPP techniques aim to enhance the classification accuracy by imposing smoothness priors in the image domain. In contrast to that, here, a relearning strategy is proposed where the initial classification outcome of a CNN model is provided to a subsequent CNN model via an extended input space to guide the learning of discriminative feature representations in an end-to-end fashion. This deep relearning convolutional neural network (DRCNN) accounts explicitly for the geospatial domain by taking the spatial alignment of preliminary class labels into account. Hereby, we evaluate to learn the DRCNN in a cumulative and non-cumulative way, i.e., extending the input space based on all previous or solely preceding model outputs, respectively, during an iterative procedure. Besides, the DRCNN can also be conveniently coupled with alternative CPP techniques such as object-based voting (OBV). Experimental results obtained from two test sites of WorldView-II imagery underline the beneficial performance properties of the DRCNN models. They can increase the accuracies of initial CNN models on average from 72.64% to 76.01% and from 92.43% to 94.52% in terms of κ statistic. An additional increase of 1.65 and 2.84 percentage points can be achieved when combining the DRCNN models with an OBV strategy. From an epistemological point of view, our results underline that CNNs can benefit from the consideration of preliminary model outcomes, and that conventional CPP techniques can profit from an upstream relearning strategy.**

***Index Terms*—deep learning, relearning, classification postprocessing, convolutional neural networks**

I. Introduction

The accurate extraction of thematic information from remote sensing data is a prerequisite for numerous applications. Consequently, it has become a major research subject for the remote sensing community [1]. Thereby, supervised classification approaches are very popular due to their accuracy and flexibility [2]. The governing principle of such methods is to infer a rule from limited but properly encoded prior knowledge, i.e., training data, to assign a class label to unseen instances of the domain under analysis. However, methods which solely exploit spectral signatures of individual pixels show frequently limited accuracy properties due to the well-known salt-and-pepper effect [3]. This is particularly relevant when analyzing remote sensing data with a ground sampling distance considerably higher than the objects of interest. The high spatial resolution can induce high intraclass and low interclass variabilities.

To address this issue, distinguishable strategies were established in the past which consider not solely the spectral properties of individual pixels but also spatial relations. In this manner, features were designed which encode the neighborhood characteristics of individual pixels such as morphological operators [4] or texture measures [5]. Alternatively, the image is partitioned with a segmentation algorithm into segments/objects and e.g., the spectral means and spatial-hierarchical context characteristics are deployed (referred to as object-based image analysis (OBIA)) [3]. Hybrid approaches were followed also, which aim to combine the aforementioned processing principles (e.g., so-called Object-based Morphological Profiles [1]).

As an alternative or additional processing step, the refinement of the classification outcome by classification postprocessing (CPP) can be followed [6]. With respect to CPP techniques, the majority of approaches aim to refine the initial classification outcome by building upon spatial occurrence and alignment of class labels and eventually *relabel them in the image domain*, based on e.g., majority filtering [7]. However, recently it was shown that the concept of *relearning* can be a more accurate CPP strategy [6], [8]-[10]. Thereby, a supervised classification model is learned for a second time with additional features derived from the initial classification outcome to enhance the discriminative properties of relearned decision functions. Here, also a relation to methods such as stacked generalization [11] can be drawn, which include information from a prior model outcome for a new prediction. However, relearning methods in the context of remote sensing account explicitly for the geospatial domain by taking the spatial alignment of class labels into account: Huang *et al.* [6] calculate a primitive co-occurrence matrix and local class histograms for quantification of the spatial alignment of class labels in the feature space. Experimental results showed better accuracies compared to a per-pixel

The work of Christian Geiß was supported by the Helmholtz Association under the grant "pre_DICT" (PD-305). Additionally, this research was funded in part by the German Federal Ministry of Education and Research (BMBF) under grant no. 03G0876 (project RIESGOS) and also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [714087]- So2Sat)

C. Geiß, and H. Taubenböck are with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Weßling-Oberpfaffenhofen, Germany (e-mail: christian.geiss@dlr.de; hannes.taubenboeck@dlr.de)

Y. Zhu is with the Department of Architecture, University of Cambridge, Cambridge, UK (e-mail: yz591@cam.ac.uk)

C. Qiu, is with the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: chunping.qiu@tum.de).

L. Mou and Xiao Xiang Zhu are with SiPEO-TUM and also with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

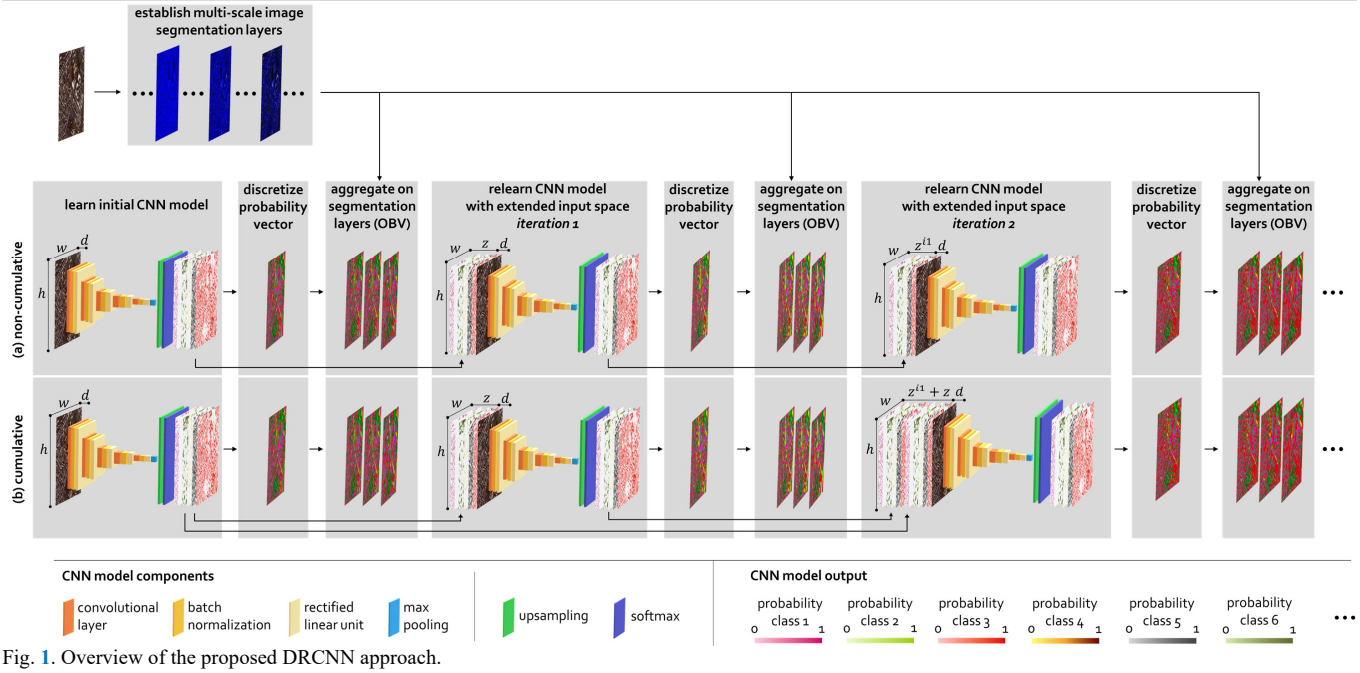


Fig. 1. Overview of the proposed DRCNN approach.

approach and traditional CPP methods. Geiß and Taubenböck [8] model spatial-hierarchical context relations with the preliminary classification outcome by computing class-related features using a triplet of hierarchical segmentation levels. Following the general principle of relearning strategies, those features are used to enlarge the initial feature space and impose spatial regularization in the relearned model which allowed for enhanced accuracy properties of the resulting model. An overarching relearning strategy was recently also employed in the context of an ensemble classifier method [9] and active learning framework [10].

In parallel, deep learning methods became increasingly popular for classifying remote sensing data [12] [13]. Models such as Fully Convolutional Neural Networks (CNN) [14] allow for learning a hierarchy of discriminative feature representations which frequently enable beneficial accuracy properties of pixel-level predictions given a sufficient amount of training data. Analogous to the aforementioned features which internalize the neighborhood characteristics of individual pixels, learned feature representations accumulate contextual information over large receptive fields [15].

In this letter, we built upon the idea of relearning and uniquely extend it in the context of a CNN model, referred to as deep relearning convolutional neural network (DRCNN). To this purpose, the initial classification outcome of a CNN model is provided to a subsequent CNN model via an extended input space to guide the learning of discriminative feature representations in an end-to-end fashion. Hereby, we evaluate the effects of learning the DRCNN in a cumulative and non-cumulative way, i.e., extending the input space based on all previous or solely preceding model outputs, respectively, during an iterative procedure. Finally, we also combine the outcome of the DRCNN with an alternative CPP strategy, i.e., object-based voting (OBV). Thereby, the imagery is partitioned with a segmentation algorithm on multiple segmentation layers. The DRCNN model outputs are aggregated on segment levels via a majority voting strategy. This strategy aims for preserving boundary information of objects which are frequently washed out by CNN models [15].

The remainder of the letter is organized as follows: section II details the proposed DRCNN relearning method. We describe the experimental setup in section III and report results of actual experiments in section IV. Concluding remarks are given in section V.

II. Proposed Methodology

The proposed DRCNN builds upon a CNN but deploys the outcome, i.e., class-conditional probabilities, for relearning the model (sec. IIA). In addition, it can be coupled with OBV to ultimately preserve boundary information of objects (sec. IIB).

A. Deep Relearning Convolutional Neural Network

Let $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$ be the input space, e.g., a multichannel multispectral image, where h, w, and d, correspond to height, width, and dimensionality, i.e., number of channels, respectively. A CNN imposes a set of learnable parameters $\mathbf{w}$ on $\mathbf{x}$ and establishes a corresponding output $\mathbf{z} = f(\mathbf{x}, \mathbf{w})$, i.e., $\mathbf{z} \in \mathbb{R}^{h' \times w' \times d'}$. The building blocks of a CNN include frequently several model components (Fig. 1). A convolutional block calculates the convolution of $\mathbf{x}$ with a set of $\bar{K}$ filters $\mathbf{w} \in \mathbb{R}^{\tilde{h} \times \tilde{w} \times d \times d'}$ given by

$$z_{i'j'k'} = f\left(b_{k'} + \sum_{i=1}^{\tilde{h}} \sum_{j=1}^{\tilde{w}} \sum_{u=1}^{d} w_{iju k'} \times x_{i'+i,j'+j,u}\right) \quad (1)$$

where b is a learned bias term and $f(\cdot)$ denotes a nonlinear activation function [16]. Regarding the latter, the rectified linear unit is deployed

$$z_{ijk} = \max\{0, x_{ijk}\}. \quad (2)$$

The convolution operation is directly followed by a batch normalization [17] to enhance stability properties of gradient

descent optimization and speed up convergence. Further, a max pooling operator can be employed to establish the maximum response of each feature channel in a $\tilde{h} \times \tilde{w}$ patch

$$z_{i'j'k} = \max_{1<i<\tilde{h},1<j<\tilde{w}} x_{i'+i,j'+j,k}. \quad (3)$$

After an upsampling module, multinomial logistic regression is deployed as classifier, whose scores represent class-conditional probabilities given by the softmax function

$$z_{ijk} = \frac{\exp(x_{ijk})}{\sum_{c=1}^{C} \exp(x_{ijc})} \quad (4)$$

for C classes. Based on the class label c of input x, the corresponding classification loss is computed via

$$z = -\sum_{i=1}^{h}\sum_{j=1}^{w} log x_{ijc} \quad (5)$$

to force the network to put all the mass on the correct labeling. Given our geospatial classification task, the loss is computed uniformly over the grid of spatial predictions. During inference, the probabilistic output can be discretized to obtain an unambiguous classification map

$$l(z) = \underset{c\in\{1,\dots,C\}}{\operatorname{argmax}} z_{ijk} \quad (6)$$

where $l(z)$ is the categorical label.

However, for the relearning procedure the class-conditional probabilities as obtained with eq. 4 are provided as separate input layers for each class via an extended input space, i.e., $\hat{\mathbf{x}} \in \mathbb{R}^{h\times w\times d\times z}$, to the DRCNN (Fig. 1). Hereby, we evaluate to learn the DRCNN in a cumulative and non-cumulative way, i.e., extending the input space based on all previous or solely preceding model outputs, respectively, during an iterative procedure. As such, the dimensionality of $\hat{\mathbf{x}}$ is enlarged by a value in proportion to the number of thematic classes to be estimated for both the non-cumulative and cumulative case. After the first iteration, the dimensionality of $\hat{\mathbf{x}}$ remains constant for the non-cumulative relearning procedure (Fig. 1a). In contrast, the dimensionality of $\hat{\mathbf{x}}$ increases linearly in the cumulative case as a function of the number of thematic classes and iterations (Fig. 1b). It can be noted that the training set is newly drawn after each iteration where labels remain the same while the input is altered or extended for the non-cumulative and cumulative case, respectively. Thereby, the number of iterations can be set according to a stopping criterion. Finally, the model which maximizes a defined accuracy measure is selected. As such, we treat the number of iterations as a hyperparameter which needs to be optimized in a data-driven way without prior constraints.

B. Object-based voting

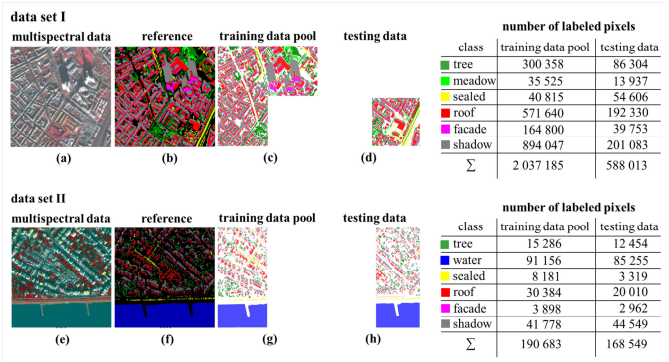
The DRCNN can be coupled with a CPP strategy such as OBV [1], [8]. $\mathbf{x}$ is partitioned with a segmentation algorithm at a certain segmentation level s in N^s objects O_i^s $(i = 1, 2, \dots, N^s)$. Thereby, the following constraint must be fulfilled to establish an unambiguous hierarchy of segmentation levels:

$$\bigcup_{O_i^{s-1}\subseteq O_j^s} O_i^{s-1} = O_j^s \quad (7)$$

This way it is ensured that an object at segmentation level $s-1$ must be included in only one object at level s. For an exhaustive description of the objects of an image, multiple hierarchical segmentation levels can be deployed, i.e., $M \in \{\dots, s-1, s, s+1, \dots\}$. For aggregation of the class labels obtained with eq. 6 to M, we deploy a crisp OBV scheme [6], what corresponds to a majority vote per object:

$$C(O) = \underset{c\in\{1,\dots,C\}}{\operatorname{argmax}} \left(\frac{1}{N^O} \sum_{v\in O} \tau(C(v) = c) \right). \quad (8)$$

Thereby, $C(O)$ refers to the final label of object O. τ is an indicator function which captures the number of times that the labeled pixels v within an object feature class label c. N^o corresponds to the number of pixels of an object.



number of labeled pixels

class	training data pool	testing data
tree	300 358	86 304
meadow	35 525	13 937
sealed	40 815	54 606
roof	571 640	192 330
facade	164 800	39 753
shadow	894 047	201 083
Σ	2 037 185	588 013

number of labeled pixels

class	training data pool	testing data
tree	15 286	12 454
water	91 156	85 255
sealed	8 181	3 319
roof	30 384	20 010
facade	3 898	2 962
shadow	41 778	44 549
Σ	190 683	168 549

Fig. 2. Experimental data for two test areas; (a/e) multispectral image from WorldView-II acquired over the city of Cologne (Germany); (b/f) reference pixels; (c/g) pool of labeled pixels used for stratified random selection and model learning; (d/h) spatially disjoint labeled pixels used for validation. The number of pixels assigned to the training data pool and testing data is also provided.

III. Experimental Setup

The experimental analysis was carried out by classifying two test areas taken from a multispectral image. This image was acquired over the city of Cologne, Germany, by the WorldView-II satellite sensor on January 31, 2014, with a geometric resolution of 0.5 m. The first data set consists of 2000 × 2000 pixels and shows an urban area of Cologne, which is dominated by buildings of commercial use (Fig. 2a; data set I). The second data set comprises 900 × 900 pixels and represents an area of residential buildings next to the river Rhine (Fig. 2e; data set II). Both image subsets feature a complex and small-scale composition of urban land cover captured by an off-nadir acquisition. The pixels of the image were grouped into six relevant thematic classes (Fig. 2b/f). The thematic classes of the individual pixels were determined based on photointerpretation analysis under consideration of additional aerial imagery and cadastral maps. In the subsequent experiments, the training patches were sampled from the training data pool uniformly with respect to class frequencies to establish a balanced training set. Thereby, it was made sure that training and testing data were compiled in a strict spatially disjoint way, while also taking the spatial extents of the receptive fields into account, to allow for unbiased estimates of model generalization capabilities [18] (Fig. 2c-d/g-h). Additionally, it was made sure that both the training data pool and testing data contains a sufficient number of samples for each class. For the experiments a subset of the training data is deployed for model learning, however, the complete set of testing samples is used for validation.

Regarding the OBV strategy, we deploy a bottom-up region-growing segmentation algorithm for partitioning of **x** (i.e., fractal net evolution approach [19]). In the experiments, we put more emphasis on shape heterogeneity rather than on gray-value heterogeneity with respect to the segmentation algorithm. This is due to the fact that man-made structures such as buildings and other elements of urban environments have distinct shape and size properties, unlike, for example, natural features. Analogously, the weights for heterogeneity of smoothness and compactness were maintained equal (i.e., shape: 0.7 and color: 0.5) and kept constant. Three hierarchical segmentation levels were created for the experiments to establish a range of potentially useful segment-based representations of the image content.

Regarding the CNN model architecture, the spatial extent of the input training samples, i.e., receptive field, should internalize the size of the objects of interest. As such, the majority of urban land cover objects feature dominant scales between three and 24 meters [20]. Consequently, the window size for training sample extraction was set to 60 × 60 pixels, i.e., 30 meters. Overall, we deploy five hidden layers with 3 × 3 kernels, and implement max pooling after the last block. The number of features was set to 10 for each convolutional layer. The learning rate was fixed to 0.0006 and deployed over 30 epochs per iteration. We want to stress that our goal here is not to set up the best possible model architecture for the data sets, but to enable *ceteris paribus*-near conditions for the comparative evaluation of the different processing principles.

The relearning procedure was implemented with five iterations. Final model selection was carried out by evaluating corresponding κ statistics as global measures of accuracy, whereas the thematic accuracies of the obtained classification maps were additionally assessed by computing the overall accuracy (OA) and producer's (PA) and user's accuracy (UA).

IV. Experimental Results

Averaged estimated κ statistics and OA with corresponding standard deviation (SD) from five model runs with independently drawn 5000 labeled training samples per class are revealed for both data sets in Fig. 3. It can be seen that the proposed DRCNN approach can on average achieve consistently higher accuracies than the CNN models while simultaneously reducing SD. An increase of 3.09 and 2.52 percentage points (p.p.) in terms of κ and OA, respectively, can be observed for data set I. A similar trend is revealed for data set II, where κ and OA can be further increased from a fairly high accuracy level by 1.95 and 1.28 p.p., respectively. Thereby, the non-cumulative relearning strategy provides on average slightly higher accuracies compared to the cumulative scheme with an additional increase of 0.56 and 0.29 p.p. in terms of κ for data set I and II, respectively. Thereby, the non-cumulative strategy is also beneficial regarding the time for learning the models. During the experiments, we observed that the runtime remained approximately constant while it increased in a linear fashion for the cumulative relearning strategy. Generally, the experimental analyses underline that CNNs can benefit from the consideration of preliminary model outcomes and that the proposed DRCNN can

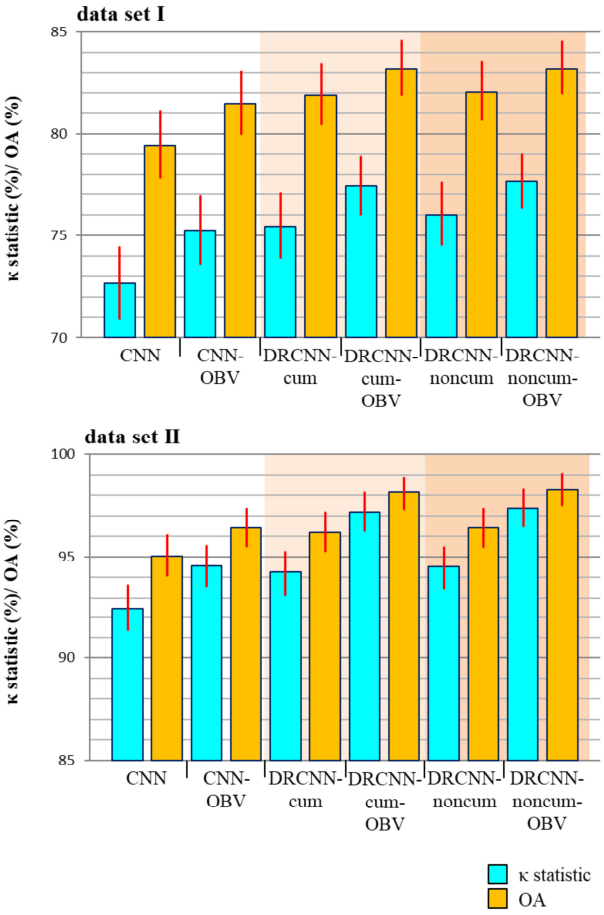


Fig. 3. Averaged κ statistic and OA with corresponding error bars representing one SD as obtained from five independent trials for the two test data sets given 5000 labeled samples per class.

unambiguously provide superior accuracies compared to the initial CNN models.

When coupling the models with the OBV scheme, additional improvements can be observed. The accuracies of the initial CNN models increase from 72.64% to 75.24% and from 92.43% to 94.53% with respect to κ for data set I and II, respectively. The DRCNN-based results can provide an additional increase to 77.45% and 77.66% as well as 97.19% and 97.36%, regarding the cumulative and non-cumulative strategy for data set I and II, respectively. Generally, these findings are in line with recent studies which couple CNN models with an OBV strategy and also obtain enhanced accuracy properties [21]. However, here we additionally show that conventional CPP techniques such as OBV can further profit from an upstream relearning strategy.

We can therefore state that, overall, the best results could be achieved with the DRCNN models which deploy also an OBV scheme over a non-cumulative relearning strategy. They allowed increasing the initial CNN model accuracies by 5.02 and 3.75 p.p., as well as by 4.93 and 3.25 p.p. regarding κ and OA for data set I and II, respectively. This underlines the beneficial performance of the proposed methods.

When inspecting the classification maps of a single model run and corresponding PA and UA values (Fig. 4), it becomes comprehensible that the DRCNN models with OBV strategy

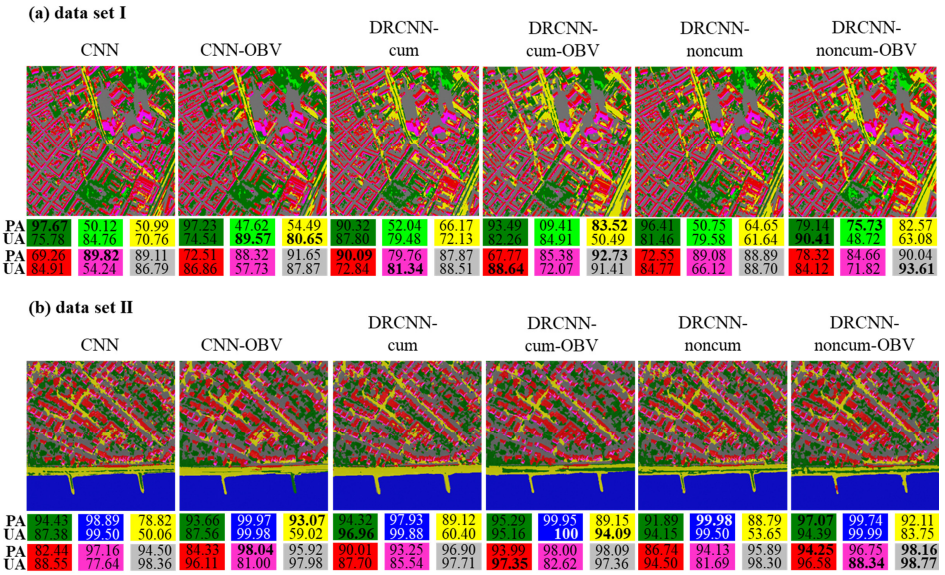


Fig. 4. Classification maps obtained from a run with 5000 samples per class for the two data sets. Class-specific values for PA and UA are also provided. Highest PA and UA values for each land cover class are marked in **bold**.

can provide spatially homogenized solutions and synergize most class-specific top scores. This confirms the capability of the proposed methodology to deploy the intrinsic patterns in initial classification outcomes for learning models with enhanced discriminative properties while simultaneously encoding adequate smoothness priors.

V. Conclusions

In this letter, we have proposed a novel post-classification relearning strategy for CNNs to enhance classification accuracy. It was inspired by the observation that relearning strategies can be superior compared to CPP methods which relabel model outcomes in the image domain. Thus, the designed DRCNN model deploys initial classification outcomes via an extended input space to guide the learning of discriminative feature representations. The experimental results unambiguously stress that CNNs can benefit from the consideration of preliminary model outcomes and also that conventional CPP techniques can further profit from an upstream relearning strategy. In the future, we aim to implement diagnostic schemes which allow to further track the improved feature representations of CNN-based relearning strategies.

References

[1] C. Geiß *et al.*, "Object-based Morphological Profiles for Classification of Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5952-5963, Oct. 2016.

[2] C. Geiß *et al.*, "Virtual Support Vector Machines with self-learning strategy for classification of multispectral remote sensing imagery," *ISPRS Int. J. Photogramm. Remote Sens*., vol. 151, pp. 42–58, 2019.

[3] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Int. J. Photogramm. Remote Sens*., vol. 65, no. 1, pp. 2-16, Jan. 2010.

[4] R. Haralick, S. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 4, pp. 532–550, Jul. 1987.

[5] R. Haralick, "Statistical and Structural Approaches to Texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786-804, May 1979.

[6] X. Huang *et al.*, "New postprocessing methods for remote sensing image classification: A systematic study," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7140–7159, Nov. 2014.

[7] J. Stuckens, P. R. Coppin, and M. E. Bauer, "Integrating contextual information with per-pixel classification for improved land cover classification," *Remote Sens. Environ.*, vol. 71, no. 3, pp. 282–296, Mar. 2000.

[8] C. Geiß and H. Taubenböck, "Object-based Postclassification Relearning," *IEEE Geosci. Remote Sens. Letters*, vol. 12, no. 11, pp. 2336-2340, Nov. 2015.

[9] X. Han *et al.*, "The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery," *ISPRS Int. J. Photogramm. Remote Sens*., vol. 138, pp. 57-73, 2018.

[10] Q. Shi, X. Liu, and X. Huang, "Active Relearning Framework for Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3468–3485, Jun. 2018.

[11] D. H. Wolpert, "Stacked Generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241-259, Jan. 1992.

[12] X. X. Zhu *et al.*, "Deep Learning in Remote Sensing: A Review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8-36, Dec. 2017.

[13] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: status and perspectives," *Natl. Sci. Rev.*, vol. 6, no. 6, pp. 1082-1086, 2019.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in Proc. IEEE Conf. Comput. Vision Pattern Recog., 2015.

[15] D. Marmanis *et al.*, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Int. J. Photogramm. Remote Sens*., vol. 135, pp. 158-172, 2018.

[16] A. Vedaldi and K. Lenc, "Matconvnet-convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015.

[18] C. Geiß *et al.* "On the Effect of Spatially Non-disjoint Training and Test Samples on Estimated Model Generalization Capabilities in Supervised Classification with Spatial Features," *IEEE Geosci. Remote Sens. Letters*, vol. 14, no. 11, pp. 2008-2012, 2017.

[19] M. Baatz, A. Schäpe (2000). Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. In J. Strobl, T. Blaschke & G. Griesebner (Eds.), Angewandte Geographische Informations-Verarbeitung XII (pp. 12-23). Wichmann Verlag, Karlsruhe, Germany.

[20] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[21] S. Liu *et al.* "Integration of Convolutional Neural Networks and Object-Based Post-Classification Refinement for Land Use and Land Cover Mapping with Optical and SAR Data," *Remote Sens.*, 11, 690, 2019.