IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Spatialising uncertainty in image segmentation using weakly supervised convolutional neural networks: a case study from historical map processing

ISSN 1751-9659
Received on 15th January 2018
Revised 11th May 2018
Accepted on 2nd July 2018
E-First on 6th September 2018
doi: 10.1049/iet-ipr.2018.5484
www.ietdl.org

Johannes H. Uhl[1] ✉, Stefan Leyk[1], Yao-Yi Chiang[2], Weiwei Duan[2], Craig A. Knoblock[2,3]

[1]*Department of Geography, University of Colorado Boulder, Guggenheim 110, 260 UCB, Boulder, CO, USA*
[2]*Spatial Sciences Institute, University of Southern California, 3616 Trousdale Parkway, Los Angeles, CA, USA*
[3]*Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA, USA*
✉ *E-mail: johannes.uhl@colorado.edu*

Abstract: Convolutional neural networks (CNNs) such as encoder–decoder CNNs have increasingly been employed for semantic image segmentation at the pixel-level requiring pixel-level training labels, which are rarely available in real-world scenarios. In practice, weakly annotated training data at the image patch level are often used for pixel-level segmentation tasks, requiring further processing to obtain accurate results, mainly because the translation invariance of the CNN-based inference can turn into an impeding property leading to segmentation results of coarser spatial granularity compared with the original image. However, the inherent uncertainty in the segmented image and its relationships to translation invariance, CNN architecture, and classification scheme has never been analysed from an explicitly spatial perspective. Therefore, the authors propose measures to spatially visualise and assess class decision confidence based on spatially dense CNN predictions, resulting in continuous decision confidence surfaces. They find that such a visual-analytical method contributes to a better understanding of the spatial variability of class score confidence derived from weakly supervised CNN-based classifiers. They exemplify this approach by incorporating decision confidence surfaces into a processing chain for the extraction of human settlement features from historical map documents based on weakly annotated training data using different CNN architectures and classification schemes.

1 Introduction

The renaissance of convolutional neural networks (CNNs) and other machine learning methods for recognition tasks in computer vision has also catalysed the application of such frameworks for information extraction tasks in the geospatial sciences. Recently, approaches for object detection, scene classification, and semantic segmentation have been applied to remotely sensed geospatial data and have shown promising results outperforming traditional methods [1, 2]. Whereas most contributions focus on the training and evaluation of their approaches using benchmark datasets (e.g. [3, 4]), only a few efforts tackle the challenging task of applying these approaches to real-world data. Many of the popular benchmark datasets come with pixel-level reference data, often generated through manual efforts that can be used to train and validate the models within a controlled environment.

However, accurate and abundant training labels at the pixel level (Fig. 1*a*, left) are often not available in real-world application scenarios. Training labels at the patch level (Fig. 1*a*, right) describing the semantic content of a patch of an image rather than a pixel are typically easier to obtain. If only patch-level labels are available to train a model for pixel-level inference, we speak of weakly supervised learning or weakly annotated training data [5]. State-of-the-art segmentation architectures that require pixel-level training annotations such as proposed in [6, 7] cannot be used directly in such cases.

Using weakly annotated training data for segmentation tasks at the pixel level has one major shortcoming, which is the translation invariance property of the CNNs. It causes a loss in spatial granularity since the image content in a (static) image patch around a centre pixel influences the predicted class of the centre pixel (Fig. 1*b*). This typically results in a spatially inflated segmentation of less spatial details (Fig. 1*c*). In response to such shortcomings, researchers have developed methods that rely on expectation–maximisation techniques [8], alternative pooling techniques [9], conditional random fields [10], and superpixel-based segmentation methods [11]. The quantitative analysis of translation invariance, its consequences for predictions, and how it is impacted by CNN architecture, training configuration and chosen hyperparameters is an active research field. In [12] for example, translation invariance is quantified using translation-sensitivity maps based on augmented (i.e. artificially shifted) training data.

However, the effects of translation invariance have not explicitly been studied from a spatial point of view which is surprising given the growing number of deep learning applications in the geospatial sciences (e.g. [13–18]). Hence, in this contribution, we propose a framework for the spatially explicit analysis of such effects and how they relate to training data, CNN architecture, and chosen hyperparameters. For this purpose, we propose a visual-analytical approach based on uncertainty surfaces derived from dense pixel-wise CNN predictions reflecting the spatial variation of class decision confidence of a weakly supervised CNN. The analysis is exemplified by the extraction of human settlement features from scanned historical topographic map documents.

Deep learning techniques have increasingly been applied for information extraction from earth observation data, and this naturally projects into the idea of applying such techniques to other types of geospatial data. Historical maps contain valuable information on the evolution of natural, environmental or human-induced geographic processes. Map processing aims to systematically extract such spatial information and transition it into analysis-ready digital data formats [19].

Efficient graphics recognition of historical maps is impeded to date, mainly due to the issues of poor graphical quality and large data volume, which is a common problem when thousands of historical map sheets are scanned and stored in digital map archives. To overcome the need for user intervention and manual training in a recognition system, we are developing techniques to fully automate the process of extracting geographic information

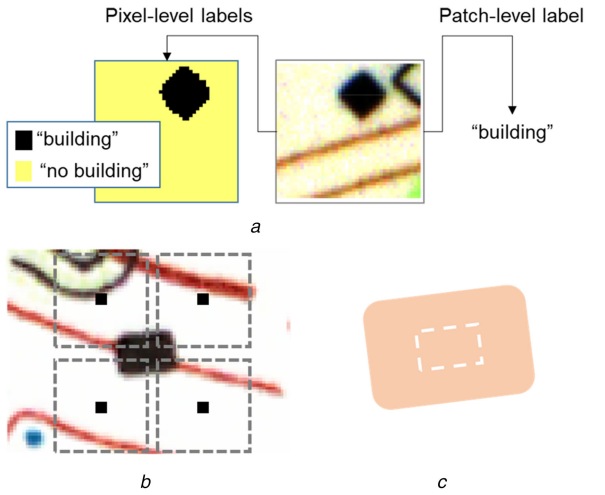

Fig. 1 *The concept of weakly annotated training labels and its effect on segmentation results*

(a) Examples of pixel-level training labels and a weakly annotated patch-level training label, ***(b)*** Spatial granularity loss in weakly supervised image segmentation: centre pixel of the CNN input patches are labelled as class 'building' due to the translation invariant property, ***(c)*** Resulting spatially inflated segmentation (red) overlaid with original object outline (dashed)

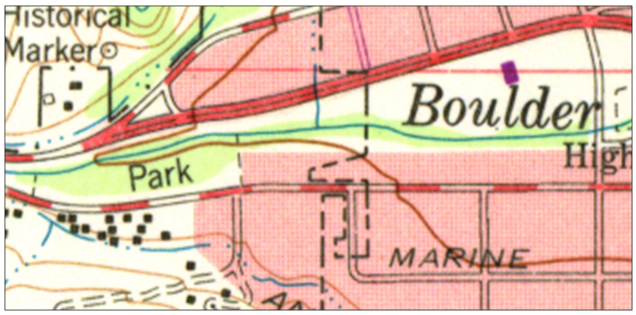

Fig. 2 *Subset of the Boulder map (1966) used in these experiments*

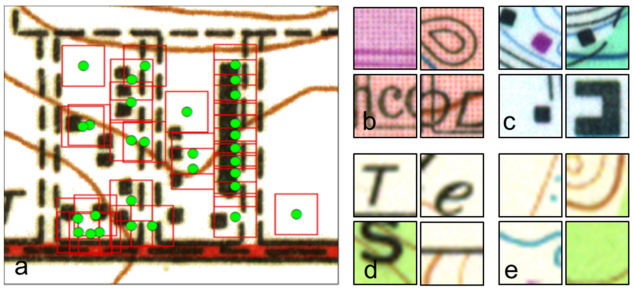

Fig. 3 *Training data collection*

(a) Contextual settlement data extracted from the Zillow database with derived sample patch extents and underlying map data, and exemplary extracted training samples at patch-level for the four classes, ***(b)*** Urban area, ***(c)*** Individual buildings, ***(d)*** Black negative, ***(e)*** Other negative content

from scanned historical cartographic documents [20, 21]. The goal of such information extraction efforts is to make the data in these documents accessible for spatial-temporal analysis of landscape patterns and their changes. One approach to improve recognition performance is to incorporate contextual geographic layers to make use of the fact that map series represent evolutionary documents that change in cumulative ways [22].

The concept of geographic context implies the effective use of ancillary geographic information that contains the feature of interest such as gazetteers or other map series for guided graphics sampling in training a recognition model [23–25]. For example, it can be assumed that many building symbols in a historical map spatially overlap or are in close proximity to building objects in a contemporary geographic dataset. Thus, sampling nearby the contemporary building objects enables a system to collect graphic examples of building symbology in historical maps and facilitates creating high-quality training data at the image patch level.

In this study, we first demonstrate the use of the geographic contextual data to guide graphics sampling for automatically generating training labels at the image patch level, which are then used for the extraction of building symbols and urban areas in historical map sheets by performing semantic segmentation at the pixel level using different CNN architectures and classification configurations. Since the training labels are obtained for sample image patches, this process constitutes a typical case of weakly annotated training data. Secondly, we validate the resulting segmentation against manually created reference data, and then demonstrate how spatial uncertainty measures derived from the CNN class scores can be employed to (i) assess translation invariance-induced effects on the segmentation results and (ii) identify potentially ill-trained CNN classifiers resulting from inappropriate configuration, training, or classification scheme. These effects on the accuracy of the resulting semantic segmentation are systematically evaluated from a spatial point of view.

2 Data and methods

2.1 Historical topographic maps

Recently, several historical (topographic) map series have been made available to the public (e.g. [26, 27]). The United States Geological Survey (USGS) has scanned more than 180,000 historical map sheets and stored the entire map series in a digital archive. While urban areas in the USGS map sheets are similarly coloured and textured areas, building symbols are shown as small black rectangles or more complex polygons (Figs. 2 and 3*c*). We tested our approach on a map sheet of Boulder, Colorado (1966) at a map scale of 1:24,000 scanned at a resolution of ∼500 dpi (dots per inch) in the RGB colour space.

2.2 Automated training data generation

In [28], we proposed an approach for automatic training data generation based on cadastral parcel records and building footprint data as contextual information. The temporal information of when a building has been established can be derived from the parcel data attributes and allows for reconstructing the spatial distribution of parcel units with built structures at a given point in time. Spatially refining these parcel boundaries with high-resolution LiDAR-derived building footprint data makes it possible to create snapshots of the existing buildings at different points in time, allowing for the creation of training samples of settlement features with relatively high reliability. However, such rich training data based on LiDAR are available for only a selected number of counties in the U.S., restricting training data generation to those regions.

To generate training data from maps covering larger spatial extents, we used a settlement location database derived from the Zillow Transaction and Assessment Dataset [29] in this experiment. This database contains geographic coordinates of approximated address points and information about the year when a structure was built, which allows for reconstructing building locations at a fine temporal resolution, but with a lower spatial accuracy (Fig. 3*a*). For this study, our automatic approach created training samples for 15 topographic map sheets of five different locations in Colorado (USA) and three points in time to test different geographic settings and changing the cartographic design that may affect the way how settlements are cartographically represented in the maps. These 15 map sheets include the test map shown in Fig. 2. Based on the locations given in the contextual data, we collected the candidate samples of the underlying map document. The cropped sample patches have dimensions of 42 × 42 pixels (corresponding to ∼50 × 50 m).

We expected the recognition of urban areas to be straightforward due to the dominating uniform background colours (Fig. 3*b*). A major expected challenge for the recognition of individual buildings (Fig. 3*c*) was the discrimination of building features from the other black content in the map, such as black text

elements, whereas in [28] the extraction was considered a three-class problem (i.e. individual buildings, urban areas, negative class), in this study, we split the negative map content into black non-building objects (Fig. 3*d*) and the remaining not-black content as a fourth class (Fig. 3*e*). This resulted in a total of two positive classes (i.e. individual building and urban area) and two negative classes (i.e. negative black and negative other). We implemented this separation using threshold-based, unsupervised image processing methods, as described in the following. The effect of reducing the in-class variability of the two negative classes will be discussed in the Results section, where we will evaluate the results based on both classification schemes. The training labels are automatically derived (i) based on the proximity of sample locations to contextual data points and (ii) based on the content of the samples. The spatial proximity of sample locations to the contextual data points indicates the likely class membership (close = positive candidate samples, distant = negative candidate samples).

However, to determine the exact label (e.g. a positive candidate sample can be a member of the individual building class or belong to the urban class, or neither of them), we used an unsupervised hierarchical classification approach to examine the content of the candidate samples, involving a variety of image processing techniques (see Fig. 4). In addition to the label assignment, this procedure also performed sample cleaning, since discrepancies between map data and the used contextual data may cause incorrect training samples (e.g. some training samples for the building class do not contain a building object).

This fully automatic multi-stage workflow first identified samples containing urban areas based on colour segmentation and identification of the dominant colour (step 1) from the set of positive candidate samples. The procedure then tests remaining samples for the black content. Gaussian filtering and, additionally to the method proposed in [28], morphologic operations were used to filter out the irrelevant black objects based on their size, and edge detection was used to quantify the complexity of the remaining black objects.

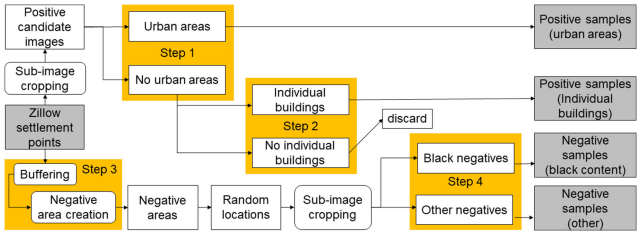

Fig. 4 *Workflow of the automated training data generation*

Here, the black objects of high complexity were considered more likely to represent text, and the remaining dark objects were considered individual building symbols. They were identified using SIFT [30] key point detection applied to the filtered sample image (step 2). Furthermore, we extended the method described in [28] to distinguish between black negative and other negative samples. From the pool of negative candidate samples (i.e. samples collected at locations distant from the contextual data where no settlements are expected), we assigned labels for the respective negative classes based on the number of black pixels in the samples. Since contrast levels determined by the scanning process and cartographic styles may vary across different map sheets, there is a potential risk of assigning incorrect labels. Hence, we visually assessed the results of this unsupervised training data generation using t-distributed stochastic neighbourhood embedding (t-SNE, [31]) plots, where the random subsets of the training samples are arranged in a two-dimensional space in a way that similar samples of the same class are located nearby (Fig. 5). This allows for a quick and efficient visual assessment of the correctness and representativeness of the created training samples in all four classes.

2.3 Semantic segmentation using weakly supervised CNNs

Following the law of parsimony, we tackled the map segmentation task, firstly, using the simplest available CNN architecture, namely the classical LeNet model with two convolutional, two pooling, and two fully connected layers that has proven good performance on simple image recognition tasks [32] and can be trained on a common desktop machine without the need for high-performance graphical processing unit servers (Fig. 6*a*). We compared this light-weight approach with a VGGNet-S model, which is a shallow variant of the VGGNet with a total of 11 layers (five convolutional layers, three pooling layers, and three fully connected layers, [33], see Fig. 6*b*).

This comparison will allow for examining how sensitive the model performance is to increase the depth of the network as an outlook for future directions. In the following experiment, we produced and compared the segmentation results for four different scenarios (Table 1).

First, we employed LeNet trained on 30,000 samples from the same individual map that is used for inference. The training labels represented three classes (i.e. urban, individual buildings, and a non-settlement class) which we named LeNet A (see [28]). In the second scenario (LeNet B), we included a fourth class as previously described to reduce in-class variability in the negative classes, keeping the same training samples as in the LeNet A scenario still applied to one map page. The third scenario (LeNet C) used the four training labels used in LeNet B but incorporated a

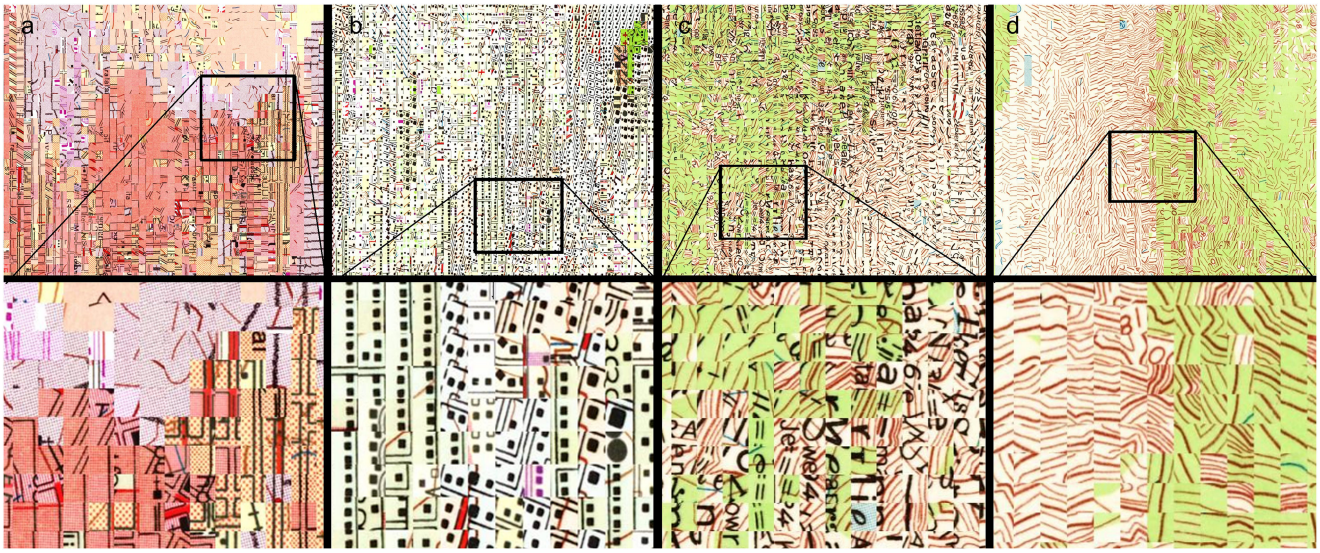

Fig. 5 *t-SNE plots and corresponding enlargements for visual assessment of the automatically generated training samples*
(*a*) Urban areas, **(*b*)** Individual buildings, **(*c*)** Negative black, **(*d*)** Negative other class

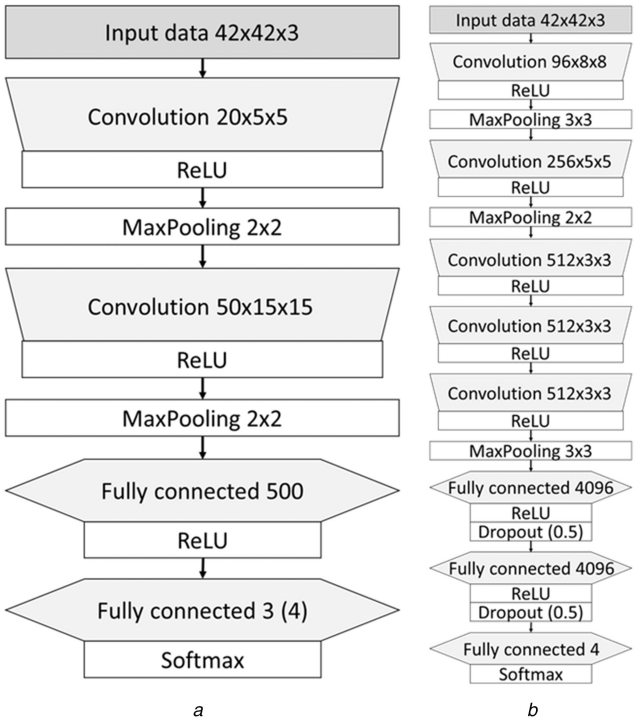

Fig. 6 *Architectures of the CNNs used in this study*
(*a*) LeNet, **(*b*)** VGGNet-S

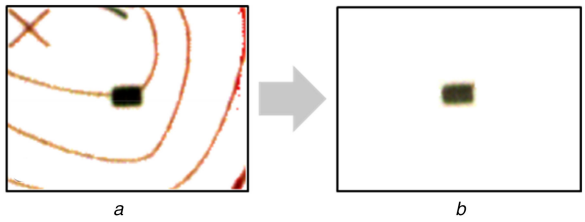

Fig. 7 *Background noise removal for explicit analysis of salient features*
(a) Original and, ***(b)*** the manually cleaned map subset to assess spatial variability of the CNN prediction behaviour

much larger training dataset ($N = 400{,}000$) from 15 different map sheets. Finally, the fourth scenario used the same four class labels and the same large training dataset ($N = 400{,}000$) from 15 map pages, but this time we trained the above described VGGNet-S model.

These four scenarios covered different combinations of training sample size, number of classes, CNN architecture, and number of training epochs and allowed for assessing the relationship between these settings and the resulting segmentations and corresponding uncertainty measures. For each trained CNN, we generated dense pixel-wise predictions with a stride of one pixel and registered the obtained class scores at each pixel location. In addition to the subset shown in Fig. 2, we generated a small de-noised test patch that only contains an individual building object. We manually removed all other map content to investigate and illustrate the spatial behaviour of the CNNs across this patch focusing on the salient feature without background noise (Fig. 7). We used this test patch to visually assess the class scores and uncertainty measures for each of the four scenarios.

2.4 Deriving uncertainty measures

To evaluate the performance of the different segmentations, we generated overall accuracy (OA) measures by performing pixel-wise map comparison to a manually generated reference dataset for the test map extent described before. We generated confusion matrices for each scenario, and derived accuracy measures, measuring *external uncertainty*, since these measures are obtained by comparison to independently generated validation data. These measures include the class-independent measures OA, the κ index of agreement [34], and the entropy-based normalised mutual information index (NMI, [35]).

Since OA (i.e. the proportion of correct predictions among all predictions), tends to yield inflated values in the case of imbalanced class proportions [36], we also reported κ index and NMI that describe overall agreement in more conservative but more robust ways, accounting for chance agreement (κ) and based on mutual dependence between validation and test data (NMI), respectively. κ index is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

with $p_0 = OA$ and p_c representing agreement by chance

$$p_c = \frac{1}{N^2} \sum_k n_{k\,\mathrm{Ref}} n_{k\,\mathrm{Test}}, \tag{2}$$

where N is the number of overall predictions and n_k are the frequencies how often class k is predicted by the test data and the reference data, respectively. NMI is based on the entropies of reference class labels $H(R)$, of predicted labels $H(P)$ and the joint entropy $H(R, P)$ as

$$\mathrm{NMI} = 1 - \frac{H(R,P) - H(P)}{H(R)}. \tag{3}$$

Additionally, we reported class-specific accuracy measures including precision and recall [37], whereas the class-independent measures characterise the overall capacity of the CNNs to reproduce the validation data and allow for quantitative comparison between different segmentation results, the class-specific measures provide insights into the degree of confusion between individual classes constituting valuable information for model improvement. These measures describe external uncertainty in a non-spatial manner. In addition to that, we created several spatial layers to analyse spatial variability in the internal confidence of the CNNs in discriminating between the different classes at each location. The spatial variables computed are (i) class score surfaces (score maps) for each class and (ii) several pixel-wise measures to characterise the CNN decision confidence, such as the difference between the highest and the second highest class score S_1–S_2, the ratio of the lowest and the highest class score S_{min}/S_{max}, and the entropy $H(S)$ of the class scores. These variables measure *internal uncertainty* since they characterise internal decision confidence without taking into account external reference data. These measures can be used to describe decision confidence at each pixel location regarding the absolute difference, the relative difference, and the variety of class scores, respectively. The resulting surfaces describing the spatial variation of class decision confidence can be used as a visual-analytical tool to assess translation invariance, but also to diagnose overfitted, underfitted, or ill-trained classifiers. We visually compared the surfaces

Table 1 CNN scenarios

CNN configuration	Training samples	Classes	Training epochs	Learning rate	Step size
LeNet A	30k	3	10	0.001	0.0001
LeNet B	30k	4	10	0.001	0.0001
LeNet C	400k	4	20	0.001	0.0001
VGGNet-S	400k	4	60	0.001	0.0001

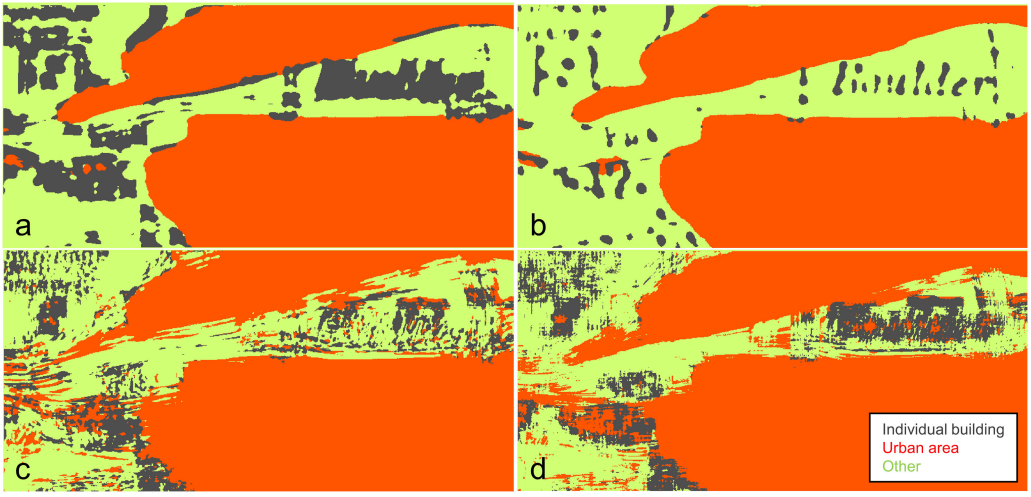

Fig. 8 *Segmentation results for the four scenarios*
(*a*) LeNet A, **(*b*)** LeNet B, **(*c*)** LeNet C, and **(*d*)** VGGNet-S

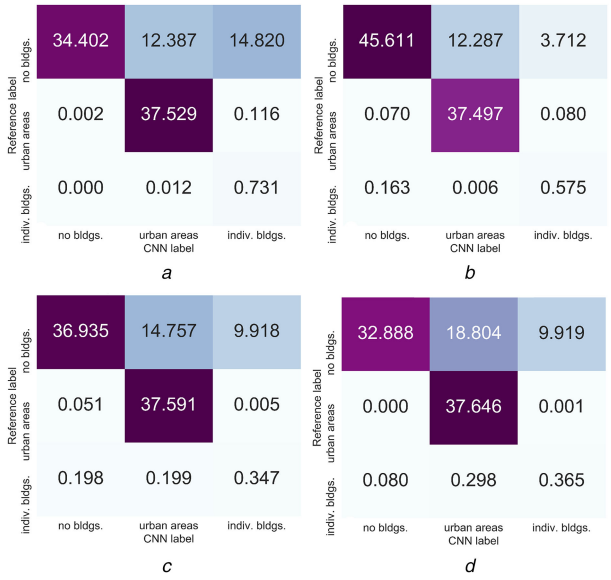

Fig. 9 *Confusion matrices for the four scenarios, displayed in per cent of the total study area*
(*a*) LeNet A, **(*b*)** LeNet B, **(*c*)** LeNet C, **(*d*)** VGGNet-S

generated from these three variables in order to assess their spatial variations and their capacity to effectively describe the decision confidence.

3 Results

3.1 Semantic segmentation

To create segmentations of the test map, we assigned the class label of the highest class score to each pixel. The segmented maps can be seen in Fig. 8. Since the two negative classes are semantically identical and the validation data is available at a semantic resolution of three classes (i.e. urban areas, individual buildings, and non-settlements), the two negative classes were visualised using the same green colour and were merged for the subsequent external accuracy assessment. Comparing the segmentation results with the original map (Fig. 2), it can be observed that the urban areas are extracted well in the four scenarios, indicating that all CNNs learned to distinguish urban areas from the remaining classes, reliably, based on the colour or texture. Differences can be observed in the smoothness of the boundaries of the urban areas, where the scenarios LeNet A and B show smooth and rounded boundaries, LeNet C and VGGNet-S show more rugged patterns, possibly because the latter two CNNs were trained on a much larger amount of training samples, which increased the complexity of the classification problem.

The detection of individual building symbols differed significantly between the four scenarios, whereas in LeNet A, the CNN apparently only learned to be receptive to colours, resulting in labelling any dark pixel as an individual building, the effect of introducing a 'black non-settlement' class is clearly visible in LeNet B, where the number of false positives (i.e. black text elements labelled as individual buildings) has dropped considerably by reducing the in-class variability of the two negative classes. Interestingly, the deeper VGGNet-S falsely detected more text elements as individual buildings than the LeNet B and LeNet C. One reason might be that VGGNet-S was overfitted due to the extensive training (60 epochs) and learned also from the noise inherent in the training data. This is also indicated by the increased confusion between the urban area and individual buildings in the VGGNet-S segmentation result.

3.2 External uncertainty assessment

We quantified external uncertainty based on the pixel-wise map comparison with the manually digitised (external) reference dataset, i.e. we generated a confusion matrix for each scenario (Fig. 9) to assess the external uncertainty in the created segmentations. Based on these confusion matrices, we derived class-independent accuracy measures (i.e. OA, κ index, and NMI), and class-specific accuracy measures (i.e. precision and recall) (Table 2). It is notable that the confusion between individual buildings and the negative class (i.e. the 'no bldgs.' column) seems to be least for the LeNet B scenario (Fig. 9*b*), resulting in the highest class-independent accuracy measures and the highest precision of the individual building class across the four scenarios.

3.3 Internal uncertainty assessment

In addition to the assessment of external uncertainty by comparing the segmentation results with independently generated reference data, we assessed the 'internal' uncertainty inherent in the decision process of the four scenarios based on the created score maps of each class (i.e. spatial visualisation of the class scores at each pixel location).

To focus on the most challenging features in this experiment (i.e. individual building symbols), we created these score maps also for the de-noised subset in Fig. 7 for each of the four scenarios (Fig. 10).

These score maps allow for analysing the spatial behaviour of the class scores for each CNN. It is notable for the individual building class scores (second column from the left) that the regions around the building symbol show higher class scores in the LeNet C and VGGNet-S scenarios, both trained on a larger amount of training data. This indicates a higher degree of translation invariance, most likely due to the increased variety in the training

Table 2 Accuracy measures derived from the confusion matrices for the four segmentation scenarios

Scenario	Accuracy measures		Class	Precision	Recall
LeNet A	NMI	0.38	no bldgs.	1.00	0.56
	OA	0.73	urban areas	0.75	1.00
	κ	0.54	indiv. bldgs.	0.05	0.98
LeNet B	NMI	0.46	no bldgs.	0.99	0.74
	OA	0.84	urban areas	0.75	1.00
	κ	0.69	indiv. bldgs.	0.13	0.77
LeNet C	NMI	0.37	no bldgs.	0.99	0.6
	OA	0.75	urban areas	0.72	1.00
	κ	0.56	indiv. bldgs.	0.03	0.47
VGGNet-S	NMI	0.33	no bldgs.	1.00	0.53
	OA	0.71	urban areas	0.66	1.00
	κ	0.50	indiv. bldgs.	0.04	0.49

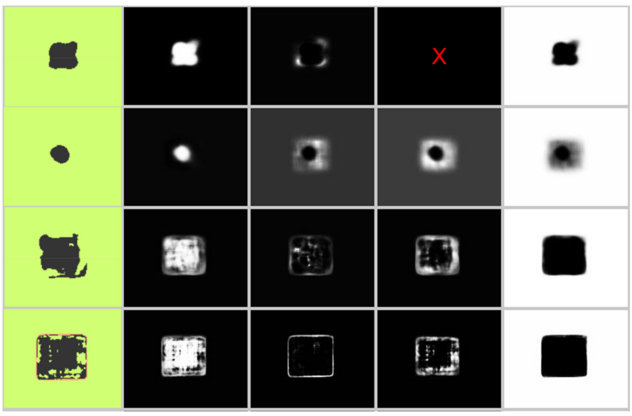

Fig. 10 *Segmentation results (left column) and class score surfaces (highest = white) for the four classes: individual building, urban area, non-settlement black, other non-settlement content (from left to right), for each scenario: LeNet A, LeNet B, LeNet C, and VGGNet-S (from top to bottom)*

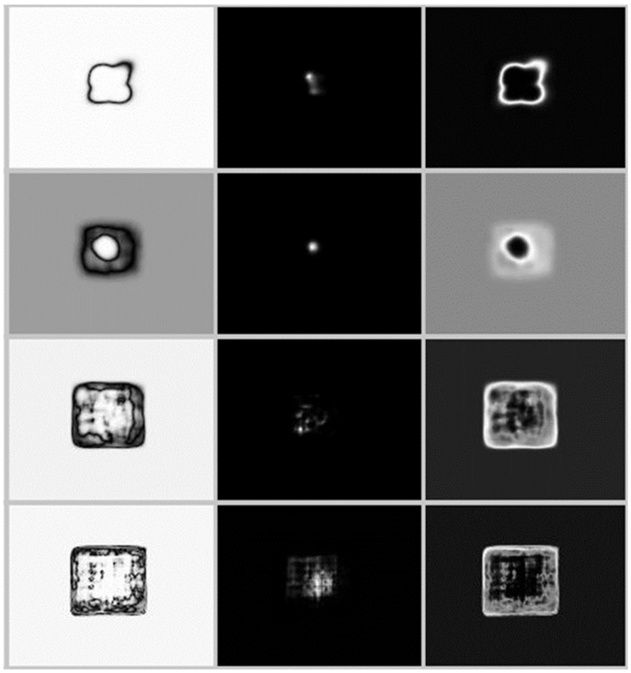

Fig. 11 *Class decision confidence surfaces S_1–S_2, S_{min}/S_{max}, and $H(S)$ (from left to right) for the scenarios LeNet A, LeNet B, LeNet C, and VGGNet-S (from top to bottom, highest = white)*

data and higher generalisation/abstraction capabilities in the case of the deeper VGGNet-S.

Besides the raw class score maps, the internal uncertainty measures derived from the class scores S_1–S_2, S_{min}/S_{max}, and $H(S)$ are shown for the de-noised subset in Fig. 11. These plots provide some insight into the decision confidence across space for each scenario. LeNet A shows linear drops in decision confidence around the building symbol, resulting in a thin dark line in the S_1–S_2, surface.

This indicates a high degree of translation invariance that can be explained by pure receptiveness to colours: as soon as a few black pixels are found in the convoluting window, the decision confidence is high, and the CNN labels the pixel as an individual building.

The S_1–S_2 surface for the LeNet B scenario (second row) shows how class confidence drops at non-settlement locations (i.e. the grey areas) due to the introduction of a second negative class, and how the class confidence increases drastically when the convoluting window is centred at the building object. LeNet B shows very low translation invariance due to the introduction of the black-negative class: only if the building is close to the centre of the search window, the centre pixel is labelled as an individual building. The increasing translation invariance in LeNet C and VGGNet-S is clearly reflected in the larger rectangle around the building location. The diffuse artefacts in the S_1–S_2 surface for LeNet C and VGGNet-S indicate that decision confidence varies with small shifts across the map. This might be a consequence of the CNN being incapable to successfully solve the given classification problem (LeNet C) or overfitted because it has also learned from noise in the data (VGGNet-S).

Additionally, we show S_1–S_2 for the entire map subset (Fig. 12). The patterns, which have to be read in conjunction with the segmentation results in Fig. 8, confirm the observations made for the de-noised subset and show again a highly localised variation in decision confidence for the LeNet C and VGGNet-S scenarios.

4 Conclusions and outlook

This study discusses a spatial approach to uncertainty assessment in image segmentation using weakly supervised CNNs, exemplified by a method for settlement recognition in historical topographic map documents. The described case study is a typical example of weakly supervised learning, where no training data is available a priori, and training annotations have to be generated at the image patch level using an unsupervised, threshold-based labelling method based on image processing techniques. The resulting training annotations are then used to train CNNs for classification at the patch level. However, the trained CNNs are employed for the pixel-level inference, which can be particularly problematic for small objects such as the building symbols. The presented segmentation results, which will systematically be improved in the next steps of this research, and the derived aspatial accuracy measures (e.g. the low-precision values for the individual building class) reflect effects of translation invariance, insufficient generalisation capability, and overfitting, and clearly demonstrate the necessity of post-processing methods such as superpixel-based segmentation approaches [11] in order to increase the correctness and the spatial granularity of the segmentation.

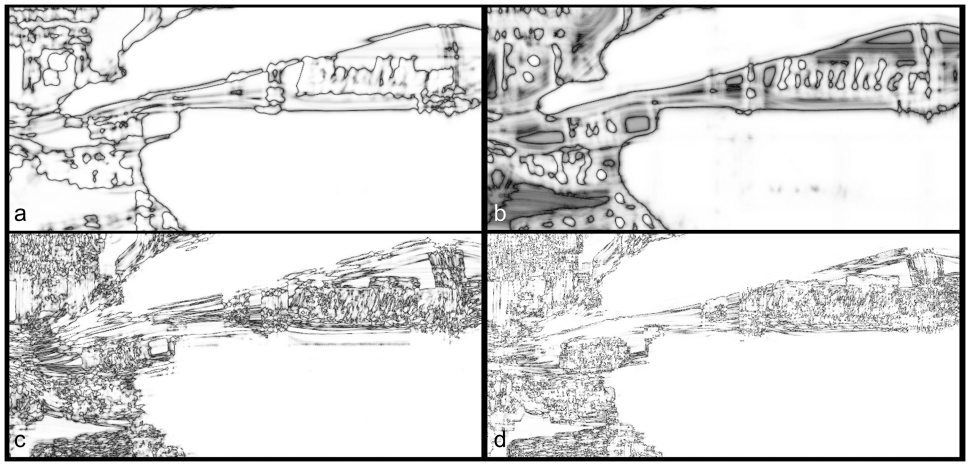

Fig. 12 *Decision confidence using the pixel-wise difference between the highest and second highest class score for the 4 scenarios*
(*a*) LeNet A, **(*b*)** LeNet B, **(*c*)** LeNet C, and **(*d*)** VGGNet-S (highest = white)

Several measures derived from the class score maps that describe class decision confidence are compared visually and show how these decision confidence surfaces indicate uncertainty in the resulting segmentation due to the translation invariance and potentially overfitted or overly shallow CNNs. The created decision confidence maps represent valuable visual-analytical tools to diagnose overfitted, ill-trained or underfitted learners, with a focus on weakly supervised CNNs, however, these tools are applicable to other probabilistic classifiers as well. In addition to that, the decision confidence surfaces illustrate the relationship between translation invariance and depth of the applied CNN architectures.

Future work will include the spatialisation of external uncertainty measures (e.g. focal κ index based on spatially constrained confusion matrices) successfully applied in uncertainty modelling in historical maps [38] and the application of superpixel methods [11] in order to spatially refine the segmented maps created based on weakly annotated training data. The spatially explicit analysis of decision confidence with respect to distance and direction from presumably salient features in the test images can identify the effects of anisotropy in translation invariance, which may indicate systematic offsets in the used training data or effects of lacking rotation invariance of certain CNN configurations. Furthermore, the potential of a multi-stage segmentation approach will be tested, where the pixel-wise predictions from a spatially refined segmentation will be used as pixel-level training data for fully convolutional networks [7], which will allow for end-to-end learning of settlement features in historical topographic maps.

5 Acknowledgments

This material is based on research sponsored in part by the National Science Foundation under grant nos. IIS 1563933 (to the University of Colorado at Boulder) and IIS 1564164 (to the University of Southern California). The authors were provided access to the Zillow Transaction and Assessment Dataset (ZTRAX) through a data use agreement between the University of Colorado Boulder and Zillow Inc. Support by Zillow Inc. is gratefully acknowledged.

6 References

[1] Zhang, L., Zhang, L., Du, B.: 'Deep learning for remote sensing data: a technical tutorial on the state of the art', *IEEE Geosci. Remote Sens. Mag.*, 2016, **4**, (2), pp. 22–40

[2] Ball, J. E., Anderson, D. T., Chan, C. S.: 'Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community', *J. Appl. Remote Sens.*, 2017, **11**, (4), p. 042609

[3] Rottensteiner, F., Sohn, G., Jung, J., *et al.*: 'The ISPRS benchmark on urban object classification and 3D building reconstruction', *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, 2012, **1**, (3), pp. 293–298

[4] Cordts, M., Omran, M., Ramos, S., *et al.*: 'The cityscapes dataset for semantic urban scene understanding'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 2016

[5] Zhou, Z. H.: 'A brief introduction to weakly supervised learning', *Nat. Sci. Rev.*, 2018, **5**, (1), pp. 44–53

[6] Badrinarayanan, V., Kendall, A., Cipolla, R.: 'Segnet: a deep convolutional encoder-decoder architecture for image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **39–12**, pp. 2481–2495

[7] Long, J., Shelhamer, E., Darrell, T.: 'Fully convolutional networks for semantic segmentation'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015, pp. 3431–3440

[8] Papandreou, G., Chen, L. C., Murphy, K. P., *et al.*: 'Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation'. Proc. IEEE Int. Conf. Computer Vision, Venice, Italy, 2017, pp. 1742–1750

[9] Durand, T., Mordan, T., Thome, N., *et al.*: 'Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawaii, USA, 2017

[10] Chen, L.C., Papandreou, G., Kokkinos, I., *et al.*: 'Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs'. 2016, arXiv preprint arXiv:1606.00915. DOI: 10.1109/TPAMI.2017.2699184

[11] Zhao, W., Jiao, L., Ma, W., *et al.*: 'Superpixel-based multiple local CNN for panchromatic and multispectral image classification', *IEEE Trans. Geosci. Remote Sens.*, 2017, **55**, (7), pp. 4141–4156

[12] Kauderer-Abrams, E.: 'Quantifying translation-invariance in convolutional neural networks', 2016. Available at http://cs231n.stanford.edu/reports/2016/pdfs/107_Report.pdf, accessed 8 January 2018

[13] Audebert, N., Le Saux, B., Lefèvre, S.: 'Semantic segmentation of earth observation data using multimodal and multi-scale deep networks'. Proc. Asian Conf. on Computer Vision, Taipei, Taiwan, 2016, pp. 180–196

[14] Maire, F., Mejias, L., Hodgson, A.: 'A convolutional neural network for automatic analysis of aerial imagery'. IEEE Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA), Wollongong, Australia, 2014, pp. 1–8

[15] Castelluccio, M., Poggi, G., Sansone, C., *et al.*: 'Land use classification in remote sensing images by convolutional neural networks'. 2015, arXiv preprint arXiv:1508.00092

[16] Marmanis, D., Datcu, M., Esch, T., *et al.*: 'Deep learning earth observation classification using ImageNet pretrained networks', *IEEE Geosci. Remote Sens. Lett.*, 2016, **13**, (1), pp. 105–109

[17] Romero, A., Gatta, C., Camps-Valls, G.: 'Unsupervised deep feature extraction for remote sensing image classification', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (3), pp. 1349–1362

[18] Scott, G. J., England, M.R., Starms, W.A., *et al.*: 'Training deep convolutional neural networks for land-cover classification of high-resolution imagery', *IEEE Geosci. Remote Sens. Lett.*, 2017, **14**, (4), pp. 549–553

[19] Chiang, Y.-Y., Leyk, S., Knoblock, C.A.: 'A survey of digital map processing techniques', *ACM Comput. Surv. (CSUR)*, 2014, **47**, (1), p. 1

[20] Uhl, J. H., Leyk, S., Chiang, Y.-Y., *et al.*: 'Map archive mining: visual-analytical approaches to explore large historical map collections', *ISPRS Int. J. Geo-Inf.*, 2018, **7**, p. 148

[21] Duan, W., Chiang, Y. Y., Knoblock, C. A., *et al.*: 'Automatic alignment of geographic features in contemporary vector data and historical maps'. Proc. First Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery, Redondo Beach, California, USA, 2017

[22] Leyk, S., Chiang, Y.-Y.: 'Information extraction based on the concept of geographic context'. Proc. AutoCarto, 2016, Albuquerque, NM, USA, 14–16 September 2016, 2016, pp. 100–110

[23] Chiang, Y.-Y., Leyk, S.: 'Exploiting online gazetteer for fully automatic extraction of cartographic symbols'. Proc. 27th Int. Cartographic Conf. (ICC), Rio de Janeiro, Brazil, 2015

[24] Chiang, Y.-Y., Leyk, S., Nazari, N.H., *et al.*: 'Assessing impact of graphical quality on automatic text recognition in digital maps', *Comput. Geosci.*, 2016, **93**, pp. 21–35

[25] Yu, R., Luo, Z., Chiang, Y.-Y.: 'Recognizing text in historical maps using maps from multiple time periods'. Proc. IEEE 23rd Int. Conf. on Pattern Recognition (ICPR), Cancun, Mexico, 2016, pp. 3993–3998
[26] Fishburn, K.A., Davis, L.R., Allord, G.J.: 'Scanning and georeferencing historical USGS quadrangles', *U.S. Geol. Surv. fact sheet*, 2017, **3048**, pp. 1–2, https://doi.org/10.3133/fs20173048
[27] Library of Congress: 'Geography and map division'. 2018. Available at https://www.loc.gov/collections/sanborn-maps/, accessed 8 January 2018
[28] Uhl, J. H., Leyk, S., Chiang, Y.-Y*., et al.*: 'Extracting human settlement footprint from historical topographic map series using context-based machine learning'. Proc. Eighth Int. Conf. on Pattern Recognition Systems (ICPRS 2017), Madrid, Spain, 2017, pp. 15–21
[29] Zillow Transaction and Assessment Dataset (ZTRAX): 'Available through a data use agreement between the university of Colorado Boulder and Zillow Inc.', 2016
[30] Lowe, D.G.: 'Object recognition from local scale-invariant features'. Proc. Seventh IEEE Int. Conf. on Computer Vision, Kerkyra, Greece, 1999, pp. 1150–1157
[31] Maaten, L. V. D., Hinton, G.: 'Visualizing data using t-SNE', *J. Mach. Learn. Res.*, 2008, **9**, pp. 2579–2605
[32] LeCun, Y., Boser, B., Denker, J.S*., et al.*: 'Backpropagation applied to handwritten zip code recognition', *Neural Comput.*, 1989, **1**, (4), pp. 541–551
[33] Chatfield, K., Simonyan, K., Vedaldi, A*., et al.*: 'Return of the devil in the details: delving deep into convolutional nets'. British Machine Vision Conf., Nottingham, UK, 2014, arXiv ref. cs1405.3531
[34] Cohen, J.: 'A coefficient of agreement for nominal scales', *Educ. Psychol. Meas.*, 1960, **20**, pp. 37–46
[35] Forbes, A.D.: 'Classification algorithm evaluation: five performance measures based on confusion matrices', *J. Clin. Monit. Comput.*, 1995, **11**, pp. 189–206
[36] Rosenfield, G., Melley, M.: 'Applications of statistics to thematic mapping', *Photogramm. Eng. Remote Sens.*, 1980, **46**, pp. 1287–1294
[37] Fawcett, T.: 'An introduction to ROC analysis', *Pattern Recognit. Lett.*, 2005, **27**, (8), pp. 861–874
[38] Leyk, S., Zimmermann, N.E.: 'A predictive uncertainty model for field-based survey maps using generalized linear models', in Egenhofer M., Freksa C., Miller H. (Eds.): Proc. Third Int. Conf. on Geographic Information Science (GIScience, 2004), Adelphi, MD, USA, 20–23 October 2004, (LNCS, 3234), pp. 191–205