CNN-BASED MULTI-SCALE HIERARCHICAL LAND USE CLASSIFICATION FOR THE VERIFICATION OF GEOSPATIAL DATABASES

C. Yang *, F. Rottensteiner, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
(yang, rottensteiner, heipke)@ipi.uni-hannover.de

Commission II, WG II/6

KEY WORDS: classification, CNN, land use database, hierarchy, multi-scale

ABSTRACT:

Land use is an important piece of information with many applications. Commonly, land use is stored in geospatial databases in the form of polygons with corresponding land use labels and attributes according to an object catalogue. The object catalogues often have a hierarchical structure, with the level of detail of the semantic information depending on the hierarchy level. In this paper, we extend our prior work for the CNN (Convolutional Neural Network)-based prediction of land use for database objects at multiple semantic levels corresponding to different levels of a hierarchical class catalogue. The main goal is the improvement of the classification accuracy for small database objects, which we observed to be one of the largest problems of the existing method. In order to classify large objects using a CNN of a fixed input size, they are split into tiles that are classified independently before fusing the results to a joint prediction for the object. In this procedure, small objects will only be represented by a single patch, which might even be dominated by the background. To overcome this problem, a multi-scale approach for the classification of small objects is proposed in this paper. Using this approach, such objects are represented by multiple patches at different scales that are presented to the CNN for classification, and the classification results are combined. The new strategy is applied in combination with the earlier tiling-based approach. This method based on an ensemble of the two approaches is tested in two sites located in Germany and improves the classification performance up to +1.8% in overall accuracy and +3.2% in terms of mean F1 score.

1. INTRODUCTION

Land use describes the socio-economic function of a piece of land. This information is frequently maintained by governmental mapping agencies. Commonly, land use data is stored in the form of polygon objects in geospatial databases, the labels of which indicate the corresponding land use. In order to verify this information automatically as a first step of a database update, current remote sensing data can be employed to predict a land use label. The predicted label can then be compared to the one contained in the database, and inconsistent predictions can be interpreted as cues for land use change.

Today, work on image-based classification is dominated by convolutional neural networks (CNN) (Krizhevsky et al., 2012). CNN require images of a fixed size as input. If the goal is to predict the current land use for every polygon in the database, a big challenge relates to the large variation of polygons in terms of their geometrical extent. In addition, object catalogues of geospatial databases typically contain a very large number of land use classes (also called categories, these terms are used interchangeably in this paper), many of which cannot be expected to be distinguishable in remote sensing imagery. On the other hand, many object catalogues, e.g. the catalogue used in the German Authoritative Real Estate Cadastre Information System (ALKIS; AdV, 2008), provide land use information in multiple semantic levels with a hierarchical structure. From the point of view of the application, it is therefore useful to obtain predictions at multiple semantic levels simultaneously.

Consequently, in (Yang et al., 2020a) we proposed a method for the hierarchical classification of land use polygons based on CNN, in which land use labels consistent with the pre-defined object-class hierarchy were predicted at multiple semantic levels simultaneously. The input consists of multispectral aerial imagery and derived height data at a resolution in the order of 0.1 to 0.2 metres. The classification is based on a two-stage process: first, a fully convolutional network (FCN) (Long et al., 2015) is applied to predict the current land cover at pixel level; the resultant land cover posteriors, the original data and a binary mask encoding the polygon shape provide the input to the second step, the CNN-based prediction of land use at multiple hierarchical levels. The evaluation has shown that the classification quality clearly depends on the size of the polygon: small polygons, i.e. polygons which fit into a window of 256 x 256 pixels (which is the input size of the CNN) are classified with considerably lower accuracy than the large ones. To a certain degree it is not a surprise for some of them to be classified incorrectly: some polygons in the database cover only about 10% of the area of the image patch, so that the image content will be dominated by the surroundings.

In this paper, we address the problem of classifying small land use objects in the context of the hierarchical classification technique presented in (Yang et al., 2020a). A simple scaling approach (Yang et al., 2019) was found not to be sufficient to solve the problem. Thus, in this paper we present a *multi-scale approach*: each small object (according to the above definition, this is an object fitting into a window of 256 x 256 pixels in the resolution of the sensor data) is presented to the CNN multiple times, each time in a different scale, and the predictions are combined afterwards. Apart from capturing context regions of multiple size and processing images that are dominated by the interior of the object, also our experience with large objects, which are split into tiles that are classified independently before determining a joint classification result, gives rise to the

* Corresponding author.

expectation that the combination of multiple predictions may act as a kind of ensemble method and improve the quality of the predictions accordingly.
The scientific contribution of this paper can be summarized as follows:

- Based on our previous work for hierarchical land use classification (Yang et al., 2020a), we propose a *multi-scale* approach for classifying small land use objects to improve the classification accuracy for these objects;
- We validate that approach by conducting a series of experiments in two test sites located in Germany. At the same time, we highlight the benefits and investigate the limits of the proposed approach in differentiating fine-grained class structures corresponding to the finest semantic level of a hierarchical object catalogue.

In section 2, we give a brief review of related work. Our new multi-scale approach is presented in section 3. Section 4 describes the experimental evaluation of our method. Conclusions and an outlook are given in section 5.

2. RELATED WORK

Since the success of AlexNet (Krizhevsky et al., 2012), CNN have been shown to outperform other classifiers by a large margin. They have also been widely adopted for classification in remote sensing applications; cf. (Zhu et al., 2017) for an overview.

Zhang et al. (2018) propose a segment-based approach to determine land use from remote sensing data. The authors start with an initial non-semantic image segmentation using mean-shift (Comaniciu and Meer, 2002), the resultant segments are then considered to correspond to objects for which land use is to be predicted. These segments are split into rectangular patches using the moment bounding box method of Zhang and Atkinson (2016). These patches, which consist of either 48 x 48 or 128 x 128 pixels, are classified independently from each other using a CNN. The final class label of each segment is determined by combining the predictions for all patches by simple majority vote. Zhang et al. (2019) propose a joint deep learning framework for classifying land cover and land use simultaneously in an iterative approach. Both (Zhang et al., 2018) and (Zhang et al., 2019) focus on 10 urban land use classes only.

In contrast to the approaches cited so far, Huang et al. (2018) rely on the availability of polygons representing urban blocks for which land use is predicted on the basis of multispectral images. Each polygon is represented by a series of rectangular processing units of 227 x 227 pixels which are positioned inside the polygon on the basis of a skeleton. These processing units are classified independently from each other using a CNN-based approach. The final prediction for a polygon is obtained by computing the arithmetic mean of the class scores of all corresponding processing units. Their work focuses solely on urban land use and differentiates 13 classes. All methods cited so far differentiate land use classes only in one semantic level.

A problem that occurs when predicting class labels for database objects is the large variability of such objects in size. One strategy to cope with this problem is an analysis of the input data at multiple scales. In the context of multi-scale analysis, many researches use a pyramidal approach to capture context areas of different size, using the image at different resolutions as input. Marmanis et al. (2018) adopted the multi-scale approach originally described in (Kokkinos, 2016) for land cover classification and gained a slight improvement (0.2%) in terms of overall accuracy. Auderbert et al. (2018) proposed an alternative way of multi-scale analysis by combining predictions of land cover at different resolutions (corresponding to different layers of the network decoder) to achieve final prediction, which also leads to a slight improvement (0.3%) in terms of overall accuracy. However, these methods apply a pixel-wise prediction of land cover, not a prediction of land use for objects of a geospatial database.

Considering classifying land use objects in multiple semantic levels while guaranteeing consistent hierarchical predictions, Yang et al. (2020a) proposed two approaches to classify land use objects in three semantic levels according to the ALKIS object catalogue. In the evaluation we found again a considerable accuracy discrepancy between large and small polygons. In this paper, we propose an additional multi-scale approach to address this problem. The small polygons are represented by multiple patches at different scales that are presented to the CNN for classification, and the classification results are combined. In this context, the size of the polygons is enlarged to mitigate the influence of the area outside the object boundaries on the classification result.

3. HIERARCHICAL CLASSIFICATION OF LAND USE

For our method, the first input required for the CNN-based hierarchical land use classification is a land use database in which all objects are represented by polygons with land use categories at multiple semantic levels according to a hierarchical object catalogue. Multispectral aerial image (RGB-IR), a normalized digital surface model (nDSM), i.e. a model of heights above the terrain, and pixel-wise class scores for land cover obtained from a first pixel-wise land cover classification step serve as additional input. To obtain land cover class scores, the CNN-based land cover classification method of Yang et al. (2021) is used. The goal of CNN-based land use classification is the prediction of one class label for every polygon at three semantic levels in a way that is consistent with the hierarchic object catalogue.

As mentioned earlier, in CNN-based land use classification a big challenge is the large variation of polygons in terms of their geometrical extent. For instance, *road* objects are commonly long and thin, and *residential* objects cover both, quite large and quite small areas. To overcome this problem, large objects have to be split into several patches first. These patches are classified by the CNN independently from each other, and finally, the individual predictions are combined to determine the class label of the compound object. In the following, we adapt the method presented in (Yang et al., 2021) for that purpose. In that paper, large polygons were split into patches by a tiling approach, whereas small polygons (i.e., polygons that fit into a window of the input size of the CNN) were only represented by a single patch at the original scale of the remote sensing data. In section 3.1, we propose an alternative patch preparation strategy for small objects, which is the main methodological contribution of this paper. In section 3.2 we briefly outline the CNN architecture for land use classification of (Yang et al., 2021) to make this paper self-contained. Section 3.3. presents several network variants. It also describes the method for combining the predictions for individual patches and gives implementation details.

3.1 Patch preparation

In (Yang et al, 2021), a window of 256 x 256 pixels centred at the centre of gravity of the object from all data (image and nDSM, binary object mask, land cover scores) is extracted and then

presented to the CNN. This procedure is unproblematic if the polygon size corresponds well to the window size at the ground sampling distance (GSD); otherwise the window is either dominated by information outside the object (for very small objects) or the object does not fit into the window. In (Yang et al., 2021), large objects not fitting into the input window of the CNN are split into tiles; this method is outlined in section 3.1.1. In section 3.1.2, we present an alternative approach based on scaling, in which for small polygons patches are generated at different scales. These strategies can be combined to achieve a classification methodology in which both small and large objects are represented by multiple patches (cf. section 3.3).

3.1.1 Tiling approach: For large objects, the window enclosing the object is split into tiles (patches) of the desired size. This might cause the number of patches to be very large. As a consequence, the training procedure is expected to lead to overfitting to patches corresponding to large objects. To avoid this, for objects with more than 3 patches, we randomly select 40% of these patches for further processing only, whereas the other patches are discarded. For all other objects, all patches are preserved. Fig. 1 illustrates this process for a large road object.

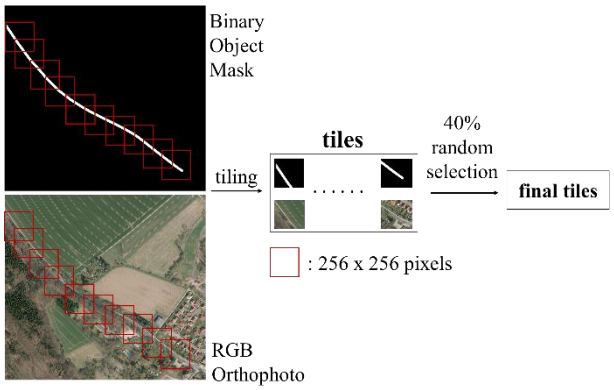

Figure 1: Illustration of the tiling process for a large road object.

3.1.2 Multi-scale approach: Using the method described in section 3.1.1 for patch generation, small objects will correspond to exactly one patch, which additionally might be more representative for the background than for the object. To improve the classification of small polygons, we propose a new method in which they are represented in different scales both in training and classification. First, the input data are scaled such that the object fits exactly into a window of the input size of the CNN using the scale

$$s_1 = \frac{256}{\max(w,h)}, \qquad (1)$$

where *w* and *h* are the width and height of a rectangle enclosing the object at the GSD of the images in [pixels], respectively. The other scales s_k are based on s_1:

$$s_k = \frac{1}{2^{k-1}} \cdot s_1, \qquad (2)$$

These scales are computed for $k \in \{2, 3, 4, 5\}$, i.e. the minimum scale that can be considered is $1/16 \cdot s_1$; however, we do not use scales $s_k < 1$, except for s_1. For objects that fit into a window of 256 x 256 pixels at the GSD of the input image, we additionally define a scale $s_0 = 1$, so that for these data, the original input is used for patch generation, too. Thus, the number of scales applied to an object depends on the object size. For each selected scale s_k, a corresponding window centred at the object centre is extracted from all the input data and up-scaled to 256 x 256 pixels. In the scaling process, the binary mask images are interpolated via nearest neighbour interpolation; for all other data, bilinear interpolation is applied. Fig. 2 presents an example for a multi-scale representation of a small polygon. In this case, three scale factors are applied.

This patch generation procedure implies that large polygons are represented by a single patch extracted using the scale s_1. Thus, the window enclosing the entire object is downscaled to the desired input size of 256 x 256 pixels. Thus, for large objects, this procedure is the identical to the one described in our previous work (Yang et al., 2019). In that paper, this scaling approach did not work very well when applied as a stand-alone procedure, though it could improve the results slightly when used in an ensemble with the tiling method.

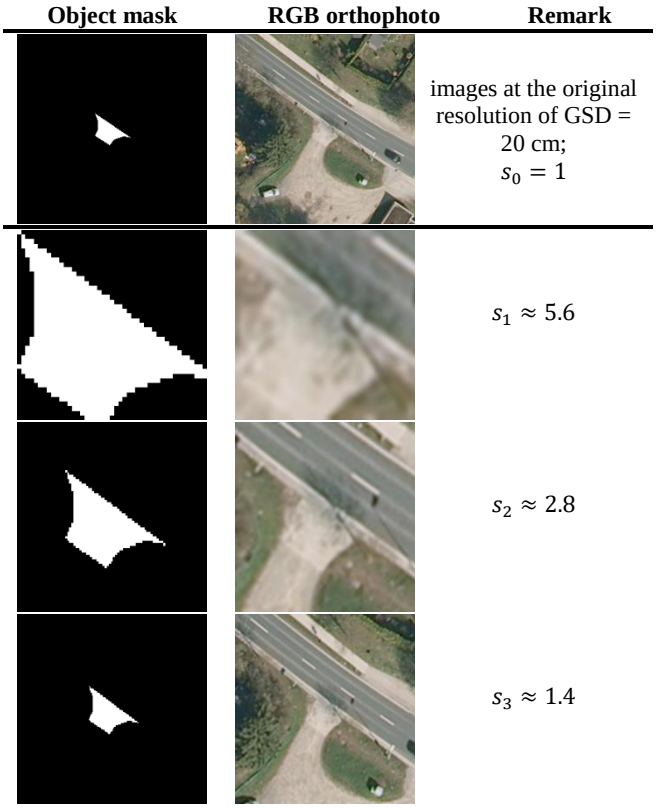

Figure 2: An example showing the contents of a small polygon at the original scale and up-scaled versions at three different scales. All images have a size of 256 x 256 pixels. The size of the polygon is about 45 x 45 pixels in the resolution of the images (about 81 m^2 in area).

3.2 Network architecture

The classification of the patches generated in one of the ways described in section 3.1 is based on the *LuNet-lite-JO* network described in (Yang et al., 2021) and presented in Fig. 3. The input image patches are processed by a series of blocks of convolution and pooling layers. Afterwards, the network is split into two branches. The first branch consists of standard convolution and pooling layers, whereas the second one extracts a ROI from the feature map of the previous joint layer that tightly encloses the object. Subsequently, rescaling of that ROI to 16 x 16 pixels is performed and a set of convolutions and poolings is applied. Finally, the feature vectors of the two branches are concatenated to form a combined 128-dimensional vector. The combined

vector is processed by three fully-connected layers to obtain raw unnormalized class scores z_{LU}^{l} for each of the three semantic levels l. These class scores are the input to a network block consisting of two layers with a specific connectivity structure designed for learning semantic dependencies between the different layers; the reader is referred to (Yang et al., 2021) for details about these layers. The output consists of raw class scores $z_{LU}^{out,l}$ per layer, which are passed to the final softmax layer to produce probabilistic class scores.

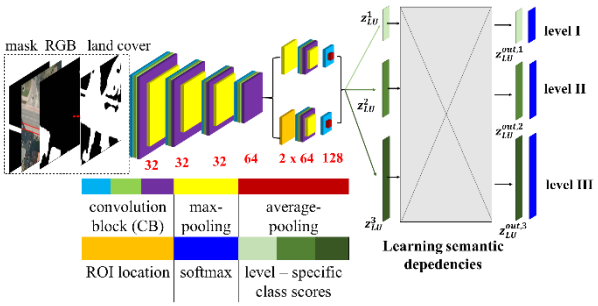

Figure 3: Architecture of *LuNet-lite-JO*. The red numbers indicate the number of filters per layer. More details are given in the main text.

Although the last layers of the network are designed to learn dependencies between classes at different semantic levels, there is no guarantee that the prediction results are consistent with the hierarchical object class catalogue. To achieve semantic consistency, the joint optimization (JO) strategy, also proposed in (Yang et al. 2021), is used. The basic idea is to maximize the joint class scores of the consistent triplets of class labels. As the class structure is hierarchical, each class at the finest semantic level corresponds to one such triplet; each triplet consists of the corresponding class in the finest level and its predecessors according to the class hierarchy. The joint class score of a triplet is the product of the scores of all labels in a triplet, and the triplet having maximum joint class score is selected as the prediction result for each patch.

3.3 Training, network variants and inference at object level

3.3.1 Training: Training of the network described in section 3.2 is based on stochastic mini-batch gradient descent. The input consists of patches with known triplets of class labels (one per semantic level). The loss function consists of two terms designed to maximise the joint class scores for triplets of predictions that match the reference and to minimize the class scores of triplets not corresponding to the reference, respectively. For more details, the reader is referred to (Yang et al., 2021).

3.3.2 Network variants: In section 3.1, two different methods for producing patches to be classified by the CNN have been described. Of course, the patches to be used for training (cf. section 3.3.1) and for testing have to be generated using the same approach. Thus, there are different network variants. The variant *LuNet-lite-JO-T* is based on patches generated by tiling (cf. section 3.1.1). It is identical to the strategy pursued in (Yang et al., 2021) and serves as a baseline in our experiments. The second variant, denoted by *LuNet-lite-JO-MS*, is based on using patches generated by the multi-scale approach (cf. section 3.1.2). As pointed out in section 3.1.2, for large polygons this variant is not expected to work too well. The third variant, referred to as *LuNet-lite-JO-ENS*, is an ensemble of the first two networks and is what we consider to be the main variant investigated in this paper. At test time, it takes the first two networks (trained independently from each other) and combines their outputs in a decision level fusion process described below. We expect this variant to combine the advantages of the two basic approaches and lead to an improved classification performance.

3.3.3 Inference at object level: Each network delivers class scores that are consistent with the class hierarchy of the object catalogue of the geospatial database for a single patch. The predictions of multiple patches have to be combined to obtain the final class scores for an object to be classified. In case of *LuNet-lite-JO-T* and *LuNet-lite-JO-MS*, these patches are generated by one of the two patch generation strategies described in section 3.1, respectively, and there may be objects corresponding to one patch only. In the variant *LuNet-lite-JO-ENS*, both patch generation strategies are applied. In this case, the set of patches generated by tiling are processed by *LuNet-lite-JO-T* and the patches generated by the multi-scale approach are processed by *LuNet-lite-JO-MS*. Consequently, all objects correspond to multiple patches in the case of *LuNet-lite-JO-ENS*.

The combination of the class scores of the individual patches is identical for all variants. For objects which are not split in the tiling process due to their size, the prediction of the related patches is directly used to define the result at object level. Of course, as pointed out earlier, this will only occur for variants *LuNet-lite-JO-T* and *LuNet-lite-JO-MS*. For objects which had to be split, we first compute combined class scores per semantic level by taking the product of the corresponding softmax outputs of all patches. These products form the basis for selecting the optimal triplet of class labels using the joint optimization procedure outlined in section 3.2. That is, the joint optimization procedure is not applied at patch level, but at object level.

3.3.4 Implementation: All networks are implemented based on the tensorflow framework (Abadi et al., 2015). We use a GPU (Nvidia TitanX, 12GB) to accelerate training and inference.

4. EXPERIMENTS

4.1 Test Data und test setup

4.1.1 Test data: Two German test sites are used for our experiments. The first one is located in Hameln, covering an area of 2 x 6 km^2 with various urban and rural characteristics. The second one is located in Schleswig. It covers an area of 6 x 6 km^2 and has similar characteristics as Hameln. For both test sites, digital orthophotos (DOP), a nDSM and land use objects from the German Authoritative Real Estate Cadastre Information System (ALKIS) are available. The DOP are multispectral images (RGB-IR) with a GSD of 20 cm. The nDSM was generated from a digital surface model generated by image matching and subtracting a given digital terrain model. The ALKIS object catalogue (AdV, 2008) is used to obtain the hierarchical class structure. There are three semantic levels with 4 classes at level I, 14 classes at level II and 21 classes at the finest level III; the class structure is presented in Tab. 1 along with the number of samples per class. The total number of land use objects is 2945 in Hameln and 4345 in Schleswig.

4.1.2 Test setup: Each test dataset is split into six blocks for cross validation. The block size is 10.000 x 5.000 pixels (2 km^2) and 30.000 x 5.000 pixels (6 km^2) for Hameln and Schleswig, respectively. In each test run one block is used for testing and the rest for training. In each run, about 15% of all training samples are used for validation and the rest is used for updating the network parameters. We report the average overall accuracy and

F1 scores over all test runs for evaluation, in both cases based on the number of correctly classified database objects.

We use the *FuseNet-lite* architecture of (Yang et al., 2021) for pixel-wise land cover classification based on the available input data. For both datasets we differentiated eight land cover classes (*building, sealed area, bare soil, grass, tree, water, car* and *others*), so that the input patches for the networks for predicting land use have 14 bands (4 DOP bands, nDSM, binary mask, 8 land cover inputs). The overall accuracy at pixel level was 88.8% in Hameln and 86.5% in Schleswig (Yang et al., 2021).

level I	level II	level III	#H	#S
settlement	*residential area (res.)*	*residential in use (res.use)*	528	803
		extended residential area (ext. res.)	34	61
	industry area (ind.)	*factory area (fact.)*	87	39
		business area (busi.)	193	158
		infrastructure (infra.)	54	62
	mixed-used area (mix)	*mixed-used area (mix)*	9	127
	special area (special)	*special area (special)*	135	207
	recreation area (recreation)	*sport & leisure area (leisure)*	27	64
		park	299	365
traffic	*road traffic (ro.traf)*	*motor-road*	530	732
		traffic-guided area (traf.area)	134	75
	path & way (path)	*path & way (path)*	477	287
	parking lot (park.lot)	*parking lot (park.lot)*	91	76
vegetation	*agriculture*	*farm land*	58	214
		garden/fallow land (garden)	100	440
	forest	*hardwood or softwood (h/s. wood)*	33	154
		hard and softwood (h&s. wood)	15	134
	grove	*grove*	51	88
	moor or swamp (moor)	*moor or swamp (moor)*	31	116
water bodies	*flowing water (flow.wat.)*	*flowing water bodies (flow.wat.bo.)*	54	41
	stagnant water (stag.wat.)	*stagnant water bodies (stag.wat.bo.)*	5	102
		Total number of objects	2945	4345

Table 1. Hierarchical class structure. Abbreviations are shown in brackets. *#H / #S*: number of samples in level III for Hameln and Schleswig, respectively.

In the training phase, the setting of the hyper-parameters is kept the same as in (Yang et al., 2021). Weight decay is 0.0005, the total number of training epochs is 8, and the minibatch size is 30. The base learning rate is 0.001 and the rate is reduced by a factor of 10 after four epochs. In addition, data augmentation (DA) is applied on both datasets. For patches generated by the tiling approach, DA is the same as described in (Yang et al., 2021). For patches generated by multi-scale approach, all patches are augmented by horizontal and vertical flipping and 36 random rotations, so that each original patch contributes 39 training patches.

4.2 Evaluation

In section 4.2.1, we firstly compare the results obtained by the three network variants described in section 3.3.2 and then take a closer look at the performance for individual classes. The results delivered by *LuNet-lite-JO-T*, corresponding to the method described in (Yang et al., 2021), serve as a baseline for comparison. In section 4.2.2 we analyse the achieved accuracies as a function of object size to assess the impact of the multi-scale patch generation approach on the results for small objects.

4.2.1 Comparison of network variants: Tab. 2 presents an overview or the results obtained by *LuNet-lite-JO-T, LuNet-lite-JO-MS* and *LuNet-lite-JO-ENS* in both test sites. Tabs. 3 and 4 give the detailed F1 scores of all categories over all levels in Hameln and Schleswig, respectively. The results in Tab. 2 show that there is a clear ranking of the methods according to the achieved quality metrics across all semantic levels. In all cases, the variant based on multi-scale patch generation (*LuNet-lite-JO-MS*) achieves the lowest quality numbers. The variant based on tiling (*LuNet-lite-JO-T*) achieves the second-best results. In Hameln, *LuNet-lite-JO-T* outperforms *LuNet-lite-JO-MS* by up to 2.4% in terms of OA and +2.8% in terms of mean F1 score; the corresponding numbers are +3.2% (OA) and +2.5% (mean F1) in Schleswig. However, the method combining the two approaches (*LuNet-lite-JO-ENS*) delivers the best results in terms of both OA and mean F1 score over all semantic levels in both sites. Compared to the baseline, the increase is up to +1.8% in OA and + 3.2% in mean F1 in Hameln (1.6% and 3.2% in OA and mean F1, respectively, in Schleswig). The improvement in OA is relatively constant across all semantic levels. However, there is a tendency for the improvement of the mean F1 scores to become larger as the semantic level increases. The largest improvements in terms of the mean F1 score occur at level III in both sites (about 3%). The main benefit of adding the multi-scale patches for small objects to the classification thus seems to be related to a better performance for underrepresented classes in the finest semantic level of the object catalogue.

Network variant	Category level					
	I		II		III	
	OA	$\overline{F1}$	OA	$\overline{F1}$	OA	$\overline{F1}$
Hameln						
LuNet-lite-JO-T	91.5	85.9	78.5	59.9	74.1	51.8
LuNet-lite-JO-MS	91.1	83.1	76.1	58.5	71.7	50.2
LuNet-lite-JO-ENS	**93.0**	**87.5**	**80.0**	**61.5**	**75.9**	**54.7**
Schleswig						
LuNet-lite-JO-T	91.0	85.4	74.8	61.8	70.4	55.3
LuNet-lite-JO-MS	90.6	83.7	71.5	59.3	67.2	53.1
LuNet-lite-JO-ENS	**92.2**	**86.5**	**76.1**	**63.8**	**72.0**	**58.5**

Table 2: Overview of the results of hierarchical land use classification achieved by *LuNet-lite-JO-T, LuNet-lite-JO-MS* and *LuNet-lite-JO-ENS* for Hameln and Schleswig. $\overline{F1}$: mean F1 score [%], OA: Overall Accuracy [%]. Best scores are shown in bold font.

Looking at the F1 scores of all classes (Tab. 3 and Tab. 4), it can again be observed that *LuNet-lite-JO-T* outperforms *LuNet-lite-JO-MS* in most indices over all levels, and the ensemble method delivers better results than the baseline in most cases. In Hameln, the F1 scores of all categories at level I are increased at least by 1.4% by the ensemble method. At level II, 10 out of 14 categories are better recognised, with increases of F1 score up to +6.7% (class *moor or swamp*). At level III, 15 out of 21 categories are also better identified, with a maximum increase of +14.3% (class *extended residential*) in terms of F1 score. Similar behaviours of improvement are observed in the results for Schleswig, where the maximum increases of F1 scores from the coarsest level to the finest level are +2.0%, +6.1% and +13.4%. *LuNet-lite-JO-MS* achieves the best results for very few classes, e.g. *parking lot* in Hameln at level II. It has to be noted that the class-wise F1 scores for some classes at levels II and III are not satisfactory yet. These problems affect underrepresented classes (e.g. *sport & leisure area* at level III) and classes which have a similar appearance in the data, as already observed in (Yang et al., 2020b). In summary, these results reveal that in principle both strategies for patch generation are well-suited for the purpose of land use classification, but the method based on tiling performs slightly better than the one based on multi-scale patch generation. This

level I				level II				level III			
class	***T***	***MS***	***ENS***	**class**	***T***	***MS***	***ENS***	**class**	***T***	***MS***	***ENS***
Settlement	92.1	92.0	**93.8**	*res.*	86.9	86.0	**88.2**	*res.use*	89.7	88.6	**90.3**
								ext. res.	36.1	43.6	**50.4**
				ind.	69.7	72.9	**74.5**	*fact.*	34.4	**49.9**	44.5
								busi.	52.2	52.9	**56.5**
								infra.	29.9	39.7	**42.3**
				mix	0	0	0	*mix*	0	0	0
				special	52.2	53.0	**53.9**	*special*	52.2	53.0	**53.9**
				rec.	75.5	72.4	**77.4**	*leis.*	6.2	6.3	**6.9**
								park	76.7	72.8	**78.9**
traffic	92.5	92.0	**93.9**	*ro.traf.*	83.6	80.7	**84.5**	*mo. road*	86.4	84.1	**86.9**
								traf. area	63.7	60.1	**67.2**
				path	**84.7**	77.4	84.1	*path*	**84.7**	77.4	84.1
				park.lot	50.9	**55.2**	50.9	*park.lot*	50.9	**55.2**	50.9
vegetation	83.0	80.7	**84.6**	*agr.*	**87.4**	83.7	87.2	*farm*	**85.4**	71.5	83.1
								garden	**68.7**	61.4	68.2
				forest	84.0	71.7	**84.3**	*h/s.wood*	59.9	59.4	**64.1**
								h&s.wood	**47.6**	13.1	44.2
				grove	54.0	57.3	**57.9**	*grove*	54.0	57.3	**57.8**
				moor	26.7	**34.6**	33.4	*moor*	26.7	**34.6**	33.4
water	76.1	67.8	**77.6**	*flow.wat.*	74.1	64.9	**74.9**	*flow.wat.bo*	74.1	64.9	**74.9**
				stag.wat.	9.2	9.0	**10.0**	*stag.wat.bo.*	9.2	9.0	**10.0**

Table 3: F1 scores [%] for individual classes at all semantic levels in Hameln, achieved by the three variants *LuNet-lite-JO-T (T)*, *LuNet-lite-JO-MS (MS)* and *LuNet-lite-JO-ENS (ENS)*. The best values per class are printed in bold font.

level I				level II				level III			
class	*T*	*MS*	*ENS*	class	*T*	*MS*	*ENS*	class	*T*	*MS*	*ENS*
settlement	93.3	91.9	**94.0**	*res.*	87.1	84.3	**87.5**	*res.use*	88.4	85.4	88.4
								ext. res.	68.4	67.1	**74.5**
				ind.	57.9	56.2	**64.2**	*fact.*	0	**17.9**	7.9
								busi.	53.9	53.0	**59.8**
								infra.	32.4	32.9	**45.6**
				mix	**27.6**	20.6	27.5	*mix*	**27.6**	20.6	27.5
				special	**38.6**	33.8	36.2	*special*	**38.6**	33.8	36.2
				rec.	68.9	63.7	**69.9**	*leis.*	52.3	31.8	**59.9**
								park	68.6	64.5	**70.1**
traffic	90.5	90.8	**92.5**	*ro.traf.*	84.8	82.6	**85.8**	*mo. road*	88.7	86.5	**89.8**
								traf. area	42.3	36.8	**45.4**
				path	72.5	63.0	**75.8**	*path*	72.5	63.0	**75.8**
				park.lot	26.9	**41.0**	33.2	*park.lot*	26.9	**41.0**	33.2
vegetation	**93.6**	91.6	**93.6**	*agr.*	92.0	88.9	**92.5**	*farm*	90.1	84.1	**90.7**
								garden	85.9	84.3	**86.5**
				forest	89.5	86.2	**89.9**	*h/s.wood*	46.0	**50.7**	47.4
								h&s.wood	58.3	52.3	**58.9**
				grove	58.6	48.9	**60.7**	*grove*	58.6	48.9	**60.7**
				moor	63.1	57.0	**67.0**	*moor*	63.1	57.0	**67.0**
water	64.2	60.5	**65.9**	*flow.wat.*	26.0	**39.6**	27.2	*flow.wat.bo*	26.0	**39.6**	27.2
				stag.wat.	72.4	64.9	**76.4**	*stag.wat.bo.*	72.4	64.9	**76.4**

Table 4: F1 scores [%] for individual classes at all semantic levels in Schleswig, achieved by the three variants *LuNet-lite-JO-T (T)*, *LuNet-lite-JO-MS (MS)* and *LuNet-lite-JO-ENS (ENS)*. The best values per class are printed in bold font.

may be due to the fact that the tiled versions preserve the geometrical resolution well, in particular for large objects. However, the, combination of both types of patch generation for the classification of database objects performs best in terms of OA and mean F1 score, indicating that both approaches are complementary to each other to a certain degree.

4.2.2 Influence of object size: In (Yang et al., 2021) object size was found to have a major impact on the classification accuracy: small objects are less frequently classified correctly. Therefore, generating multiple patches for small polygons based on the multi-scale approach is expected to improve the classification of small polygons. To validate this assumption, the differences of the OA and mean F1 scores between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-T* and between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-MS* at all semantic levels are computed. These differences are presented as a function of object area in Figs. 4 and 5. The area unit $A = 2621\ m^2$ is the area of a CNN patch of 256 x 256 pixels at a GSD of 20 cm. In the figures, positive differences correspond to an increase of the quality metric. Tab. 5 presents the area categories and the statistics about the numbers of polygons in each category.

Looking at the differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-MS* (checkered bars), using the ensemble leads to an increase in OA at all levels and for polygons of different sizes in both sites, and the improvement for large polygons corresponding to more than one patch according to the tiling approach (which are in the categories 2A-3A and >3A) is larger than the one for smaller polygons only corresponding to a single tiled patch (category < A). The maximum increase is about 6% in Hameln, occurring at level III in category 2A-3A, and about 8% in Schleswig, occurring at level II and also in category 2A-3A.

There is a similar picture for the mean F1 scores, except that in Hameln there is decrease at level III in category A-2A.

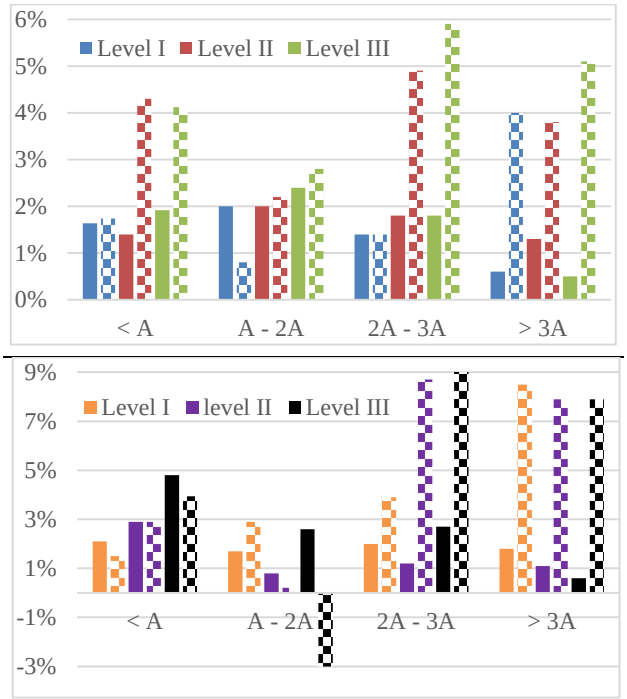

Figure 4: Differences of OA (top) and mF1 (bottom) between different network variants in Hameln. Solid bars: differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-T*; checkered bars: differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-MS*.

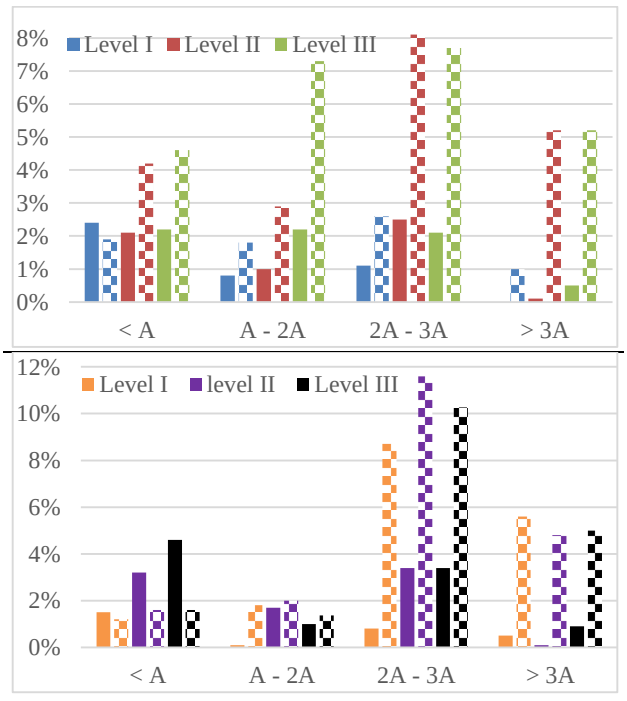

Figure 5: Differences of OA (top) and mF1 (bottom) between different network variants in Schleswig. Solid bars: differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-T*; checkered bars: differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-MS*.

Turning the focus on the differences between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-T* (solid bars), the increase caused by using the ensembe is lower than the one between *LuNet-lite-JO-ENS* and *LuNet-lite-JO-MS* in most cases, because the network based on the tiling approach perform better than the multi-scale one (cf. Section 4.2). In Hameln, the accuracy for polygons having an area smaller than A are improved by at least 1.5% over all semantic levels. The maximum increase of 2.3% at level III occurs in the category of polygons with an area of A-2A. As the object size increases further, there is still an improvement in accuracy, but it becomes smaller. In Schleswig, there is a similar tendency for the increase in OA due to using the ensemble method. We see that in both sites, the maximum increase occurs with polygons having an area smaller than A, and it is at least 2.1%. Switching the focus to the mean F1 score, we observe a similar behaviour in both test sites. The most significant increase occurs at level III with polygons smaller than A in both sites; this increase is 4.8% in Hameln and 4.6% in Schleswig. Note that it is exactly this group of polygons for which the multi-scale approach will generate additional patches. In conclusion, the proposed multi-scale approach helps in the classification of all polygons when it is combined with the tiling approach, and on average, small polygons benefit more than large ones from the combination.

Object Size	**Hameln**		**Schleswig**	
	#Polygons	#avg.Tiles	#Polygons	#avg.Tiles
< A	1757	1.4	1754	1.4
A – 2A	513	3.0	768	2.8
2A – 3A	222	4.6	393	4.5
> 3A	453	10.1	1430	15.5

Table 5: Number of polygons (#Polygons) and avegrage tiles (#avg.Tiles) generated in tiling approach as a function of object size in Hameln and Schleswig.

5. CONCLUSION

In this paper, we have proposed an additional multi-scale approach for land use classification to address the problem of a poor classification performance for small polygons. The experimental results show that the integration of the multi-scale approach does improve the classification performance indeed, with improvements of up to +1.8% in terms of OA and +3.2% in terms of mean F1 score, and the categories at the finest semantic level are improved most. Furthermore, the integration of the multi-scale approach improves the classification of polygons differently according to their size. The average of the mean F1 scores over all semantic levels increases by the largest amount for small polygons, i.e. those for which the new approach generates additional multi-scale patches. We believe that this observation validates the effectiveness of the proposed approach. In addition, we also observed an increase in performance for larger polygons.

In the current version of our method we train and test the networks for patches genereated using the tiling and multi-scale approaches and combine the results by decision level fusion in the ensemble method. To achieve an end-to-end learning framework, in future work we strive to combine both types of patches in one unified CNN model, e.g. by combining the patches directly to form a larger training dataset or by developing a joint network architecture with two branches. Another interesting point is to increase the number of training samples, which is a pre-requisite for reliable results. However, manual annotation of large areas is time-consuming and expensive. One possibility is

to derive the training labels from an outdated geospatial databases, though in this case one has to cope with annotation errors (label noise) (Kaiser et al., 2017). Strategies to mitigate these errors in the class labels of training samples can be developed and integrated in the learning model, e.g. (Maas, et al., 2019).

ACKNOWLEDGEMENT

We thank the Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), the Landesamt für Vermessung und Geoinformation Schleswig-Holstein (LVermGeo) and Landesamt für innere Verwaltung Mecklenburg-Vorpommern (LaiV-MV) for providing the test data and for their support of this project.

REFERENCES

Abadi, et al., 2015. Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org (accessed 09/04/2021).

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2008. ALKIS®-Objektartenkatalog 6.0. Available online (accessed 27/01/2021): http://www.adv-online.de/GeoInfoDok/GeoInfoDok-6.0/Dokumente/

Audebert, N., Saux, B. L., Lefevre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140: 20-32

Comaniciu, D., Meer, P., 2002: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5):603-619.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017: Densely connected convolutional networks. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700-4708.

Long, J., Shelhamer, E., Darrell, T., 2015: Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017: Learning aeiral image segmentation from online maps. IEEE Transactions on Geoscience and Remote Sensing 55(11): 6054-6068.

Kokkinos, I., 2016. Pushing the boundaries of boundary detection using deep learning. In: International Conference on Learning Representations (ICLR).

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS'12), Vol. 1, pp. 1097-1105.

Maas, A., Rottensteiner, F., Heipke, C., 2019. A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. Computer Vision and Image Understanding 188, paper 102782.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing 135: 158–172.

Yang, C., Rottensteiner, F., Heipke, C., 2018: Classification of land cover and land use based on convolutional neural networks. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-3, pp. 251-258

Yang, C., Rottensteiner, F., Heipke, C., 2019: Towards better classification of land cover and land use based on convolutional neural networks. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W13, pp. 139-146

Yang, C., Rottensteiner, F., Heipke, C., 2020a: Exploring semantic relationships for hierarchical land use classification based on convolutional neural networks. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences V-2, pp. 599-607

Yang, C., Rottensteiner, F., Heipke, C., 2020b: Investigations on skip-connections with an additional cosine similarity loss for land cover classification. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. V-3, pp. 339-346.

Yang, C., Rottensteiner, F., Heipke, C., 2021: A hierarchical deep learning framework for the consistent classification of land use objects in geospatial databases. ArXiv pre-print: arXiv:2104.06991.

Zhang, C., Atkinson, P.M., 2016. Novel shape indices for vector landscape pattern analysis. International Journal of Geographic Information Science 30, 2442–2461.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural networks (OCNN) for urban land use classification. Remote Sensing of Environment 216: 57-70.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019. Joint deep learning for land cover and land use classification. Remote Sensing of Environment 221: 173-187.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5(4): 8-36.