IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2019, accepted December 21, 2019, date of publication December 31, 2019,
date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963213

Automated Extraction of Human Settlement Patterns From Historical Topographic Map Series Using Weakly Supervised Convolutional Neural Networks

JOHANNES H. UHL[1,4], **STEFAN LEYK**[1,4], **YAO-YI CHIANG**[2], **WEIWEI DUAN**[2], **AND CRAIG A. KNOBLOCK**[2,3], **(Senior Member, IEEE)**

[1]Department of Geography, University of Colorado Boulder, Boulder, CO 80309, USA
[2]Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089, USA
[3]Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
[4]Institute of Behavioral Science, University of Colorado Boulder, Boulder, CO 80309, USA

Corresponding author: Johannes H. Uhl (johannes.uhl@colorado.edu)

This work was supported in part by the National Science Foundation through the University of Colorado Boulder under Grant IIS 1563933, in part by the National Science Foundation through the University of Southern California under Grant IIS 1564164.

ABSTRACT Information extraction from historical maps represents a persistent challenge due to inferior graphical quality and the large data volume of digital map archives, which can hold thousands of digitized map sheets. Traditional map processing techniques typically rely on manually collected templates of the symbol of interest, and thus are not suitable for large-scale information extraction. In order to digitally preserve such large amounts of valuable retrospective geographic information, high levels of automation are required. Herein, we propose an automated machine-learning based framework to extract human settlement symbols, such as buildings and urban areas from historical topographic maps in the absence of training data, employing contemporary geospatial data as ancillary data to guide the collection of training samples. These samples are then used to train a convolutional neural network for semantic image segmentation, allowing for the extraction of human settlement patterns in an analysis-ready geospatial vector data format. We test our method on United States Geological Survey historical topographic maps published between 1893 and 1954. The results are promising, indicating high degrees of completeness in the extracted settlement features (i.e., recall of up to 0.96, F-measure of up to 0.79) and will guide the next steps to provide a fully automated operational approach for large-scale geographic feature extraction from a variety of historical map series. Moreover, the proposed framework provides a robust approach for the recognition of objects which are small in size, generalizable to many kinds of visual documents.

INDEX TERMS Convolutional neural networks, digital humanities, digital preservation, document analysis, geospatial analysis, geospatial artificial intelligence, human settlement patterns, image analysis, weakly supervised learning.

I. INTRODUCTION

Historical maps constitute unique sources of retrospective geographic information. Recently, several archives containing historical map series covering large spatial and temporal extents have been systematically scanned and made available to the public (e.g., [1]–[4]). The spatial-temporal information contained in such archives represents valuable information for a myriad of scientific applications [5]. To provide this geographic information in analysis-ready geospatial data formats, it needs to be unlocked from scanned maps using adequate recognition and extraction techniques that can handle very large volumes and varieties of complex data and provide a high degree of automation. Thus, traditional topographic map processing techniques based on manually created templates of the cartographic symbols of interest cannot be applied for information extraction from such large archives, holding map content of high levels of heterogeneity.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. F. Abate.

The renaissance of convolutional neural networks (CNNs) and the increasing use of other machine learning methods for recognition tasks in computer vision have catalyzed the application of such frameworks for information extraction tasks in the geospatial sciences [6]. In recent years, CNN-based approaches for object detection, scene classification, and semantic segmentation have been applied to remotely sensed geospatial data and have shown promising results that outperform traditional methods [7]. This indicates the potential of applying CNN-based semantic segmentation methods for information extraction from historical maps as well. However, encoder-decoder CNNs (e.g., [8]–[11]), which perform semantic segmentation at a fine level of spatial granularity (i.e., at the pixel level), require large amounts of pixel-level training data. While in remote sensing applications such pixel-level training labels can be generated efficiently, using ancillary spatial data such as land cover data [12], in the case of historical maps, it is more difficult. Often, spatial offsets between geographic features of interest in the map and the ancillary data are caused by inaccurate georeferencing, map distortions, map design or scale-induced displacements and impede the straight-forward generation of pixel-level training labels by overlaying the georeferenced map with ancillary data directly using their geocoordinates and projection metadata. In such cases, weakly supervised learning can be applied. Weak supervision refers to supervised learning when the granularity of annotations used for training is coarser than the granularity of the predicted annotations [13]. This is typically the case when pixel-level semantic segmentation is desired, but the location of the feature[1] of interest in the training data can only be determined approximately, e.g., within a certain spatial range.

In image processing, weakly supervised learning, for example, consists of the training of a CNN for image classification, i.e., learning image patch level annotations, indicating the presence of the object of interest somewhere within an image patch, and subsequent spatially dense inferences based on sliding windows, allowing for pixel-level labelling, typically of the center pixel in each sliding window. The result is a semantically segmented output image (see [14]–[16] for some examples). Thus, weakly supervised learning potentially allows for pixel-level semantic segmentation, in cases when the location of salient features for training can only be determined approximately, e.g., within a subset or patch of the image. However, such weakly supervised segmentation approaches may result in a loss of spatial detail due to heterogeneous image content and the translation invariance property of CNNs, as previous work has demonstrated [17], [18].

Thus, in order to successfully apply a CNN-based semantic segmentation framework for information extraction from historical maps, such a framework needs to ***1)*** reliably and automatically generate sufficiently large amounts of training data and labels for the recognition of geographic features from historical maps, and ***2)*** allow for the extraction of geographic features at sufficiently fine spatial granularity.

In this paper, we present and evaluate an improved approach for the extraction of human settlement features (i.e., building locations and urban area delineations), using publicly available (i.e., down-sampled), early United States Geological Survey (USGS) historical topographic maps published between 1893 and 1949. The proposed method aims to provide a framework for automated, weakly supervised CNN-based feature extraction from historical maps at fine spatial granularity without the availability of pixel-level training labels. To overcome the absence of training data, the proposed framework employs ancillary spatial data to automatically collect training data. Locational settlement information given in the ancillary data enables to spatially constrain the regions in which objects of interest can be found and allows for automated sampling of a map underlying the ancillary locations by cropping the map image at those locations.

However, potential spatial and temporal offsets between map content and ancillary spatial data may result in collected samples being ***a)*** not centered at the object of interest (e.g., at a building symbol), or ***b)*** being annotated with an erroneous label (e.g., labelled as ''building'' where the map does not contain a building symbol). These effects are mitigated by a hierarchical, spatially-stratified random sampling scheme in combination with image processing techniques and feature detection and description methods to obtain reliable training samples centered at the object of interest.

The generated training data consist of small samples of each training map (''map patches'') found in proximity of the ancillary locations, cropped around the objects of interest, and corresponding labels describing their content according to a given classification scheme. We then employ these data to train CNNs commonly used for image classification, which we use for subsequent semantic segmentation in a weakly supervised manner. The previous centering of the training patches to the likely location of the objects of interest reduces the loss of spatial granularity when these weakly supervised CNNs are employed for dense, pixel-level inferences to extract human settlement features at fine spatial resolution. Moreover, the proposed method represents a generalizable strategy for the recognition of small objects from visual documents in general.

II. BACKGROUND

A. MAP PROCESSING

Map processing, a branch of document analysis, focuses on developing methods for the extraction and recognition of information in scanned map documents such as printed engineering drawings, floor plans, cadastral and topographic maps published prior to the era of digital cartography and systematic earth observation. It combines elements of computer vision, pattern recognition, geographic information science, cartography, and geoinformatics. The main goal of map processing is to unlock spatial information from those

[1] Herein, the term ''feature'' is used for geographic features, i.e., the spatial objects depicted in topographic maps.

(mainly historical) scanned map documents, to provide this information in digital, machine-readable data formats and thus to preserve the data digitally, facilitating their use for analytical purposes [5]. The application of recognition methods to map documents often faces specific challenges compared to traditional document analysis due to low graphical quality and complex, human-made map content (e.g., overlapping cartographic symbols) [19]. Example applications of map processing include the extraction of buildings [20]–[22], residential areas [23], road networks [24] contour lines [25], [26], composite forest symbols [27], text [28] as well as the digitization of cadastral maps [29]. Successful map processing requires georeferencing ([30]–[35]) and the alignment of georeferenced maps and ancillary spatial data ([36], [37], see [5], [38] for detailed overviews). Three recent developments are currently changing the field of map processing: ***1)*** An increasing availability of large amounts of scanned, often georeferenced historical maps [39], ***2)*** advances in computer-vision based information extraction using (deep) machine learning [40], and ***3)*** increasing availability of digital geospatial data [41] that can be used as ancillary data to support symbol sample collection.

B. DIGITAL HISTORICAL MAP ARCHIVES

There is an increased availability of large map collections holding thousands of map documents as digital and georeferenced archives hosted by map agencies including the USGS topographic map archive, holding approximately 200,000 topographic maps published between 1884 and 2006 [1], the Sanborn fire insurance map collection which contains approximately 700,000 sheets of large-scale maps of approximately 12,000 cities and towns within the U.S., Canada, Mexico, and Cuba published since 1867 [2], the United Kingdom topographic map archive (>200, 000 maps, dating back to the 1840s [3], or the historical map archive of Switzerland (approximately 52,000 maps dating back to 1844 [4]. Moreover, several digital map collections[2345] and data infrastructure efforts [42] have been established. Given this vast amount of valuable historical information, there is an urgent demand to preserve map contents through efficient information extraction while reducing or eliminating user interaction.

C. CURRENT TRENDS IN HISTORICAL MAP PROCESSING

Deep-learning-based models such as convolutional neural networks (CNNs) have revolutionized many scientific fields and offer great potential for numerous applications in the geospatial sector, such as map processing [43]. Such efforts include the use of deep learning and data mining for automated map georeferencing [44], [45], for text recognition [46], for the extraction of road intersections [47], or for map archive content exploration [48], [49]. Furthermore, training and benchmark datasets tailored to information extraction from scanned maps or plans have increasingly been made available to the public [49]–[51]. Several contributions propose the use of advanced machine learning for information extraction from map documents including topographic maps [52]–[54], cadastral maps [55], [56] or floor plans [57]. The need for large amounts of training data can be overcome by using crowdsourcing [58] or by employing (contemporary) ancillary spatial data [59]–[64]. For example, the use of building footprint data, cadastral parcel data and other settlement-related geospatial databases has proven useful for the extraction of human settlement features from historical topographic maps [17], [18].

III. DATA

The experiment in this study is based on USGS topographic maps, publicly available through the USGS TopoView web application[6], at a spatial resolution of approximately 5 × 5m, downsampled from original scans by an approximate factor of 5. More specifically, we choose 18 map sheets of scale 1:62,500, organized in six adjacent map quadrangles covering Greater Albany (NY), and three epochs (1893 - 1903, 1927 - 1938, and 1949 - 1954) including the earliest editions of available USGS topographic maps. Since the Albany region is characterized by relatively early settlements, a high proportion of built-up area can be expected and thus, provides a suitable study area to test extraction methods for early cartographic products. These 18 maps are shown in Figure 1, including respective enlargements for the town of Mechanicville (NY).

The ancillary spatial data used for training data collection are settlement locations (i.e., approximate centroids of built-up cadastral parcels) derived from the ZTRAX (Zillow Transaction and Assessment Dataset) data [65], containing approximately 230,000 settlement locations in the study area in 2016. Moreover, we use a metadata file from USGS[7] to extract quadrangle boundary coordinates for each map sheet and perform subsequent map sheet edge removal by clipping the georeferenced maps to the respective quadrangle boundaries. Moreover, we employ a set of building and urban area outlines manually digitized from selected map sheets (i.e., a total number of approximately 4,700 building outlines and urban area delineations) to validate the semantic segmentation results.

IV. METHOD

The method proposed in this study consists of the following stages: ***a)*** spatial data preprocessing and automated training data collection at map patch level for building symbols, urban areas, and non-settlement classes using a hierarchical, spatially stratified random sampling scheme, and ***b)*** CNN training and semantic segmentation using weakly supervised

[2] David Rumsey Map Collection: https://www.davidrumsey.com

[3] Mapire: https://mapire.eu

[4] Old maps online: https://www.oldmapsonline.org

[5] Pahar - the Mountains of Central Asia Digital Dataset: http://pahar.in

[6] https://ngmdb.usgs.gov/topoview/viewer

[7] https://thor-f5.er.usgs.gov/ngtoc/metadata/misc

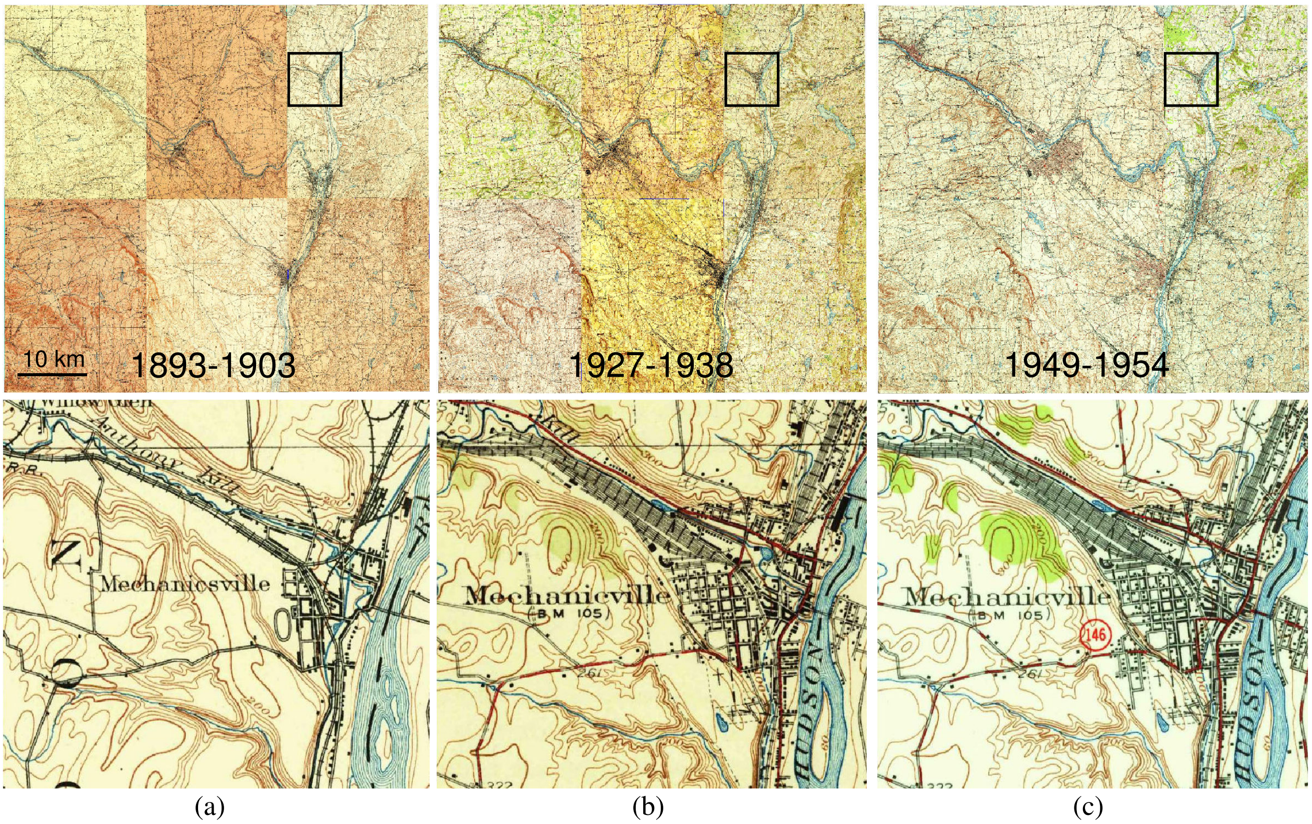


FIGURE 1. Study area for training data creation: Six USGS 1:62,500 map quadrangles covering the area of Greater Albany (NY, USA) from three time periods: (a) 1893-1903, (b) 1927-1938, and (c) 1949-1954, with respective enlargements for the town of Mechanicville (black rectangles) in the bottom row.

CNNs. The latter stage is motivated by the absence of a-priori knowledge about the precise location of building symbols, impeding the automated generation of pixel-level training labels. Thus, we train CNNs on map patches for image classification and conduct pixel-level semantic segmentation using the trained CNNs in a weakly supervised manner. Lastly, segmentation results are vectorized. Figure 2 illustrates the proposed framework.

A. SPATIAL DATA PREPROCESSING AND AUTOMATED TRAINING DATA COLLECTION

In USGS topographic maps, human settlement is depicted as black polygons for individual building symbols (Figure 3a), or as dotted areas for dense urban settlements (Figure 3b). Herein, the classification problem consists of separating these two settlement classes from non-settlement map content. As previous work has shown, a main challenge is the separation of individual buildings from other black map content, such as text (Figure 3c), whereas the separation from remaining (i.e., non-black) non-settlement related map content (Figure 3d for an example) is expected to be more straightforward. In our classification problem, black and other non-settlement content are considered two separate negative classes. This separation of non-settlement content facilitates the discrimination between building symbols and other black map content (e.g., text elements), since classification tasks are generally easier to solve if intra-class data variability is low [18]. The examples in Figure 3 illustrate that the recognition of buildings and urban area is expected to be challenging, since, in the case of buildings, the salient features are small in size (approximately 5-10 pixels) compared to other map content, and, in the case of urban areas, the salient features (i.e., composite features consisting of regularly spread red dots) are considered poorly-defined background information. In order to extract large amounts of such training samples automatically, we developed a workflow making use of several commonly used image processing techniques. This workflow uses contemporary settlement locations derived from the ZTRAX ancillary data, assuming a certain degree of coherence between these settlement locations and settlement symbols in the underlying maps, resulting in sufficient spatial proximity between the settlement features from the two datasets.

However, positional and temporal offsets and discrepancies between the two datasets need to be taken into account. These include:

- positional inaccuracy inherent in the topographic map, due to paper distortions, scanner miscalibrations, or cartographic displacements,
- positional inaccuracy introduced during georeferencing of the scanned map,
- positional discrepancies due to the nature of the ancillary settlement locations, representing approximate cadastral

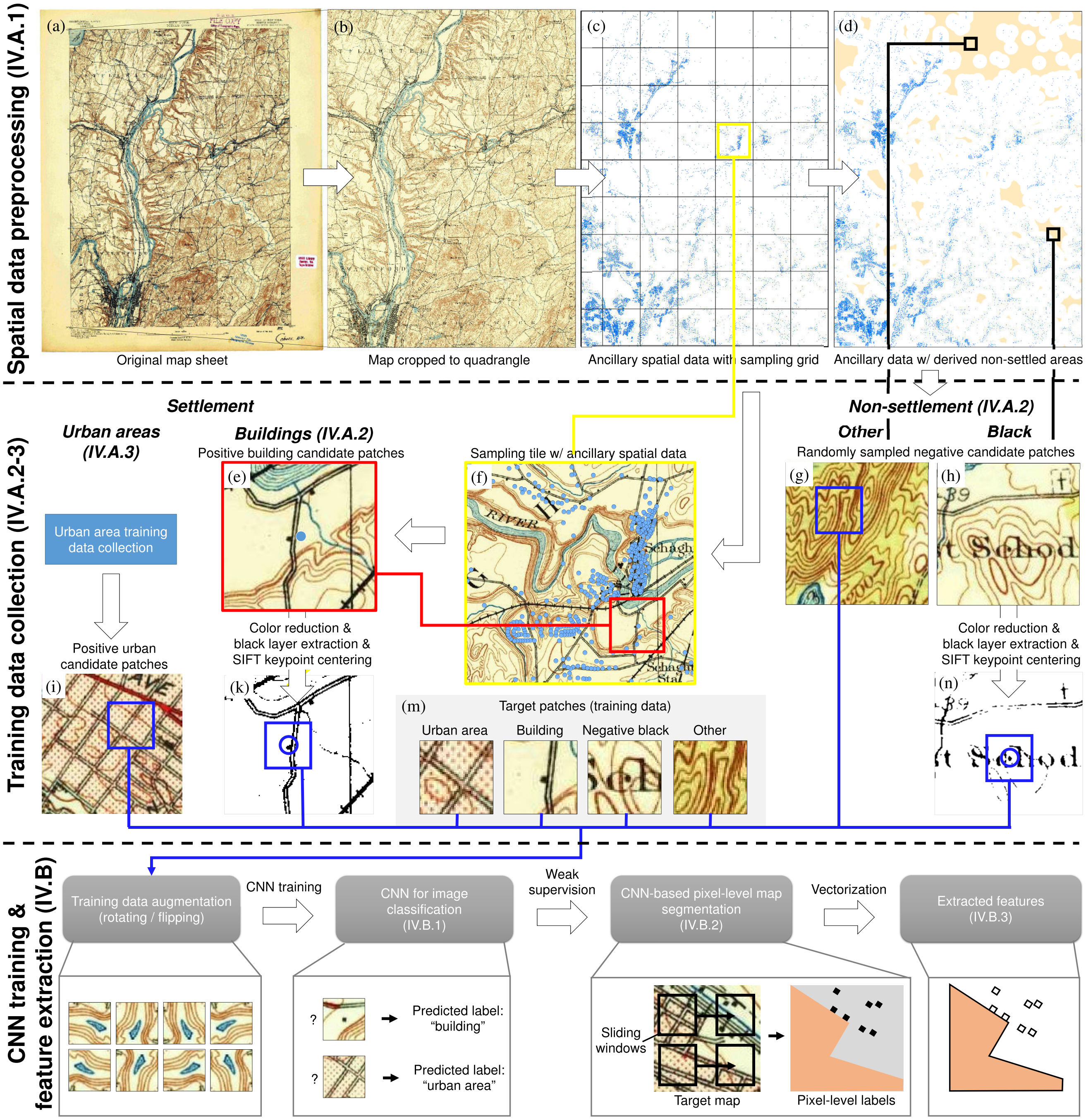

Spatial data preprocessing (IV.A.1)
(a)
Original map sheet
(b)
Map cropped to quadrangle
(c)
Ancillary spatial data with sampling grid
(d)
Ancillary data w/ derived non-settled areas
Training data collection (IV.A.2-3)
Settlement
Urban areas (IV.A.3)
Buildings (IV.A.2)
Positive building candidate patches
(e)
Sampling tile w/ ancillary spatial data
(f)
Non-settlement (IV.A.2)
Other
Black
Randomly sampled negative candidate patches
(g)
(h)
Urban area training data collection
Positive urban candidate patches
(i)
Color reduction & black layer extraction & SIFT keypoint centering
(k)
(m)
Target patches (training data)
Urban area
Building
Negative black
Other
Color reduction & black layer extraction & SIFT keypoint centering
(n)
CNN training & feature extraction (IV.B)
Training data augmentation (rotating / flipping)
CNN training
CNN for image classification (IV.B.1)
Weak supervision
CNN-based pixel-level map segmentation (IV.B.2)
Vectorization
Extracted features (IV.B.3)
?
Predicted label: "building"
?
Predicted label: "urban area"
Sliding windows
Target map
Pixel-level labels

FIGURE 2. **The proposed framework for settlement feature extraction from historical topographic maps. The upper part shows the spatial data preprocessing, the center part illustrates the hierarchical, spatially stratified random sampling scheme for training data collection and label assignment, and the lower part shows the CNN training and feature extraction stage: (a) Original Cohoes (NY) map at scale 1:62,500 from 1893, (b) map sheet after removing the map collar by clipping to the map quadrangle boundaries, (c) the ancillary spatial data (i.e., ZTRAX settlement locations) in blue and 3 × 3km tiles for spatially stratified random sampling, (d) ancillary settlement locations (blue) and derived non-settlement areas (light red), (e) a positive building candidate patch extracted from the sampling tile shown in (f), (g) randomly sampled negative candidate patches for non-black, and (h) for black non-settlement content, (i) a candidate patch containing urban area, (k) the black layer extracted from the patch shown in (e) including an identified SIFT keypoint and the target patch centered around it, (n) corresponding visualization of the negative black candidate patch shown in (h), and (m) examples of target patches (48 × 48 pixels) for the four classes used as CNN training data. Note that the method for urban area training data collection (blue box in the center left) is illustrated separately in Figures 6 and 7. Section numbers corresponding to each methodical step are shown in parenthesis.**

parcel locations, partially integrated with address point data, and

- temporal offsets between contemporary ancillary settlement locations from the ZTRAX database and the map content, i.e. contemporary settlements that did not exist when the historical maps were created. The number of locations affected by such temporal offsets is assumed to increase towards early map editions.

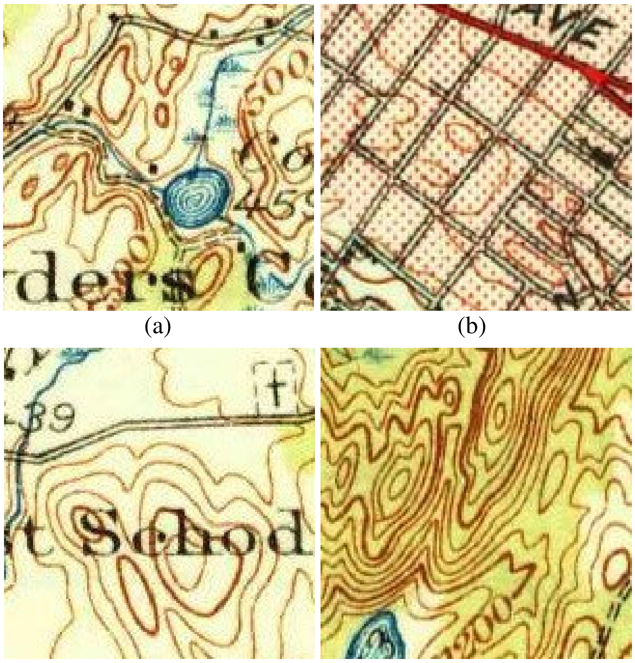


FIGURE 3. **Exemplary map patches of the four training classes: cropped patches (150 × 150 pixels) from a USGS map at scale 1:62,500 for (a) building symbols, (b) urban area, (c) black non-settlement map content, and (d) other non-settlement map content.**

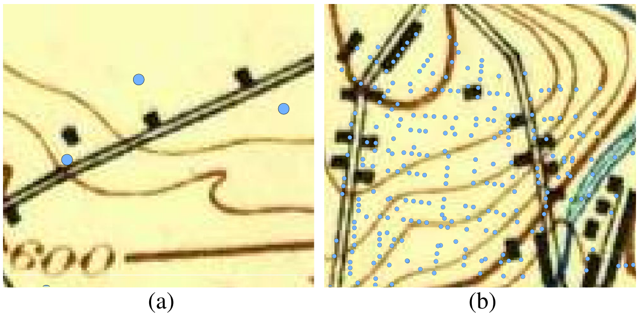


FIGURE 4. **Examples of (a) spatial offset, and (b) temporal offset between building locations in historical maps and contemporary settlement locations (blue dots) from the ZTRAX ancillary spatial data.**

Figures 4a and 4b show examples of spatial and temporal offsets between map and ancillary spatial data. Due to these discrepancies a sample of the map cropped at a contemporary settlement location may contain a building, may contain urban area, or may contain both, or neither of them. The proposed automated training data collection procedure accounts for these uncertainties using a variety of image processing techniques to keep levels of label noise (i.e., mislabeled training samples) to a minimum. The training data collection procedure is done after a spatial data preprocessing step, conducted for each training map individually, and consists of ***a)*** building and non-settlement training data collection, and, separately, ***b)*** urban area training data collection.

1) SPATIAL DATA PREPROCESSING

Firstly, the collar of each map sheet used for training data collection is removed. This process is automated based on the previously mentioned quadrangle extent data file, and is realized by clipping the georeferenced map to a generated quadrangle polygon (Figure 2a,b). Then, ancillary settlement locations from the ZTRAX database are retrieved for the map quadrangle extent and a regular sampling grid is generated for each input map to partition the map into spatial bins (i.e., tiles) of $3 \times 3km$. These ancillary data and the sampling grid are shown in Figure 2c. The sampling grid is used for spatially stratified random sampling, as described in the following section. Moreover, the ancillary settlement locations are employed to identify likely non-settled sample areas. To do so, all areas within the map quadrangle farther away than 500m from any ancillary settlement location are selected using a spatial buffering operation, which helps avoiding mislabeling in boundary regions between settled and non-settled areas (Figure 2d).

2) COLLECTING BUILDING AND NON-SETTLEMENT TRAINING DATA

To account for potential discrepancies (i.e., spatial and temporal offsets between ancillary data and map content), the training data collection step is designed as a hierarchical process, involving spatial units of three levels of granularity, which we call: ***sampling tiles***, ***candidate patches***, and ***target patches***. Within each sampling tile, a random subset (N = 500) of ancillary settlement locations is selected (Figure 2f, blue points). Such a spatially stratified random sampling scheme (i.e., random sampling within fixed spatial bins) makes it possible to obtain a sample of locations equally representative for all sub-regions covered by the map, accounting for density variations of settlement locations between urban and rural areas. This strategy mitigates the imbalance between urban and rural regions, and thus, increases the training data variability and minimizes the probability of generating duplicate or heavily overlapping samples. For each selected settlement location in each $3 \times 3km$ sampling tile (Figure 2f), the underlying map is cropped within 144 × 144 pixels (px) (approximately 750 × 750m) centered at the ancillary settlement location (i.e., candidate patch). The large patch size is chosen to capture building symbols even if spatial offsets between map and ancillary data exist. These intermediate patches are called candidate patches since they may not contain a building symbol, if the area was not settled during the map edition year, or if spatial offsets between building symbol and ancillary location are too large. In the exemplary candidate patch in Figure 2e, a building symbol is captured with a slight offset with respect to the ancillary settlement location. Whether a candidate patch contains a building or not, is determined using the following procedure:

First, a color-space segmentation of the candidate patches is carried out using k-means clustering with k = 5 in the RGB space. While historical USGS topographic maps were initially printed in three colors and later using five colors [66], the texture of the paper and the chosen bit depth during the scanning process artificially increase color complexity in the scanned document. Thus, color reduction decreases the complexity of the scanned image without losing important information. Since building symbols are expected to be black, the cluster with the centroid closest to (0,0,0) in RGB space (i.e., the "darkest" cluster) is considered the black layer, if $R < 80 \cap G < 80 \cap B < 80$. Subsequently, the black layer is tested for the presence of dark blobs, potentially representing building symbols. This is done by searching for maxima in the Difference-of-Gaussian scale space (i.e., Scale-Invariant Feature Transform - SIFT keypoint detection, [67]). It has been shown that SIFT keypoints reliably detect dark blobs such as building symbols [17], [18]. If multiple keypoints are detected, only one of them is randomly selected and retained. If such a keypoint is found, a small patch (48 × 48 pixels, approximately 250 × 250m) is cropped, centered around the keypoint location (see Figure 2k,m). This is the target patch, and its training label is "building".

To create non-settlement training samples, the map is cropped at random locations within the non-settled areas (Figure 2d). These map patches are called negative sample candidates (Figure 2g,h). To determine whether these patches contain black or other non-settlement related map content, the patches are processed in analogy to the building candidate patches (i.e., color reduction, black layer extraction, SIFT keypoint detection). If no keypoint is found, a target patch is cropped at a random location within the candidate patch, and the target patch is labelled "other non-settlement map content" (Figure 2g,m), otherwise, if a keypoint is found on the black layer, a target patch is cropped at the keypoint location, and the target patch is annotated "negative black" (Figure 2h,m,n). In latter case, the eight patches adjacent to the respective candidate patch are examined for salient negative black content as well, assuming that larger text elements typically extend over multiple candidate patches. Concluding, this unsupervised procedure yields annotated training data of building symbols and the two negative classes, with buildings and negative black map content geometrically centered in the map patches.

3) COLLECTING URBAN AREA TRAINING DATA

Starting from approximately 1950, the USGS depicts densely built-up urban areas using a uniform texture-based signature rather than individual residential buildings or building blocks [68] (see Figure 5 for some examples from different time periods and map scales).

The textural characteristics of map patches containing urban area are expected to differ significantly from non-urban areas (cf. Figure 3). Thus, in order to identify samples representing urban areas from the pool of positive candidate patches (i.e., all candidate patches potentially containing either building symbols or urban area), we apply an unsupervised texture-based classification method to the positive candidate patches, involving texture descriptors based on the local binary patterns (LBP) method [69]. LBP makes use of differences in grey values between a center pixel and its neighbors within a convolving structural element. Signs of the differences to the neighboring pixels are encoded as 0 and 1 and are used to form the binary representation of a number, which is assigned to the center pixel, to generate an LBP surface (see Figure 6, middle row for some examples). It is common practice to use the histogram of the LBP surface as a texture descriptor (Figure 6 bottom row). Thus, the LBP histogram is expected to hold high degrees of discriminatory power with respect to the urban symbols, as can be seen in the histogram of the urban area example (Figure 6a) showing higher frequencies for low LBP values than the other examples.

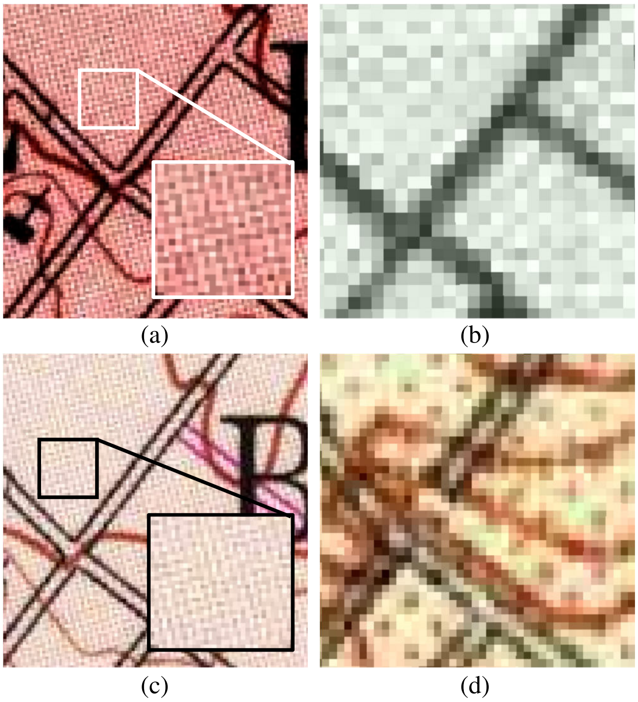


FIGURE 5. Examples of urban area signatures used in USGS topographic maps. Maps from Albany (NY), (a) year 2000, scale 1:24,000, (b) year 1986, scale 1:100,000, (c) year 1981, scale 1:24,000, and (d) year 1950, scale 1:62,500. Examples at scale 1:24,000 are enlarged for better visibility.

Figure 7 illustrates the method to extract urban area training samples. For each candidate patch cropped around ancillary settlement locations (blue dots in Figure 7a) and converted to a grayscale image, the LBP histogram is computed (number of histogram bins = 10). This allows to locate each candidate patch in a 10-dimensional LBP space, visualized in two dimensions for a subsample of candidate patches in Figure 7b using t-distributed stochastic neighbor embedding (t-SNE, [70]). T-SNE is a non-linear dimensionality reduction method, allowing for mapping and visualizing high-dimensional data in a low-dimensional space, while the proximity between nearby data points in the mapped space reflects their similarity in the original space. To separate patches containing urban area from non-urban area, the positive candidate patches are grouped into clusters using k-means clustering in the LBP space. We choose k = 3 to account for three textural types occurring in the maps, or for

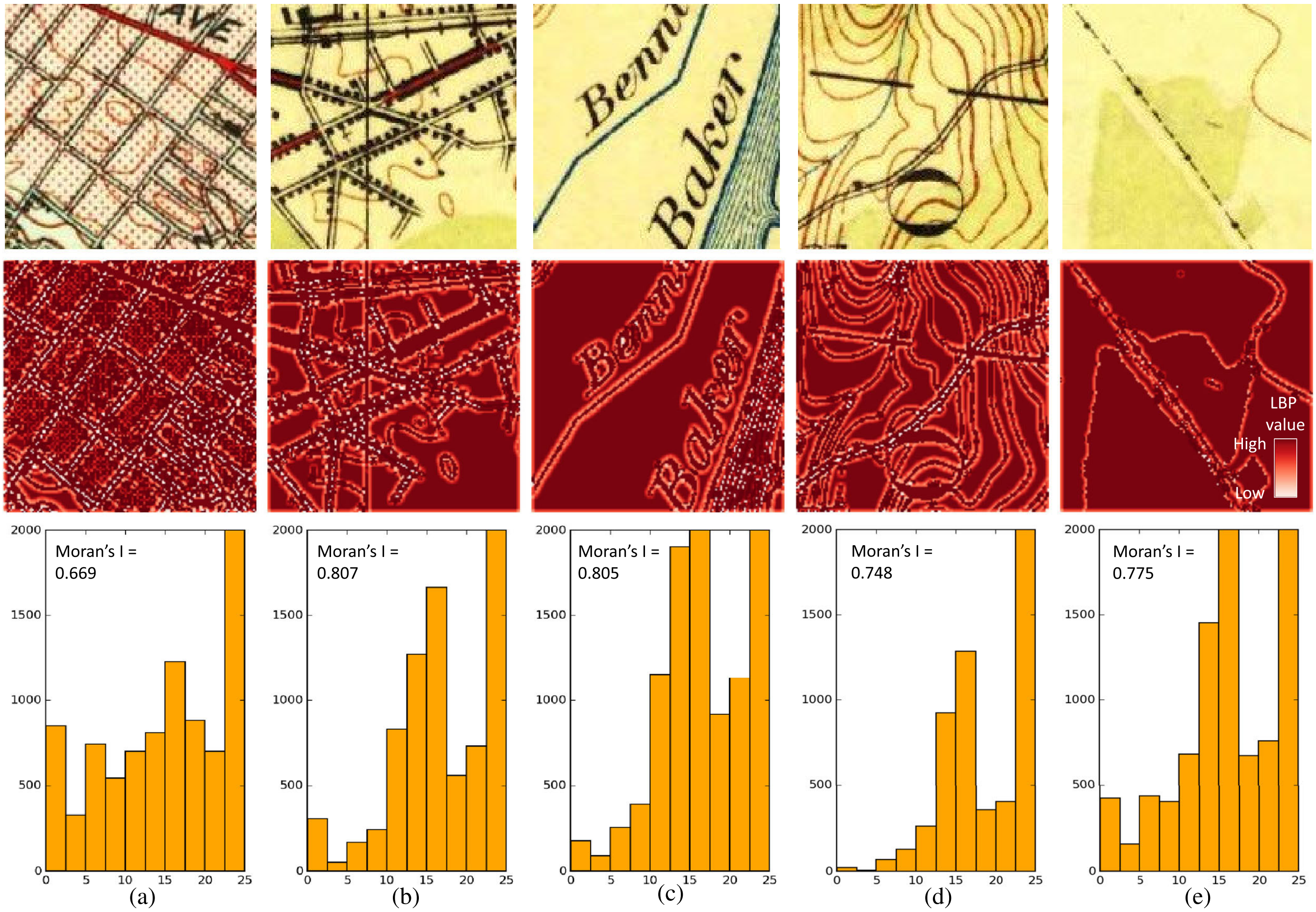


FIGURE 6. Examples of local binary patterns (LBP) histogram descriptors for texture classification used to identify urban map signatures in an unsupervised manner: (a) patch containing urban area, corresponding LBP surface, LBP histogram with corresponding Moran's I global spatial autocorrelation measure indicated (from top to bottom), and
(b) - (e) examples of non-urban map content.

edge effects caused by patches of mixed textural characteristics sampled at the edges of urban areas. In order to identify the cluster of patches likely to contain urban areas, we compute an average spatial autocorrelation measure across the patches assigned to each cluster.

Spatial autocorrelation is a fundamental concept in geospatial analysis to quantify the similarity between nearby objects in geographic space [71]. While nearby objects exhibiting high levels of similarity are considered positively spatially autocorrelated, a regularly spread grid of dots representing urban areas (Figure 5) is expected to yield negative spatial autocorrelation or at least low degrees of positive spatial autocorrelation, i.e., neighboring pixels have statistically significant differences in gray values. Thus, the average Moran's I (a common spatial autocorrelation measure, [72]) is computed over all patches per cluster, and the cluster of lowest average Moran's I (i.e., the cluster of lowest spatial autocorrelation) is assumed to contain the urban areas, as demonstrated by the Moran's I values reported for the examples in Figure 6 and shown in Figure 7d. Based on this method, we identify candidate patches that are likely to contain urban area (i.e., the patches contained in the cluster of lowest average spatial autocorrelation, cluster 3 in Figure 7c). This method is integrated into the main training data collection process, as illustrated in Figure 2i, showing an exemplary urban candidate patch. Within those urban candidate patches, the urban target patches are cropped at random locations, labelled as "urban area" and added to the pool of training samples (Figure 2i,m).

B. CNN TRAINING, SEMANTIC SEGMENTATION AND FEATURE EXTRACTION

We conduct training data augmentation by rotating and flipping the generated training images in all possible and meaningful combinations of rotations and directions and adding these variants to the pool of training data in order to increase the number of training samples and their variability which is expected to improve the generalizability of the trained CNNs. Underrepresented classes (i.e., classes which are represented by fewer training samples than the class with the highest number of training samples) are randomly oversampled to achieve balanced training samples.

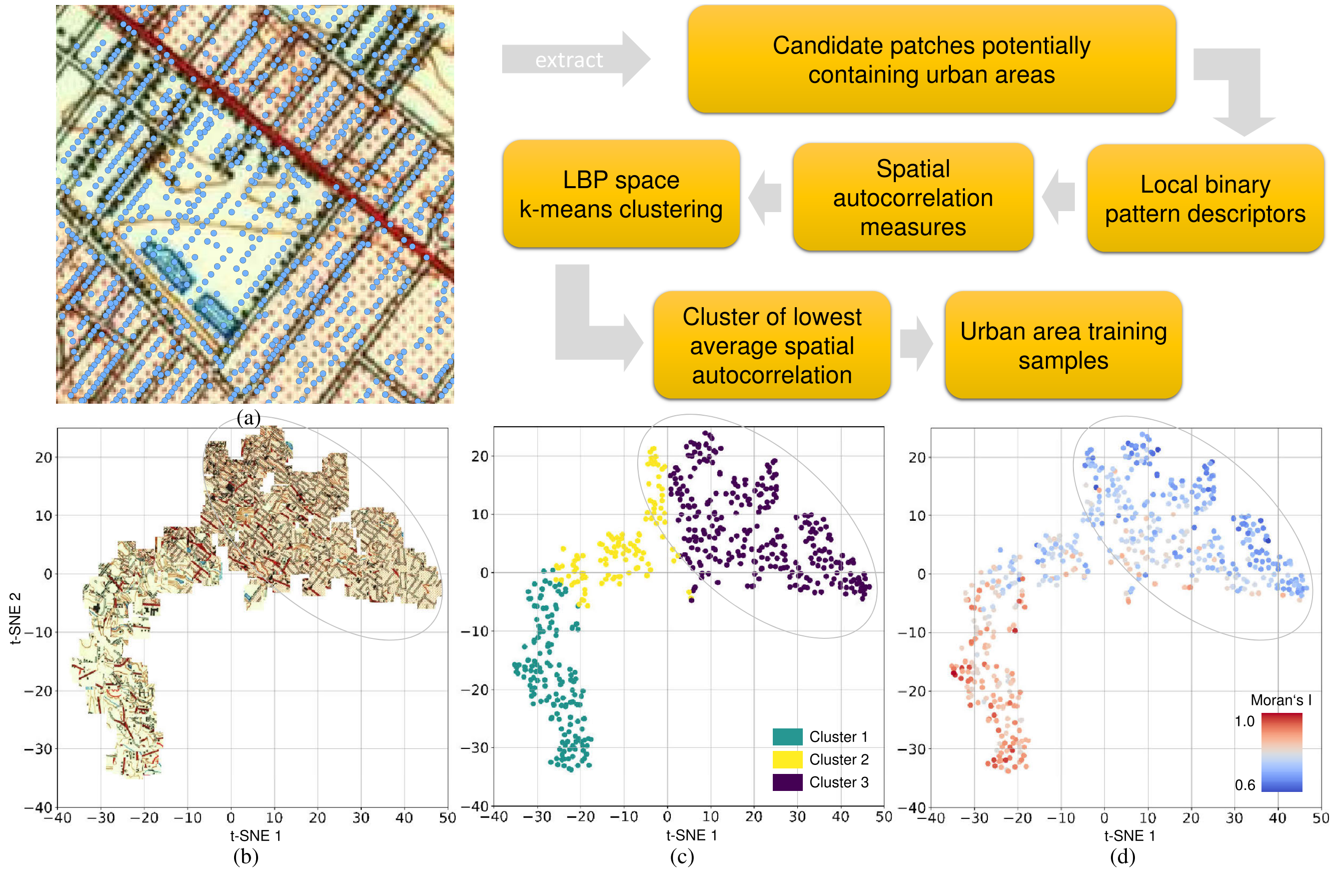

extract
Candidate patches potentially containing urban areas
Local binary pattern descriptors
Spatial autocorrelation measures
LBP space k-means clustering
Cluster of lowest average spatial autocorrelation
Urban area training samples
(a)
t-SNE 2
t-SNE 1
20
10
0
−10
−20
−30
−40
−30
−20
−10
0
10
20
30
40
50
(b)
Cluster 1
Cluster 2
Cluster 3
(c)
Moran's I
1.0
0.6
(d)

FIGURE 7. The automated workflow to collect urban area training data. (a) Example map overlaid with ancillary settlement locations, (b) illustration of extracted candidate patches potentially containing urban areas, (c) k-means clusters of the candidate patches in LBP space, and (d) corresponding Moran's I spatial autocorrelation measure for each patch, Figures (b)-(d) are transformed from the original 10-dimensional LBP space into a two-dimensional space using t-SNE. The grey ellipse in (b)-(d) indicates the cluster of lowest average spatial autocorrelation and the respective urban area training samples.

1) CNN TRAINING FOR MAP PATCH CLASSIFICATION

The automatically generated pool of training data for the four classes (i.e., urban areas, buildings, black and other non-settlement content) consists of the extracted target patches and their associated patch-level annotations. Based on these training data, we train CNNs commonly used for image classification (see Section V for details).

2) SEMANTIC SEGMENTATION

We use a CNN trained for image classification to perform pixel-level semantic segmentation in a weakly supervised manner, since the spatial and temporal offsets between map symbols and ancillary spatial data impede the generation of pixel-level training data and thus, hinder an end-to-end feature extraction using fully connected neural networks. Instead, we perform semantic segmentation by cropping patches of the map to be segmented around each pixel using a sliding window, predicting the label of the patch using the trained CNN, and assigning the predicted label to the center pixel of the sliding window (see Figure 2, bottom row). In this step, the labels of the two negative classes (i.e., black and other non-settlement content) are merged. Moreover, the georeference information of the input map is automatically transferred to the segmentation output raster dataset in order to generate a georeferenced segmentation result.

3) FEATURE EXTRACTION

Lastly, the resulting raster dataset containing the segmentation result is vectorized, i.e., generating vector-based polygon objects for contiguous groups of pixels of the building class, and the urban area class, respectively. The extracted polygon objects can then be used in an analytical environment, such as in a Geographic Information System (GIS).

V. EXPERIMENT AND RESULTS

From the multi-temporal set of historical USGS topographic maps from 1893 to 1954 for Greater Albany (NY), USA (see Figure 1) we automatically generate training data at two different target patch sizes i.e. 48 × 48 pixels (approximately250 × 250m) and 96 × 96 pixels (approximately 500 × 500m). We choose these two patch sizes to evaluate the CNN classification performance for different levels of spatial context given in the training samples. Based on the automated training data collection procedure (Section IV.A), over 130,000 labelled images are generated of size 48 × 48px, and over 120,000 labelled images of size 96 × 96px, respectively, both at a map scale 1:62,500. We test and evaluate the proposed method in a variety of ways: ***1)*** We visually assess the automatically generated training data (see Section V.A). ***2)*** We design a set of image classification scenarios using several CNN architectures and input data dimensions, and evaluate them by holding back training data for testing (see Section V.B). ***3)*** We use large amounts of manually digitized building outlines and urban areas from a set of validation maps to be employed as external validation data (see Section V.C). For the two best-performing CNNs, we conduct semantic segmentation of the validation maps, as described in Section IV.B, and use Receiver-Operator-Characteristic (ROC, [73]) diagnostics to assess the suitability of these two CNNs for semantic segmentation of each validation map, and ***4)*** for the CNN yielding best Area-under-the-curve (AUC, [74]) values, we compute a variety of accuracy measures to evaluate the segmentation results across the validation maps using pixel-based and object-based map comparison techniques (see Section V.C). For this experiment, spatial data processing is done in Python, mainly using GDAL/OGR[8] and ESRI ArcPy[9] geoprocessing python packages. Training data collection is implemented using GDAL/OGR and OpenCV[10]. CNN training and subsequent inferences are implemented and conducted using Keras[11] on an AWS EC2[12] instance of type g2.8xlarge (4 NVIDIA GRID GPUs, 32 vCPUs, 60 GB of memory).

A. VISUAL TRAINING DATA INSPECTION

The proposed automated training data collection method is unsupervised, and potentially prone to mislabeled training samples due to the mentioned spatial and temporal offsets between maps and ancillary spatial data (Figure 4), as well as due to heterogeneous map content within an extracted candidate patch. Hence, we use a visual inspection method (see [48] for details), arranging the training patches in a two-dimensional space based on color moments derived from the color histogram [75] and t-SNE. These visualizations and respective enlargements for each class (Figure 8) indicate high levels of precision, i.e., small proportions of mislabeled training samples. As Figure 8a suggests, most samples labelled as building are of high quality, i.e., show a building object at the patch center, which is a result of the previously described SIFT keypoint centering. While some false positives (i.e., text elements labeled as buildings) occur, almost no building symbols are mislabeled as ''negative black''. However, the texture-based clustering employed to generate the urban class samples falsely labels some patches containing bathymetry lines as urban (Figure 8b). This could be an edge effect (i.e., rivers crossing urban cores), or due to the characteristics of dense bathymetry lines, yielding textural descriptors similar to the urban textural signature, and low levels of spatial autocorrelation.

B. CNN-BASED MAP PATCH CLASSIFICATION

Based on the generated training data (i.e., map patches and corresponding labels describing their content) we test and evaluate the performance of three commonly used CNN architectures, separately for the image classification task and the segmentation task. These three CNNs include the classical LeNet model with 2 convolutional, 2 pooling, and 2 fully connected layers that has shown good performance on simple image recognition tasks [76], AlexNet, consisting of 5 convolutional layers, 3 pooling layers, and 3 fully connected layers [77], and VGGNet-16, consisting of 13 convolutional, 5 pooling, and 3 fully connected layers [78]. To prevent overfitting in case of the deeper CNNs (i.e., AlexNet and VGGNet-16), we used dropout regularization with a rate of 40%. We include LeNet to test the performance of shallow CNNs for the given classification task. The chosen CNN architectures are illustrated in the Appendix. Besides evaluating the different CNN architectures against each other, we use two input sizes (48 × 48px and 96 × 96px), testing for effects based on different levels of content heterogeneity and spatial context. For the shallower CNNs (LeNet and AlexNet), the training data are provided ***a)*** as RGB color images and ***b)*** as grayscale images, to test for potential effects on classification accuracy. We do this since it can be assumed that the salient features allowing for discriminating between buildings, urban areas, and other map content, are mostly shape and texture based, and likely independent from color information. However, for VGGNet-16, only color input is used, as grayscale images are expected to provide insufficient support to train the 138 million parameters used in VGGNet-16. We train all CNNs from scratch, i.e., using a random weights initialization rather than using pre-trained weights. Table 1 summarizes the experiment configuration used in this paper.

For training and evaluation of the CNN-based patch classification, the input patches are split into 80% training patches and 20% test patches in order to evaluate the internal classification accuracies as shown in Table 2, revealing some interesting details for the different CNN configurations.

Generally, most configurations yield high classification accuracies (average F-measures >0.8). As expected, the shallow LeNet achieves lowest accuracy levels (F-measure = 0.72 for the large input size and using color images), indicating that the classification problem is too complex to be modelled by LeNet. Most CNNs yield higher classification accuracies for the configurations using large input sizes (i.e., 96 × 96px) compared to the small ones. While with increasing input patch size the heterogeneity in map content is expected to increase and thus, constitutes a more challenging classification task, it is likely that due to the chosen sampling scheme, which includes oversampling of underrepresented classes, the risk of overfitting may be higher, despite the dropout regularization used. Comparing the effects of using RGB input data versus greyscale data,

[8] https://pcjericks.github.io/py-gdalogr-cookbook

[9] https://pro.arcgis.com/en/pro-app/arcpy

[10] https://docs.opencv.org/master/d0/de3/tutorial_py_intro.html

[11] https://keras.io

[12] https://aws.amazon.com/de/ec2

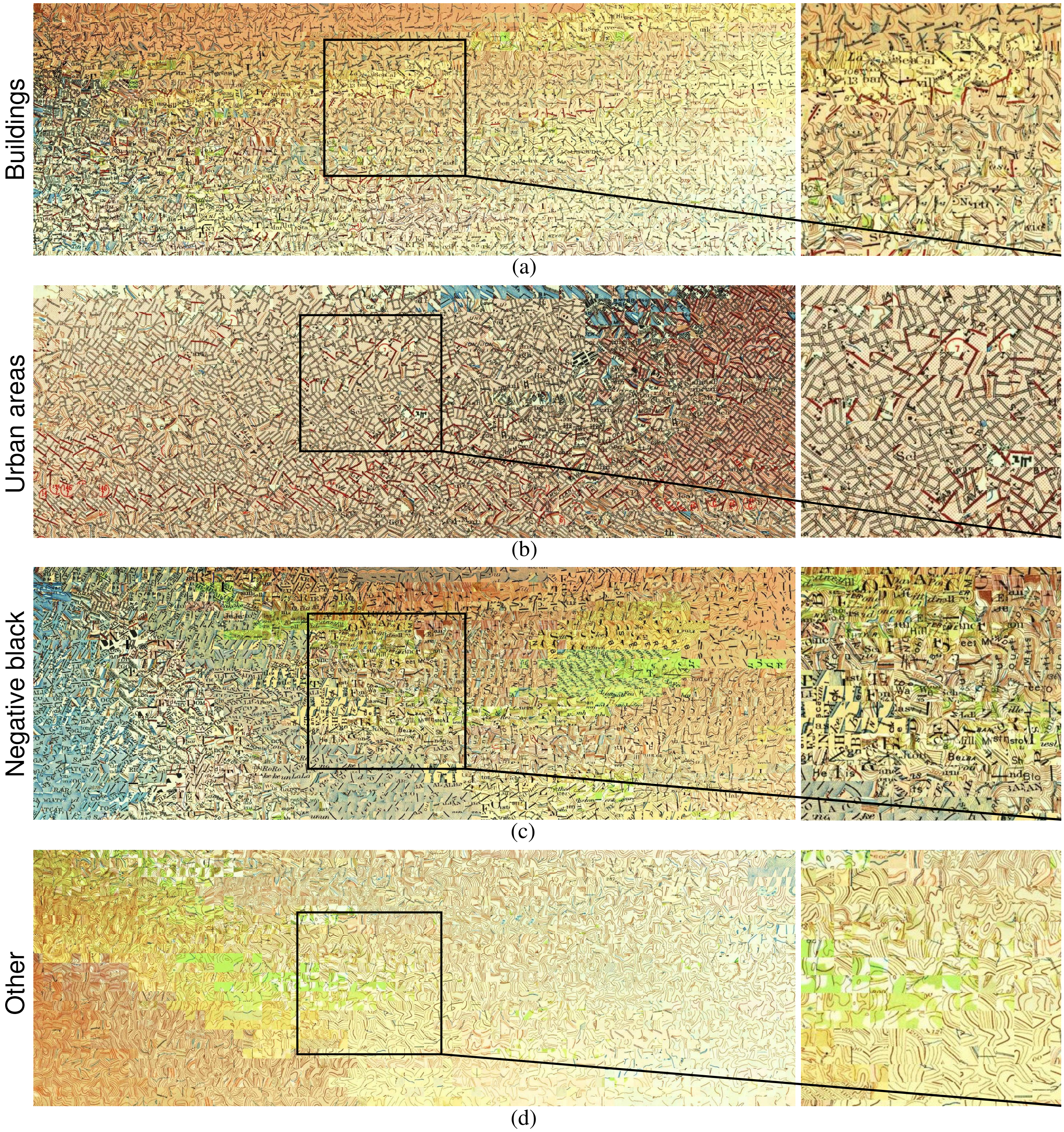


FIGURE 8. T-SNE based visualization of a random sample of the generated training data at size 48 × 48px for the four classes: (a) buildings, (b) urban areas, (c) negative black, and (d) other map content. Respective enlargements to the right.

it is notable that in the case of 48 × 48px input data, AlexNet performs very differently on the classification of urban area: while it underestimates urban area using grayscale data, it tends to overestimate urban area using color images. This is possibly due to overfitting when using grayscale data. When averaging the results across the two input sizes, VGGNet-16 outperforms the other configurations using color input data. When the results are further averaged across the grayscale-color configurations, VGGNet-16 outperforms AlexNet slightly, in average, which can be explained by deeper architecture and thus, higher learning ability.

C. SEMANTIC SEGMENTATION

In the following, we evaluate the segmentation results for VGGNet-16 only, which achieves the highest accuracy levels for the image classification, averaged across the

TABLE 1. Experiment configuration for CNN-based map patch classification.

CNN architecture	Training samples	Sample maps	input images	input size [px]	Training epochs	Learning rate
AlexNet	130k	18	color	48 × 48	10	1×10^{-6}
LeNet	130k	18	color	48 × 48	10	1×10^{-5}
VGGNet16	130k	18	color	48 × 48	3	1×10^{-6}
AlexNet	130k	18	grayscale	48 × 48	10	1×10^{-5}
LeNet	130k	18	grayscale	48 × 48	10	1×10^{-5}
AlexNet	120k	18	color	96 × 96	10	1×10^{-5}
LeNet	120k	18	color	96 × 96	20	1×10^{-6}
VGGNet16	120k	18	color	96 × 96	5	1×10^{-6}
AlexNet	120k	18	grayscale	96 × 96	10	1×10^{-5}
LeNet	120k	18	grayscale	96 × 96	10	1×10^{-6}

TABLE 2. Patch-based classification results for the CNN configurations used in this experiment.

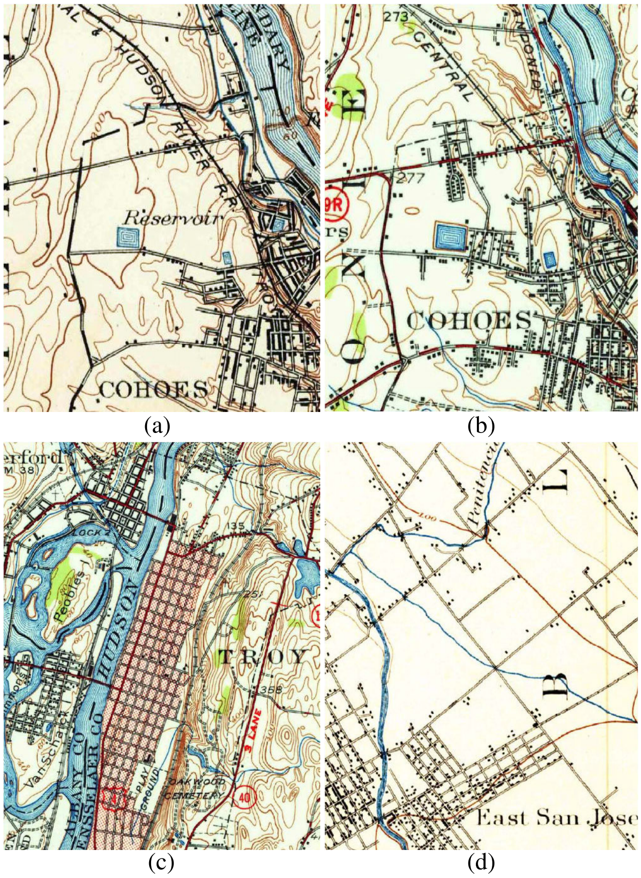
CNN architecture	input images	input size	Overall accuracy	Average F-measure	Precision (Building)	Recall (Building)	Precision (Urban)	Recall (Urban)
AlexNet	color	48 × 48	0.81	0.81	0.93	0.84	0.67	0.93
LeNet	color	48 × 48	0.79	0.79	0.82	0.87	0.82	0.83
VGGNet16	color	48 × 48	0.90	0.90	0.87	0.97	0.88	0.89
AlexNet	grayscale	48 × 48	0.87	0.87	0.86	0.92	1.00	0.79
LeNet	grayscale	48 × 48	0.78	0.77	0.81	0.87	0.78	0.74
AlexNet	color	96 × 96	0.97	0.97	0.95	0.99	1.00	0.99
LeNet	color	96 × 96	0.72	0.72	0.80	0.82	0.74	0.78
VGGNet16	color	96 × 96	0.95	0.95	0.97	0.97	0.99	0.94
AlexNet	grayscale	96 × 96	0.96	0.96	0.95	0.98	0.99	0.99
LeNet	grayscale	96 × 96	0.82	0.82	0.88	0.91	0.85	0.79
AlexNet	color	average	0.89	0.89	0.94	0.92	0.84	0.96
LeNet	color	average	0.76	0.76	0.81	0.85	0.78	0.81
VGGNet16	color	average	0.92	0.93	0.92	0.97	0.94	0.92
AlexNet	grayscale	average	0.92	0.92	0.91	0.95	1.00	0.89
LeNet	grayscale	average	0.80	0.80	0.85	0.89	0.82	0.77
LeNet	average	average	0.78	0.78	0.83	0.87	0.80	0.79
AlexNet	average	average	0.90	0.90	0.92	0.93	0.92	0.93
VGGNet16	color	average	0.92	0.93	0.92	0.97	0.94	0.92

tested scenarios. To assess the performance of the pixel-wise semantic segmentation based on the weakly supervised (i.e., patch-level trained) CNN, we define three test cases (i.e., maps and study areas). These test cases include:

- A multi-temporal test case, using two maps (i.e., from 1893 and 1949) in a peri-urban environment near Cohoes (NY), see Figure 9a,b,
- an urban test case, using a map covering Troy (NY) from the year 1949 (Figure 9c), and
- a test case on an unseen map (i.e., a map not used for training data collection) of East San José (CA) from 1893 (Figure 9d).

We carried out dense pixel-wise inferences for the two trained VGGNet-16 models, which showed best overall performance in the patch-based classification scenarios. Since dense inferences are computationally intensive, we chose a stride of 2, resulting in probability surfaces of a spatial resolution of approximately 10m. Based on the results of previous experiments [17], [18] the detection sensitivity for buildings can be considered as crucial herein, while we expect the extraction of urban area to be less challenging. We evaluate the sensitivity of VGGNet-16 in detecting building symbols for each of the four test maps using ROC plots, based on the manually digitized validation features as categorical variable and the underlying pixel-wise building class probabilities as continuous variable to which a range of discrimination thresholds is applied.

These ROC plots are shown separately for VGGNet-16 based on 48 × 48px input (Figure 10a) and 96 × 96px input (Figure 10b), respectively. These plots show a clear trend, indicating higher degrees of receptiveness for VGGNet-16 trained on small input images, resulting in an average AUC across the four test maps of 0.85, as compared to 0.73 for the large input image configurations. This supports the observation of possible overfitting of the 96 × 96px input data configurations, possibly due to a smaller amount of originally collected training samples, resulting in lower degrees of complexity in the data, alongside with less representative training data, in particular for underrepresented classes.



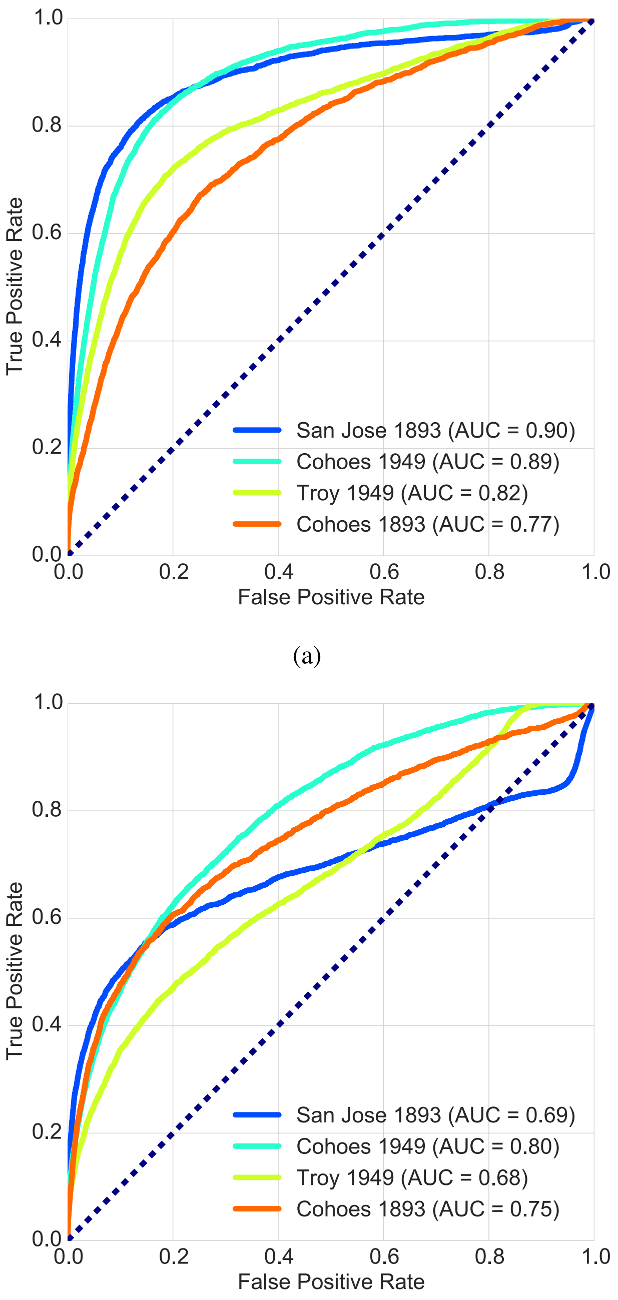
(a) (b)

(c) (d)

FIGURE 9. **Four validation areas (a) Cohoes (NY) 1893 and (b) 1949, (c) Troy (NY) 1949, and (d) East San José (CA) 1893.**

Thus, we evaluate the semantic segmentation performance of VGGNet-16 trained on 48 × 48px input images in more detail for the three study areas. We measure the performance in two ways: Firstly, we carry out a pixel-wise map comparison between the validation data and the hard classification resulting from the CNN class probability surfaces. Secondly, we perform object-based validation, in order to account for the different levels of spatial granularity in validation data and segmentation results based on the weakly supervised CNN. Figure 11 illustrates this problem and highlights three building groups of different levels of spatial granularity loss.

Thus, besides assessing the pixel-wise agreement between segmentation results and validation data, we generate polygonal vector objects from both datasets based on contiguous patches of the target class using raster vectorization. The vectorization of the segmentation results represents the last step in the proposed processing chain (cf. Figure 2), i.e., the generation of analysis-ready geospatial, polygonal vector data. Spatial intersections of polygonal objects from the predicted and the validation dataset are considered correctly classified regardless the actual number of overlapping pixels. Note that object-based validation is not carried out for the urban areas, since due to their large size, the coarser-grained segmentation results do not affect the pixel-based accuracy measures as much as they do in case of the building objects, which often cover only a few pixels. Table 3 shows the accuracy measures for the four maps and the two validation methods.



(a)

(b)

FIGURE 10. **Receiver Operator Characteristic (ROC) curves based on building probabilities estimated by VGGNet-16 and shown for each validation area (a) using 48 × 48px input size and (b) using 96 × 96px input size. The dashed line corresponds to a randomly guessing classifier.**

The pixel-based validation shows relatively low accuracies for all study areas. Comparing the building class performance to previous experiments ([18], precision of 0.07 and recall of 0.78) remarkable increases in precision are observed (average pixel-based building precision of 0.29), indicating that the centering of the building samples around the SIFT

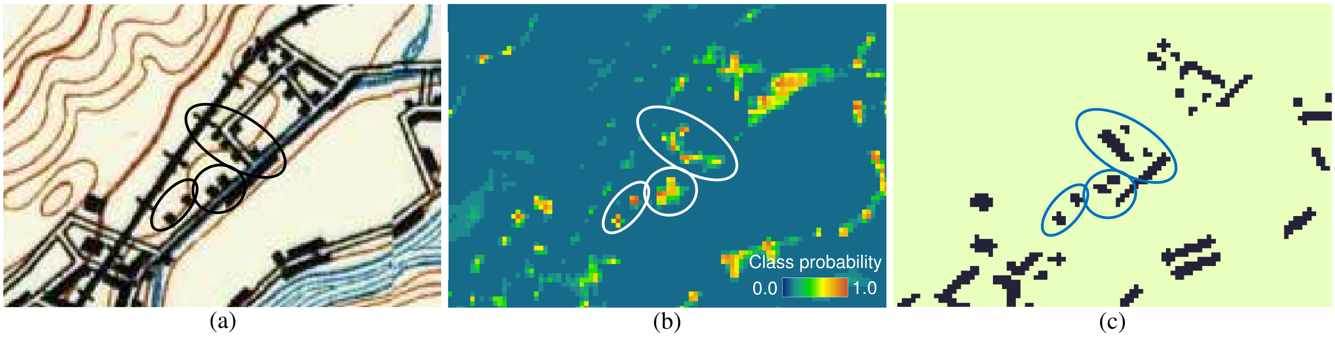


FIGURE 11. Illustrating spatial incompatibilities: (a) original map, (b) corresponding building probability surface from VGGNet-16, and (c) corresponding validation data. Highlighted are three exemplary building groups of different levels of spatial granularity loss.

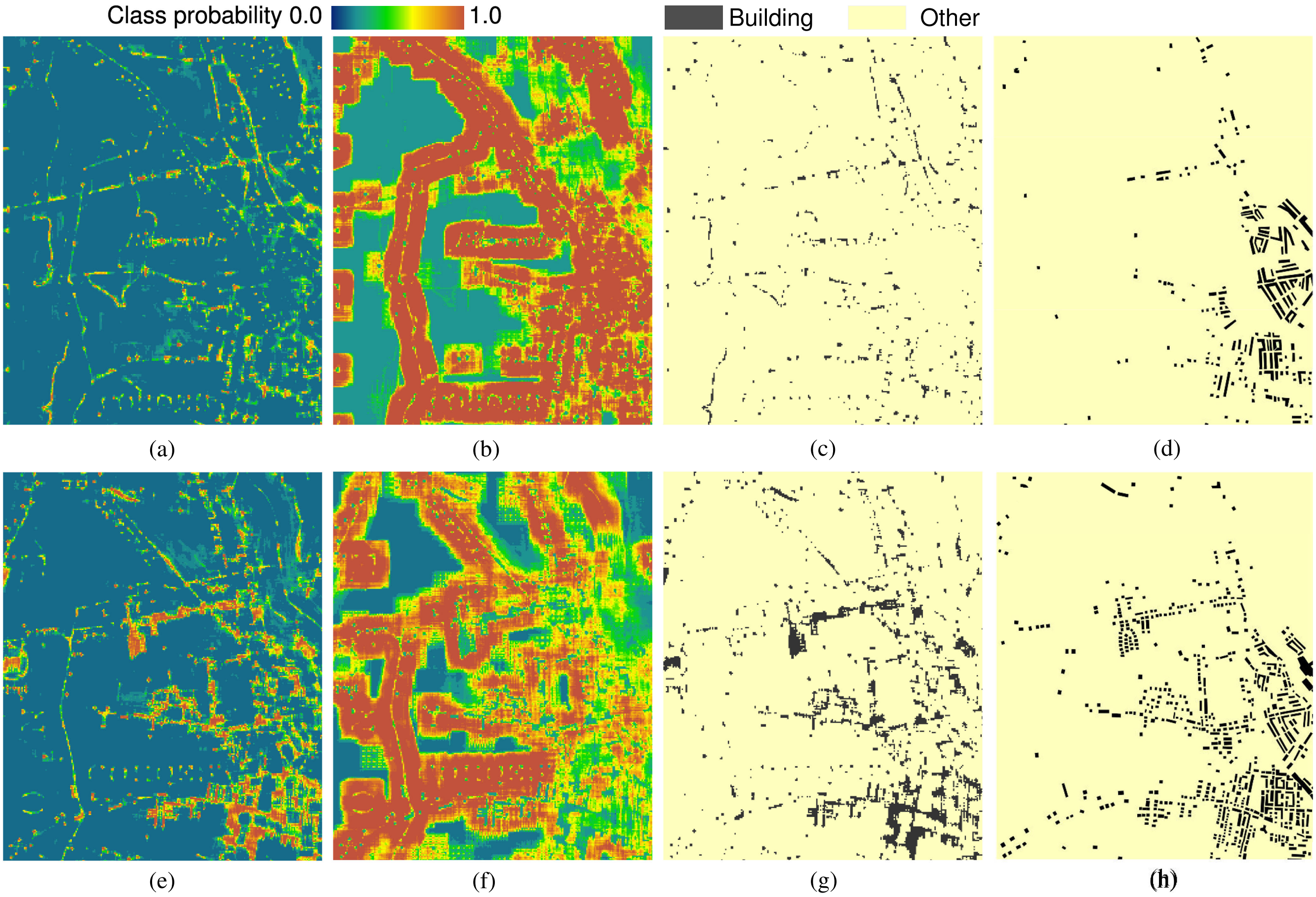


FIGURE 12. Multi-temporal accuracy assessment of the semantic segmentation in Cohoes (NY) (a) VGGNet-16 building probability surface, (b) black non-settlement probability surface, (c) semantic segmentation result, and (d) validation data based on the 1893 map. (e) to (h) show the respective datasets based on the 1949 map.

keypoint reduced effects of translation invariance with respect to buildings, and thus, results in an increase of segmentation granularity, which is a desired effect in this case. Looking at the multi-temporal test case (Figure 12, Cohoes 1895 and 1949 maps) we observe a notable increase in precision and recall of the building class over time. Both maps show good separation between buildings and other black map content such as text. As can be seen in Figure 12c and d, a considerable amount of large buildings are not detected in the 1893 map, but are detected in the 1949 map, likely an effect of increased graphical quality. This effect is also clearly visible when looking at the object-based validation results, where the accuracy measures increase considerably.

The performance of urban area extraction shows a different picture (Figure 13a-d). The recognition of the urban texture seems more challenging than expected, resulting in high

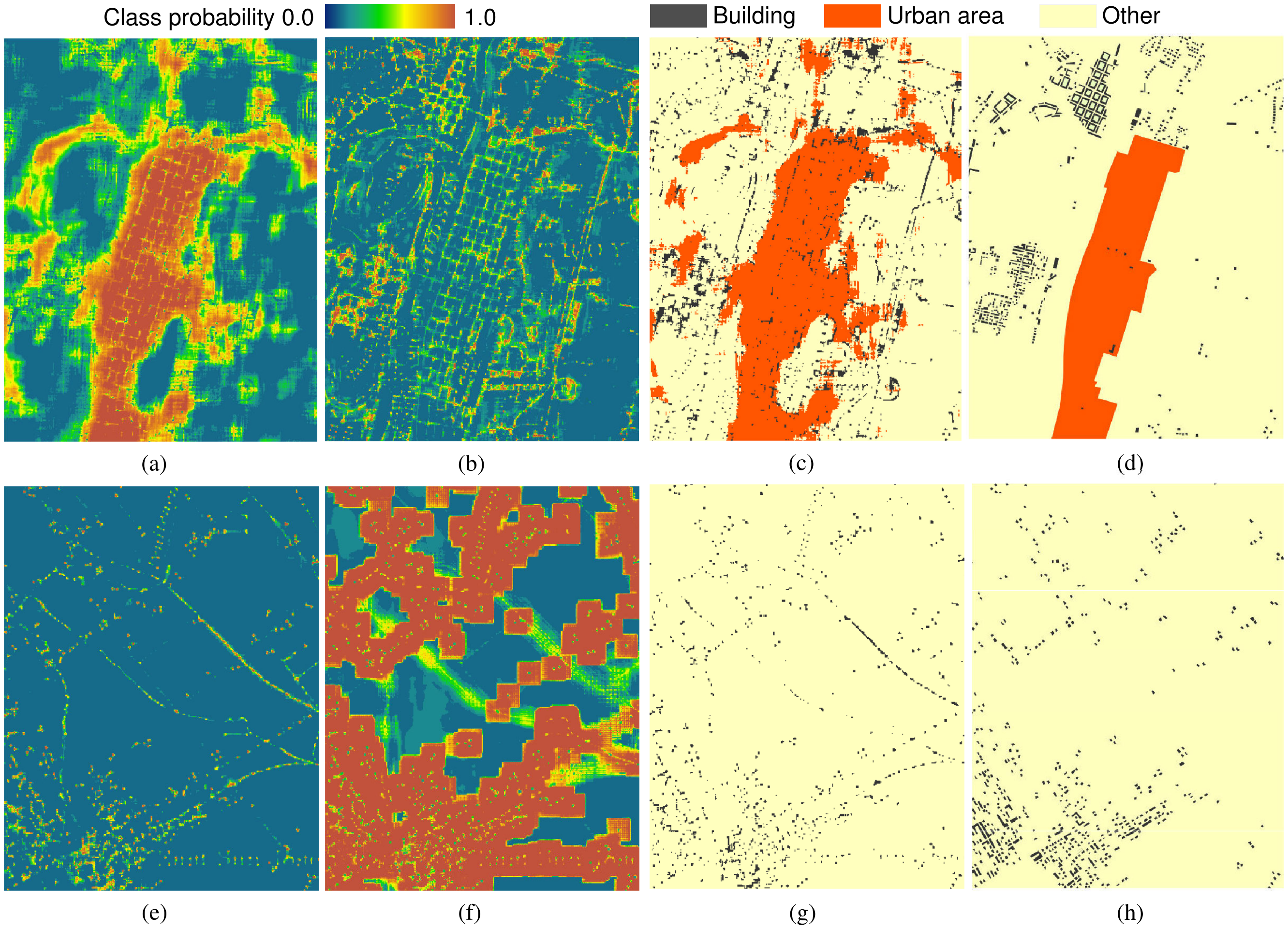


FIGURE 13. Top row: accuracy assessment for the Troy (NY) study area: (a) VGGNet-16 urban area probability surface, (b) building probability surface, (c) corresponding semantic segmentation result, and (d) validation data. Bottom row: Evaluation of VGGNet-16 on a completely unseen map in San José (CA): (e) Building probability surface, (f) black non-settlement content probability surface, (g) corresponding semantic segmentation result, and (h) validation data.

TABLE 3. Pixel-based and object-based validation results.

Site	Cohoes (NY)	Cohoes (NY)	Troy (NY)	Troy (NY)	San José (CA)
Year	1893	1949	1949	1949	1897
Class	building	building	building	urban	building
Pixel-based validation					
Precision	0.20	0.38	0.18	0.50	0.43
Recall	0.16	0.53	0.39	0.95	0.40
F-measure	0.18	0.44	0.25	0.66	0.42
Overall accuracy	0.96	0.92	0.80	0.80	0.97
Kappa	0.15	0.40	0.50	0.50	0.40
Object-based validation					
Precision	0.31	0.53	0.23	-	0.67
Recall	0.71	0.97	0.97	-	0.96
F-measure	0.44	0.68	0.37	-	0.79
IoU	0.28	0.52	0.23	-	0.65

recall values, but lower levels of precision, which may be related to the noise in the training samples in the urban class (cf. Figure 8). Also in this test case, large building blocks are not very well detected. Lastly, we evaluate the performance of the trained VGGNet-16 on the unseen map from San José (CA). Interestingly, building accuracy measures are highest for this map (precision of 0.67 and recall of 0.96 for the object-based validation), indicating remarkable levels of model generalizability. Worth noting is the linear feature in the eastern part of the map misclassified as building (Figure 13g), which represents a brown contour line. We do not observe such misclassifications in large quantity in the Troy and Cohoes maps, possibly indicating that VGGNet-16 has not learned how to classify straight contour lines, since the Albany sampling region is characterized by rather hilly terrain and thus, by more complex contour line geometry.

VI. DISCUSSION AND CONCLUSION

In this study, we described and evaluated a largely automated process chain for information extraction from historical topographic map series using novel training data sampling approaches and deep learning. The application of this framework to the case of human settlement feature extraction achieved promising results given that a priori no training data was available. The use of ancillary spatial data in a hierarchical, spatially stratified sampling scheme has proven to be an effective and efficient way to generate large amounts of training data automatically, exhibiting low levels of label noise. The results of the conducted extraction experiment show a two-fold picture: while high recall values for the building class were achieved, precision is relatively low, partially due to the weak supervision and resulting spatial inflation of segmentation results.

Additionally, we observed a trend of increasing accuracy over time in the multi-temporal experiment (Figure 12). Such difference in performance could be due to inferior graphical quality of older maps, but also a consequence of increasing temporal discrepancies between contemporary ancillary data and the map data, possibly yielding higher levels of label noise in the training data sampled from older maps. At the same time, we observed increasing precision for the building class in the segmentation results as compared to previous experiments [17], [18], [79], confirming that the centering of building training samples applied herein reduces translation-invariance induced decreases in spatial granularity.

While we observed promising extraction results for small, individual building objects, the proposed method did not perform as expected on large buildings or building blocks. Thus, the training data sampling procedure needs to be improved in order to account for these variations in building symbol size, possibly by considering small buildings, large buildings and building blocks as separate classes. Additionally, a terrain-adaptive training data sampling scheme could be developed and tested, to create representative training data for a wide range of terrain-related features such as contour lines or streams. Furthermore, we will test the combination of weakly supervised semantic segmentation with spatial refinement methods such as superpixel-based methods (i.e., unsupervised color-based segmentation of target maps and estimation of the semantics of each segment using a trained CNN, e.g., [26], [80]) which are expected to further improve classification accuracy.

We will revise the sampling scheme for large training patches (i.e., the 96 × 96px data) in order to create more representative training data to prevent overfitting, and, ultimately, test a scale-adaptive ensemble CNN approach. Measures to prevent overfitting will also involve fine-tuning the CNN architecture and training setup by systematically modifying regularization techniques, early-stopping techniques or using different dropout ratios. Additionally, we will implement and test a combined extraction method for different kinds of geographic features and map elements (e.g., text elements). Such a multi-class problem could potentially further improve the precision values for building features achieved in these experiments. In such a case, multi-class training data could potentially be generated in an unsupervised manner using an ensemble of different descriptors, besides SIFT and LBP which we employed herein.

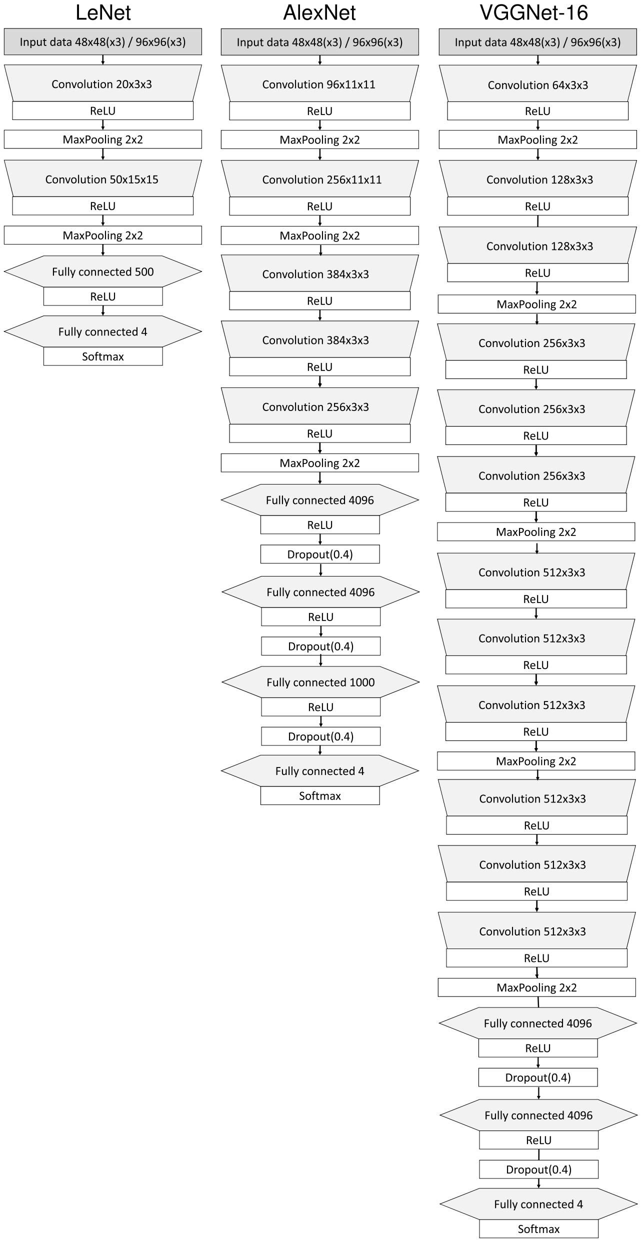


FIGURE 14. The CNN layers and their characteristics in LeNet, AlexNet, and VGGNet-16.

In future work, the outputs of spatially refined multi-class semantic segmentation could be used as pixel-level training data for encoder-decoder CNN architectures in order to perform fully supervised semantic segmentation. The manually digitized pixel-level validation data could also be employed to train an encoder-decoder CNN. In addition to that, the observed trends of increasing extraction accuracy over time will be exploited for explicit, sequential multi-temporal modelling, if the change direction of the process of interest is known (i.e., settlements are expected to grow over time), by sequential processing of temporal stacks of maps chronologically backwards and deriving spatial constraints based on contemporary data (cf. [81], [82]). Such approaches make use of the assumption that spatial and semantic discrepancies between map content and ancillary data are a function of the temporal gap between these datasets, and thus, less prevalent if this temporal gap is small. Hence, sequential retrospective extraction of geographic features starting from the most recent available data could potentially increase the reliability of extracted information and mitigate some of the shortcomings of the presented approaches when applied to earlier maps. Finally, we will test the generalizability of such an approach to other map series. In particular, we will examine the applicability of the presented method for the texture-based training data collection of urban areas on other poorly defined cartographic elements, i.e., composite elements of vaguely defined areal extent, such as forests or swamps.

Lastly, it should be mentioned that we exemplarily applied the proposed framework to human settlement features and historical maps, however, the main concept, which includes the use of contemporary, and possibly publicly available ancillary spatial data for automated training data collection can be transferred to other geographic features of interest, and to other geospatial data sources, such as remote sensing data. The proposed method represents a novel and generalizable strategy for the recognition of small objects in complex visual documents, in cases when only approximate and uncertain a-priori locational information is available to generate graphical examples of the features of interest.

APPENDIX

CNN ARCHITECTURES USED IN THIS WORK

See Figure 14.

ACKNOWLEDGMENT

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. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

REFERENCES

[1] K. A. Fishburn, L. R. Davis, and G. J. Allord, "Scanning and georeferencing historical USGS quadrangles," U.S. Geol. Surv., Reston VA, USA, Tech. Rep. 2017-3048, 2017.

[2] The Library of Congress, Geography and Map Division. *Sanborn Maps*. Accessed: Nov. 25, 2019. [Online]. Available: https://www.loc.gov/collections/sanborn-maps/

[3] The National Library of Scotland. *Ordnance Survey Maps*. Accessed: Nov. 25, 2019. [Online]. Available: https://maps.nls.uk/os/

[4] Swiss Federal Office of Topography Swisstopo. *A Journey Through Time—Maps*. Accessed: Nov. 25, 2019. [Online]. Available: https://www.swisstopo.admin.ch/en/maps-data-online/maps-geodata-online/journey-through-time.html

[5] Y.-Y. Chiang, S. Leyk, and C. A. Knoblock, "A survey of digital map processing techniques," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 1:1-1:44, May 2014.

[6] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, p. 1, Sep. 2017.

[7] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[10] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[12] C. J. Henry, C. D. Storie, M. Palaniappan, V. Alhassan, M. Swamy, D. Aleshinloye, A. Curtis, and D. Kim, "Automated LULC map production using deep neural networks," *Int. J. Remote Sens.*, vol. 40, no. 11, pp. 4416–4440, Jun. 2019.

[13] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?—Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 685–694.

[15] T. Durand, N. Thome, and M. Cord, "WELDON: Weakly supervised learning of deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4743–4752.

[16] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1796–1804.

[17] J. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan, and C. Knoblock, "Extracting human settlement footprint from historical topographic map series using context–based machine learning," in *Proc. 8th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, 2017.

[18] J. H. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan, and C. A. Knoblock, "Spatialising uncertainty in image segmentation using weakly supervised convolutional neural networks: A case study from historical map processing," *IET Image Process.*, vol. 12, no. 11, pp. 2084–2091, Nov. 2018.

[19] S. Leyk and R. Boesch, "Colors of the past: Color image segmentation in historical topographic maps based on homogeneity," *Geoinformatica*, vol. 14, no. 1, pp. 1–21, Jan. 2010.

[20] T. Miyoshi, W. Li, K. Kaneda, H. Yamashita, and E. Nakamae, "Automatic extraction of buildings utilizing geometric features of a scanned topographic map," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2004, pp. 626–629.

[21] S. Laycock, P. Brown, R. Laycock, and A. Day, "Aligning archive maps and extracting footprints for analysis of historic urban environments," *Comput. Graph.*, vol. 35, no. 2, pp. 242–249, Apr. 2011.

[22] M. G. Arteaga, "Historical map polygon and feature extractor," in *Proc. 1st ACM SIGSPATIAL Int. Workshop MapInteraction*, 2013, pp. 66–71.

[23] J. Wu, P. Wei, X. Yuan, Z. Shu, Y.-Y. Chiang, Z. Fu, and M. Deng, "A new Gabor filter–based method for automatic recognition of hatched residential areas," *IEEE Access*, vol. 7, pp. 40649–40662, 2019.

[24] Y.-Y. Chiang, S. Leyk, and C. A. Knoblock, "Efficient and robust graphics recognition from historical maps," in *Graphics Recognition. New Trends and Challenges* (Lecture Notes in Computer Science), vol. 7423. Berlin, Germany: Springer, 2013, pp. 25–35.

[25] Y. Chen, R. Wang, and J. Qian, "Extracting contour lines from common-conditioned topographic maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 4, pp. 1048–1057, Apr. 2006.

[26] Q. Miao, T. Liu, J. Song, M. Gong, and Y. Yang, "Guided superpixel method for topographic map processing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6265–6279, Nov. 2016.

[27] S. Leyk and R. Boesch, "Improving feature extraction of composite cartographic information in low-quality maps," in *Proc. 17th Int. Res. Symp. Comput.-Based Cartogr. (AutoCarto)*, 2008, pp. 8–11.

[28] Y.-Y. Chiang, S. Moghaddam, S. Gupta, R. Fernandes, and C. A. Knoblock, "From map images to geographic names," in *Proc. 22nd ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (SIGSPATIAL)*, 2014, pp. 4–7.

[29] E. Katona and G. Hudra, "An interpretation system for cadastral maps," in *Proc. 10th Int. Conf. Image Anal. Process.*, Jan. 2003, pp. 792–797.

[30] A. J. Saalfeld, "Conflation: Automated map compilation," *Int. J. Geogr. Inf. Sci.*, vol. 2, no. 3, pp. 217–228, 1988.

[31] Y. Li and R. Briggs, "Automated georeferencing based on topological point pattern matching," in *Proc. Int. Symp. Automat. Cartogr. (AutoCarto)*, Vancouver, WA, USA, 2006.

[32] Y. Li and R. Briggs, "An automated system for image-to-vector georeferencing," *Cartogr. Geogr. Inf. Sci.*, vol. 39, no. 4, pp. 199–217, Jan. 2012.

[33] H. Herold, P. Roehm, R. Hecht, and G. Meinel, "Automatically georeferenced maps as a source for high resolution urban growth analyses," in *Proc. ICA 25th Int. Cartographic Conf.*, 2011, pp. 1–5.

[34] J. Weinman, "Toponym recognition in historical maps by gazetteer alignment," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1044–1048.

[35] J. E. Burt, J. White, and G. J. Allord. *QUAD-G: Automated Georeferencing of Scanned Map Images*. Accessed: Nov. 25, 2019. [Online]. Available: https://geography.wisc.edu/research/projects/QUAD-G/files/QUAD-GUserManualver2.10.pdf

[36] W. Duan, Y.-Y. Chiang, C. A. Knoblock, V. Jain, D. Feldman, J. H. Uhl, and S. Leyk, "Automatic alignment of geographic features in contemporary vector data and historical maps," in *Proc. 1st Workshop Artif. Intell. Deep Learn. Geograph. Knowl. Discovery (GeoAI)*, 2017.

[37] W. Duan, Y.-Y. Chiang, S. Leyk, J. H. Uhl, and C. A. Knoblock, "Automatic alignment of contemporary vector data and georeferenced historical maps using reinforcement learning," *Int. J. Geogr. Inf. Sci.*, pp. 1–26, Dec. 2019.

[38] T. Liu, P. Xu, and S. Zhang, "A review of recent advances in scanned topographic map processing," *Neurocomputing*, vol. 328, pp. 75–87, Feb. 2019.

[39] Y.-Y. Chiang, W. Duan, S. Leyk, J. H. Uhl, and C. A. Knoblock, *Using Historical Maps in Scientific Studies: Challenges and Best Practices*. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-319-66908-3.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[41] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Newbury Park, CA, USA: Sage, 2014.

[42] S. Tavakkol, Y.-Y. Chiang, T. Waters, F. Han, K. Prasad, and R. Kiveris, "Kartta labs: Unrendering historical maps," in *Proc. 3rd ACM SIGSPATIAL Int. Workshop AI Geograph. Knowl. Discovery (GeoAI)*, 2019, pp. 48–51.

[43] T. Vopham, J. E. Hart, F. Laden, and Y.-Y. Chiang, "Emerging trends in geospatial artificial intelligence (geoAI): Potential applications for environmental epidemiology," *Environ. Health*, vol. 17, no. 1, p. 40, Dec. 2018.

[44] J. Weinman, "Geographic and style models for historical map alignment and toponym recognition," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 957–964.

[45] L. Dong, F. Zheng, H. Chang, and Q. Yan, "Corner points localization in electronic topographic maps with deep neural networks," *Earth Sci. Inf.*, vol. 11, no. 1, pp. 47–57, Mar. 2018.

[46] J. Weinman, Z. Chen, B. Gafford, N. Gifford, A. Lamsal, and L. Niehus-Staab, "Deep neural networks for text detection and recognition in historical maps," in *Proc. IAPR Int. Conf. Document Anal. Recognit.*, 2019, pp. 902–909.

[47] M. Saeedimoghaddam and T. F. Stepinski, "Automatic extraction of road intersection points from USGS historical map series using deep convolutional neural networks," *Int. J. Geogr. Inf. Sci.*, pp. 1–22, Nov. 2019.

[48] J. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan, and C. Knoblock, "Map archive mining: Visual–analytical approaches to explore large historical map collections," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 4, p. 148, Apr. 2018.

[49] X. Zhou, W. Li, S. T. Arundel, and J. Liu, "Deep convolutional neural networks for map-type classification," 2018, *arXiv:1805.10402*. [Online]. Available: https://arxiv.org/abs/1805.10402

[50] Y. Kiyota, "Promoting open innovations in real estate tech: Provision of the LIFULL HOME'S data set and collaborative studies," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2018, p. 6.

[51] A. Ray, Z. Chen, B. Gafford, N. Gifford, J. J. Kumar, A. Lamsa, L. Niehus-Staab, J. Weinman, and E. Learned-Miller, "Historical map annotations for text detection and recognition," Grinnell College, Grinnell, IA, USA, Tech. Rep., 2018. Accessed: Nov. 25, 2019. [Online]. Available: https://www.cs.grinnell.edu/~weinman/data/complete-map-dataset.pdf

[52] B. Budig and T. C. Van Dijk, "Active learning for classifying template matches in historical maps," in *Discovery Science* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2015, pp. 33–47.

[53] B. Budig, T. C. Van Dijk, and A. Wolff, "Matching labels and markers in historical maps: An algorithm with interactive postprocessing," *ACM Trans. Spat. Algorithms Syst.*, vol. 2, no. 4, p. 13, 2016.

[54] H. Li, J. Liu, and X. Zhou, "Intelligent map reader: A framework for topographic map understanding with deep learning and gazetteer," *IEEE Access*, vol. 6, pp. 25363–25376, 2018.

[55] S. A. Oliveira, I. D. Lenardo, B. Tourenc, and F. Kaplan, "A deep learning approach to cadastral computing," presented at the Digit. Hum. Conf., Utrecht, The Netherlands, 2019.

[56] J. Ignjatić, B. Nikolić, A. Rikalović, and D. Ćulibrk, "Deep learning for historical cadastral maps digitization: Overview, challenges and potential," in *Proc. WSCG Poster Papers*, Aug. 2018.

[57] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Raster-to-vector: Revisiting floorplan transformation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2195–2203.

[58] B. Budig, T. C. Van Dijk, F. Feitsch, and M. G. Arteaga, "Polygon consensus: Smart crowdsourcing for extracting building footprints from historical maps," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (GIS)*, 2016, p. 66.

[59] L. Hurni, C. Lorenz, and L. Oleggini, "Cartographic reconstruction of historic settlement development by means of modern geodata," in *Proc. 26th Int. Cartograph. Conf.*, Dresden, Germany, 2013, pp. 25–30.

[60] A. Tsorlini, I. Iosifescu, C. Iosifescu, and L. Hurni, "A methodological framework for analyzing digitally historical maps using data from different sources through an online interactive platform," *e-Perimetron*, vol. 9, no. 4, pp. 153–165, 2014.

[61] S. Leyk and Y.-Y. Chiang, "Information extraction based on the concept of geographic context," in *Proc. AutoCarto*, 2016, pp. 100–110.

[62] Y. Y. Chiang and S. Leyk, "Exploiting online gazetteer for fully automatic extraction of cartographic symbols," in *Proc. 27th Int. Cartograph. Conf. (ICC)*, 2015, pp. 23–28.

[63] Y.-Y. Chiang, S. Leyk, N. Honarvar Nazari, S. Moghaddam, and T. X. Tan, "Assessing the impact of graphical quality on automatic text recognition in digital maps," *Comput. Geosci.*, vol. 93, pp. 21–35, Aug. 2016.

[64] R. Yu, Z. Luo, and Y.-Y. Chiang, "Recognizing text in historical maps using maps from multiple time periods," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3993–3998.

[65] Zillow. *ZTRAX: Zillow Transaction Assessment Dataset*. Accessed: Nov. 25, 2019. [Online]. Available: https://www.zillow.com/research/ztrax/

[66] E. L. Usery, D. E. Varanka, and L. R. Davis, "Topographic mapping evolution: From field and photographically collected data to GIS production and linked open data," *Cartogr. J.*, vol. 55, no. 4, pp. 378–390, Oct. 2018.

[67] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[68] U.S. Geological Survey. *Topographic Map Symbols*. Accessed: Nov. 25, 2019. [Online]. Available: https://pubs.usgs.gov/GIP/TopoMapSymbols/

[69] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[70] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[71] A. D. Cliff and K. Ord, "Spatial autocorrelation: A review of existing and new measures with applications," *Econ. Geogr.*, vol. 46, no. 1, pp. 269–292, Jun. 1970.

[72] P. A. P. Moran, "The interpretation of statistical maps," *J. Roy. Stat. Soc., B, Methodol.*, vol. 10, no. 2, pp. 243–251, Jul. 1948.

[73] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, vol. 1. New York, NY, USA: Wiley, 1966.

[74] J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.

[75] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[76] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[78] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[79] J. H. Uhl, "Spatio-temporal information extraction under uncertainty using multi-source data integration and machine learning: Applications to human settlement modelling," Ph.D. dissertation, Dept. Geogr., Univ. Colorado at Boulder, Boulder, CO, USA, 2019.

[80] W. Zhao, L. Jiao, W. Ma, J. Zhao, J. Zhao, H. Liu, X. Cao, and S. Yang, "Superpixel–based multiple local CNN for panchromatic and multispectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4141–4156, Jul. 2017.

[81] J. H. Uhl and S. Leyk, "Towards a novel backdating strategy for creating built-up land time series data using contemporary spatial constraints," *Remote Sens. Environ.*, Art. no. 111197, Jun. 2019.

[82] H. Taubenböck, T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech, "Monitoring urbanization in mega cities from space," *Remote Sens. Environ.*, vol. 117, pp. 162–176, Feb. 2012.

JOHANNES H. UHL received the Diploma degree in surveying and geomatics from the Karlsruhe University of Applied Sciences, Germany, in 2009, and the double M.Sc. degrees in geomatics from the Karlsruhe University of Applied Sciences, Germany, and in cartography and geodesy from the Polytechnic University of Valencia (UPV), Spain, in 2011, and the Ph.D. degree in geographic information science from the University of Colorado, Boulder, USA, in 2019. From 2008 to 2009, he was a Student Intern with the Department of Photogrammetry and Image Analysis (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. From 2012 to 2015, he worked as a Geospatial Data Analyst and a Software Developer at Pfalzwerke Netz AG, Ludwigshafen, Germany, and as a Graduate Research Assistant with the Department of Geography, University of Colorado, Boulder, USA, from 2016 to 2019. Since 2019, he has been a Postdoctoral Research Associate at the University of Colorado Population Center (CUPC), Institute of Behavioral Science, University of Colorado, Boulder. His research interests include the efficient integration and analysis of large geospatial datasets, spatio-temporal modeling and information extraction based on multisource geospatial data using machine learning, image processing and statistical analysis, and uncertainty quantification and modeling of spatio-temporal data. He was a recipient of the Best Paper Award at the International Conference of Pattern Recognition Systems (ICPRS) 2017, Madrid, Spain, and received the Gilbert F. White Fellowship from the University of Colorado, in 2018.

STEFAN LEYK is currently an Associate Professor with the Department of Geography, University of Colorado Boulder, and a Research Fellow with the Institute of Behavioral Science. He is also a Geographical Information Scientist with research interests in information extraction, spatio-temporal modeling, and socio-environmental systems. In his work, he uses various sources of historical spatial data to better understand the evolution of human systems and how the built environment interacts with environmental processes in the context of land use and natural hazards.

YAO-YI CHIANG received the bachelor's degree in information management from National Taiwan University and the Ph.D. degree in computer science from the University of Southern California. He is currently an Associate Professor (Research) in Spatial Sciences, the Director of the Spatial Computing Laboratory, and the Associate Director of the NSF's Integrated Media Systems Center (IMSC), University of Southern California (USC). He is also a Faculty in data science at the USC Viterbi Data Science M.S. program. Before USC, he worked as a Research Scientist with Geosemble Technologies and Fetch Technologies, CA, USA. Geosemble Technologies was founded based on a patent on geospatial data fusion techniques, where he was a Co-Inventor. His current research combines spatial science theories with computer algorithms to enable the discovery of useful insights from heterogeneous data for solving real-world problems. His research interests include information integration, machine learning, data mining, computer vision, and knowledge graphs.

WEIWEI DUAN is currently pursuing the Ph.D. degree in computer science with the University of Southern California (USC). She is working on building a computer-vision-based system for extracting information from georeferenced images and storing them in a structured format for analysis, making use of geospatial data integration and using limited amounts of noisy labeling data. Her research interests are computer vision, knowledge graphs, and machine learning.

CRAIG A. KNOBLOCK received the B.Sc. degree from Syracuse University and the master's and Ph.D. degrees in computer science from Carnegie Mellon University. He is currently the Executive Director of the Information Sciences Institute, University of Southern California (USC), a Research Professor of both computer science and spatial sciences at USC, and the Director of the Data Science Program at USC. His research focuses on techniques for describing, acquiring, and exploiting the semantics of data. He has worked extensively on source modeling, schema and ontology alignment, entity and record linkage, data cleaning and normalization, extracting data from the Web, and combining all of these techniques to build knowledge graphs. He has published more than 300 journal articles, book chapters, and conference papers on these topics and has received seven best paper awards on this work. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) and the Association of Computing Machinery (ACM). He was the past President and Trustee of the International Joint Conference on Artificial Intelligence (IJCAI). He received the Robert S. Engelmore Award.