International Journal of Geo-Information

MDPI

Article

Model and Data Integrated Transfer Learning for Unstructured Map Text Detection

Yanrui Zhai [1,2], Xiran Zhou [1,3,*] and Honghao Li [3]

[1] Key Laboratory of Land Environment and Disaster Monitoring, Ministry of Natural Resources, China University of Mining and Technology, Xuzhou 221116, China
[2] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China
[3] School of Computer Science & Technology, China University of Mining and Technology, Xuzhou 221116, China
* Correspondence: xrzhou@cumt.edu.cn

Abstract: The emergence of the third information wave makes extensive maps available to be generated by volunteered ways, never specially designed and generated by professional institutes alone. These large-scale images-based volunteered maps created by the public provide plentiful geographical information regarding a place while posing a challenge for recognizing the unstructured text in these maps for previous approaches to standard map text detection. Map text or map annotations denote the critical element of map content. To achieve the detection of unstructured map text, this paper proposed an integrated data-based and model-based transfer learning model, which mainly respectively included data augmentation techniques and adaptive fine-tuning, to reinforce the state-of-the-art CNNs by transferring the OCR knowledge for detecting the unstructured text units in volunteered maps. The experiment proved that our proposed framework can effectively reinforce the state-of-the-art CNN in detecting unstructured map text. We hope our research results can contribute to unstructured map text detection and recognition.

Keywords: transfer learning; volunteered map; map text detection; map text annotation

Citation: Zhai, Y.; Zhou, X.; Li, H. Model and Data Integrated Transfer Learning for Unstructured Map Text Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 106. https://doi.org/10.3390/ijgi12030106

Academic Editors: Florian Hruby and Wolfgang Kainz

Received: 28 December 2022
Revised: 21 February 2023
Accepted: 1 March 2023
Published: 3 March 2023

1. Introduction

Image-based volunteered maps, which are the results of the digitization of map data, significantly change the ways for map generation and utilization [1,2]. Map text or map annotations refer to the critical element of map content, which could improve and support map content in recognition, map generation, and map retrieval. Text extraction and recognition based on map images or map files have become an important tool for map content discovery, supporting maps in information extraction, historical map reconstruction, and a series of map applications [3–5]. For text detection of professional maps drawn by unified standards, early research mainly utilized classical methods of digital image processing (e.g., morphological operations, digital signal transformations, image pyramids, form templates, etc.) with human–computer interaction techniques [6]. With the successful application of object-oriented techniques in pattern recognition, scholars have proposed object-oriented methods such as image clustering, superpixel, and multi-scale segmentation for map text extraction [7,8], which improve the robustness of extracting map text with varying types and forms. Besides drawing on image processing methods, some scholars have also focused on map mapping features, including geometric features of map characters, string width, the spatial distribution of map characters in horizontal and vertical dimensions, distances between different map characters, and so on [4,9,10].

However, non-standard, non-uniform, and non-professionally generated volunteered maps contained map text in various forms, types, and complex arrangements. The approaches mentioned above generally require salient feature selection and optimized pa-

rameter determination by labor work, which is possible to accomplish for massive various volunteered map text. To further deal with detailed and complex map text localization and detection effectively and to define accurate thresholds and scales to extract effective sparse features, pattern recognition, and machine learning techniques have started to be applied to map text extraction and recognition [11–14]. Although deep learning techniques have significantly improved the level of automation, intelligence, and precision of standardized map text extraction in recent years, the precision of map text extraction for volunteered maps is still hardly more than 70% to support text detection in complex backgrounds in volunteered maps [15,16]. Meanwhile, deep learning also relies on accurately labeled sample data at scale, and it is very difficult to complete large-scale benchmark database samples of volunteered maps if high-quality sample labeling is required for characters, text of variable types, and sizes in volunteered map images.

In comparison to the standard map text annotations being generated by professional mapping rules and approaches, the text units in the volunteered map are unstructured generated, being varied in patterns, sizes, and so on, posing a big challenge for the state-of-the-art techniques to deal with the map text detection.

Currently, transfer learning has become a mainstream branch of small sample-based deep learning (or few-shot learning). The essence of this type of learning is to transform knowledge from the known domain to the unknown domain [17]. In other words, it focuses on measuring the common and different features between the known and the unknown domain and fully exploiting the limited amount of labeled data in the unknown domain to accomplish the task. Reference [18] has successfully developed a CNN for optical character recognition (OCR) to detect the map text in USGS geological maps. Their research proved that the OCR-based CNNs was the potential for detecting the text units in map imagery.

To achieve the detection of unstructured map text, based on the previous research results [18], this paper proposes an integrated data-based and model-based transfer learning model to reinforce the state-of-the-art CNNs for effective detection of the unstructured text units in volunteered maps.

The reminders of this paper are organized as follows. Section 2 discusses the characteristics of unstructured map text in volunteered maps. Section 3 presents the proposed data-based and model-based transfer learning approach. Section 4 presents a series of experiments in terms of OCR-powered CNNs for unstructured map text detection. Section 5 summarizes the content of this paper and related prospects.

2. Unstructured Map Text from Volunteered Maps

Unlike the cartographical products generated by professional institutes, volunteered maps were generated by various sources with different types, patterns, configurations, and even mapping rules [1]. Figure 1 illustrates the text of volunteered maps we collected; these maps' text units hold several characteristics in type, arrangement, and structure.

Figure 1 shows how the map text in the volunteered maps varies in form, color, glyph, being covered by noises, being curved and rotated, and being separated by space. As shown in Figure 1, the map text from those volunteered maps is unstructured, varying in text types in terms of form, glyph, size, and so on. Deep learning always requires massive well-labeled training data to learn the data features for reaching its expected performance, and this means that a great number of labeled map text should be well prepared for deep learning regarding map text detection. However, the number of a majority of these unstructured maps is limited, and the cutting-edge web crawler cannot retrieve all volunteered maps due to the timely updating. That makes it impossible to develop a benchmark dataset including all types of volunteered maps, posing a big challenge for state-of-the-art deep learning in volunteered map text detection.

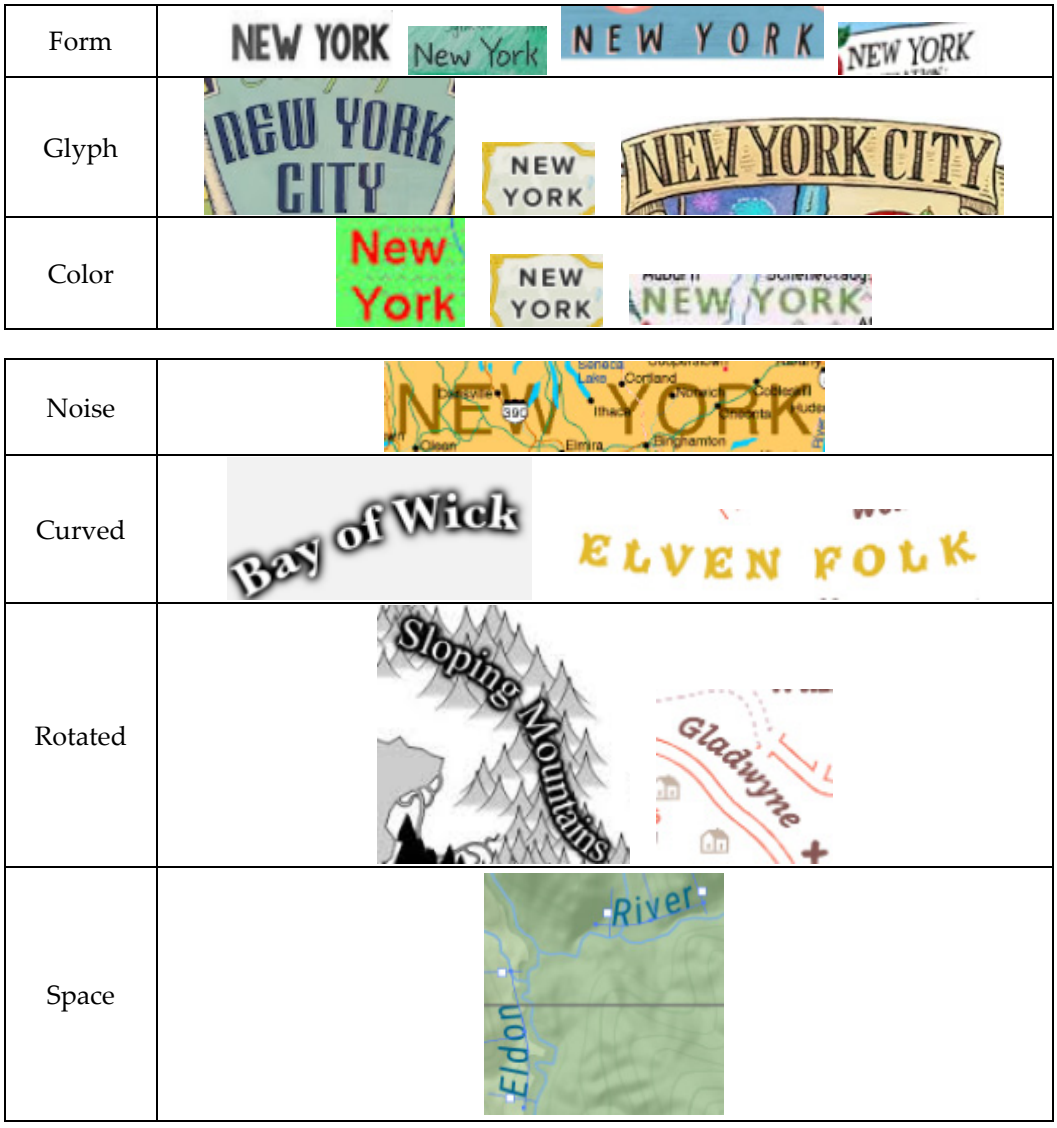

Figure 1. Illustration on unstructured map text from volunteered maps.

3. Methodology

3.1. Basic Architecture of OCR-Powered CNN

The CNNs for OCR mainly include two bribes: regression-based approaches, such as the YOLO-series models, and region proposal-based approaches, such as Fast R-CNN and Faster-R-CNN [19]. Figure 2 illustrates the general architecture of these two bribes of OCR-powered CNN, composed of three main modules: feature extraction module, character target extraction module, and classification module. The feature extraction module employs convolutional layers and pooling layers to extract the visual features from the image file. The classification module exploits a classifier to identify whether the possible text characters are true with the features derived from the feature extraction part. These two bribes hold different character target extraction modules. The character target extraction module of region proposal-based OCR creates a region proposal for possible text characters, and the character target extraction module of regression-based OCR regresses the features of possible text characters. Figure 2a shows the detailed architecture of region proposal-based OCR [19] from the input image to the output result. The keys of this OCR model include the regional proposal network and ROI layer, which are highlighted in red text. Figure 2b shows the detailed architecture of regression-based OCR [20] from the input image to the

output result. The keys of this OCR model include RELU modules and feature pyramid network, which are highlighted in red text.

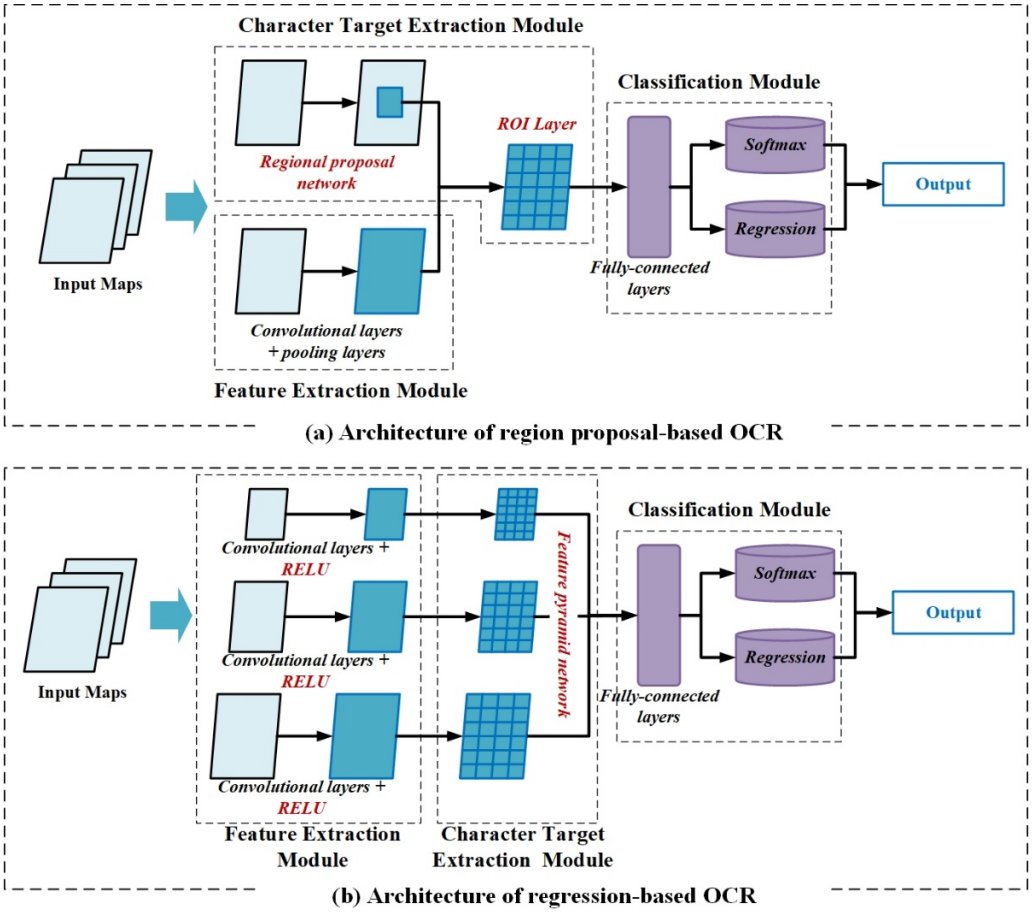

Figure 2. General architecture of an OCR-powered CNN.

3.2. Model-Based and Data-Based Feature Transferring

To enhance the robustness and scalability of CNN in extracting and learning features, researchers proposed transfer learning techniques based on the architecture of deep learning. Based on the architectures of two classical OCR-powered CNN shown in Figure 2, we propose a methodology to integrate data-based transfer learning and model-based transfer learning to transfer the OCR knowledge for unstructured text detection. The architecture of our proposed method is shown in Figure 3, which includes two main parts: data augmentation techniques for transferring data-based features and adaptive fine-tuning for optimizing the model structure of feature transferring. Additionally, data augmentation techniques are similar for region proposal-based OCR and regression-based OCR. For region proposal-based OCR, we employ adaptive fine-tuning to optimize its key part—the regional proposal network. For regression-based OCR, we employ adaptive fine-tuning for optimizing its key part—pyramid feature networks.

(1) Data augmentation for map text

Data Augmentation is a technique that increases the variation and amount of training data within limited available well-labeled data. Expanding the variance of original well-labeled data can promote the capability of discovering the distinctive features for differentiating categories while reducing the interference of irrelevant features. Previous research has proved that data augmentation could effectively improve the performance of CNNs because of the property of the CNN in terms of invariance [21].

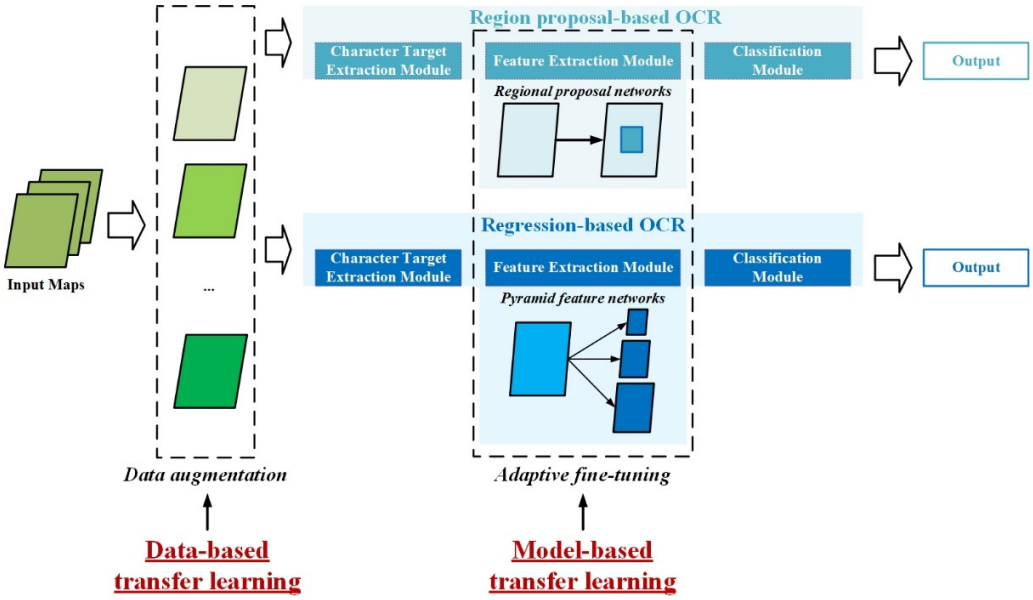

Figure 3. Illustration on our proposed transfer learning for CNN regarding OCR.

The data augmentation for map text detection holds several differences from the ones regarding computer vision tasks, such as scene classification, semantic segmentation, etc. Based on the state-of-the-art data augmentation techniques and cartographical principles, we design a framework for data augmentation shown in Figure 4. The left column shows the original unstructured map text, which varies in colors, styles, and so on. The right column shows the results generated by the augmentation approaches used by the proposed method, including rotation, flipping, contrasts, brightness, color, and noise.

Data Augmentation	Original Map Text	Augmented Map Text
Rotation		
Flip		
Contrast		
Brightness		
Color		
Noise		

Figure 4. Illustration on original instructed map text and their corresponding augmented results.

(2) Adaptive fine-tuning

The conventional fine-tuning method is mainly used on the condition that both the original domain and target domain hold a similar CNN architecture. In detail, this approach focuses on optimizing the CNN by slightly tuning the structure of higher layers of the

CNN architecture while fixing other layers. In practice, people always employ different CNNs for their tasks. Under this condition, it is impossible to determine the layers to be fine-tuned by manual operation. Thus, we propose the adaptive fine-tuning approach, which can determine the layers to be fine-tuned by learning the features between different map text characters.

Assuming that the pre-trained CNN for OCR is $C_o|1, 2, \ldots m|$, where $1, 2, \ldots m$ refers to the sequence of layers in the multiple processing layers of CNN. The CNN transferred for instructed map text detection is $C_m|1, 2, \ldots n|$, where $1, 2, \ldots n$ refers to the sequence of layers in the multiple processing layers of CNN. The objective of adaptive fine-tuning is shown as follows:

$$\theta < \partial(C_o|k_m| - C_m|k_n|), k_m \in \{1, 2, \ldots, m\}; k_n \in \{1, 2, \ldots, n\} \quad (1)$$

where $C_o|k_m|$ denotes the k_m layer of C_o, and $C_m|k_n|$ denotes the k_n layer of C_m. ∂ refers to the algorithm that measures the distance (e.g., Euclidean distance, cosine distance, etc.) between the features extracted from $C_o|k_m|$ and $C_m|k_n|$.

Moreover, in Equation (1), θ refers to the threshold difference between the features extracted from $C_o|k_m|$ and $C_m|k_n|$. When the difference of the features extracted from $C_o|k_m|$ and $C_m|k_n|$ is greater than θ, these features would be transferred, or the parameters in the k_n layer of C_m would be tuned.

Then, to uses T to determine the feature to be transferred from $C_o|k_m|$ to $C_m|k_n|$, we design a weighted matrix for learning neural network, which is shown as follows:

$$\theta < \partial(C_o|k_m| - C_m|k_n|) = > \tau(S_c, w_{s_c,m \to n}) \quad (2)$$

where τ refers to the parameter set in the CNN, S_c refers to the samples labeled as C category (e.g., text character, map text, etc.), and $w_{s_c,m \to n}$ denotes the weights obtained in training $C_o|k_m|$ and $C_m|k_n|$ for identifying S_c.

Based on this learning neural network and feature difference, we further obtain the training loss of transfer features.

$$t_l = L[\tau(S_c, w_{s_c,m \to n})] \times L[\tau(S_c, w_{s_c,m})] \quad (3)$$

where t_l refers to the training loss for transfer learning, L[] denotes the loss function of training. $w_{s_c,m}$ denotes the weights obtained in training $C_o|k_m|$ for identifying S_c.

4. Experiment

4.1. Classical OCR-Powered CNN

(1) CTPN (Connectionist Text Proposal Network)

CTPN [22] is the initial stage network for OCR. It integrates the CNN architecture and bi-directional LSTM into a framework including three parts: convolutional layer, bi-directional LSTM layer, and classification layer. This model exploits a bidirectional LSTM to propose a framework for extracting features by long-term memories. Moreover, this model designs a set of anchors within various height and equal width to locate the positions of text (or generate text proposals) and then employs a text line construction approach to combine neighbored text proposals. This network can detect the text being placed straight at a vertical orientation but might not perform well in dealing with the text being placed in other directions.

(2) DRRG (Deep Relational Reasoning Graph Network)

The architecture of DRRG [23] includes five components: shared convolution layer, text component prediction layer, local graph layer, relational reasoning layer, and link merging layer. The features extracted from the shared convolution layer are given to the text component prediction layer and relational reasoning layer. The text component prediction layer predicts the geometric properties of the text proposal, including center

position, width, height, and rotation angle. The local graph and relational reasoning layer identify the neighborhood among different text proposals based on their center position to determine whether each text proposal should be combined into a new text proposal. This network can effectively deal with grouping text characters. However, it also might not perform well in dealing with the text being placed by non-vertical orientations.

(3) EAST (Efficient and Accurate Scene Text Detector)

EAST [24] is a two-stage end-to-end architecture based on DenseBox and FCN. This network extracts feature maps within various levels and then merges these feature maps with the idea of U-net. The text proposal generated by this network holds two shapes: quadrilateral (QUAD) or rectangle (RBOX). For each text proposal, EAST designs a locality-aware NMS to identify whether the text in a QUAD or an RBOX would be filtered based on its geometry map and score map. Since EAST fuses the features from multiple scales, it can effectively detect multi-scale text proposals and support identifying the text with various angles and orientations. However, this network might have the challenge of dealing with long text and curved text due to the limitation of perceptual field and anchor size in its architecture.

(4) FCENet

The architecture of FCENet [25] includes three parts: deformable convolutions-enhanced residual network layer, feature pyramid network layer, and Fourier contour embedding layer. In detail, Fourier contour embedding layer includes a classification component for predicting the text proposa, and a regression component for predicting the Fourier feature vector of the text in the Fourier domain. Since the Fourier feature vector is compact, making robust to achieve the approximation of true text through reconstructing the text contour point sequence by inverse Fourier transform. This means that FCENet supports fitting the closed text proposals by taking advantage of the Fourier transform. Thus, this network can effectively detect text within various shapes.

(5) DBNet (DB++)

In recent years, segmentation-based approaches have become popular in the field of text detection because of their capability of accurately extracting text characters with various shapes (e.g., curved, multi-orientations). DBNet [26] is one of the classical segmentation-based CNNs for OCR. The architecture of DBNet includes three parts: the pyramid feature layer, the predictive layer, and the reasoning layer. Pyramid feature layers extract multi-scale features and cascade these features into a feature map. Then, the predictive layer generates a probability map and a threshold map based on the feature map. In detail, a probability map creates text region proposals, and a threshold map operates an adaptive binarization to optimize the result of text region proposals. The reasoning layer generates an approximate binary map to identify the true text character and group any possible text characters into a text. DBNet can effectively deal with detecting the text being curved, orientated horizontally, and multi-directional.

(6) PSENet

PSENet [27] is another segmentation-based CNNs for OCR. The architecture of this network includes a feature extraction layer, feature concertation layer, and progressive scale expansion layer. Based on the Resnet, the feature extraction layer obtains four-scale features through downsampling and then upsamples these four scale features. The feature concertation layer concats the four scale features derived from the feature extraction layer into the feature map. The progressive scale expansion layer generates the labels of possible text regions and then uses a breadth-first search algorithm to optimize the size and shape of the text region proposal. PSENet can effectively detect text with various shapes since it has the capability of adjusting the shape of text region proposal.

(7) TextSnake

TextSnake [28] proposes a text snake-like approach for detecting text with any shapes. Its architecture includes an integrated CNNs layer, feature concertation layer, and predictive layer. Integrated CNNs layer generates 2 text regions, 2 text center lines, 1 radius of the circle, and cosine of an angle and sine of an angle. The feature concertation layer generates masked text center lines by multiplying text region and text center and then operates a disjoint set on masked text center lines to obtain the segmentation result of text. The predictive layer uses a striding algorithm to obtain the skeleton line of the text region proposal and employs the predictive radius to generate the text region proposal. TextSnake novelty proposes the geometric properties of the sequence of the snake-like shape allowing addressing the irregular text shapes. Thus, this network can effectively deal with detecting text with irregular curved shapes.

4.2. Dataset and Results

(1) Dataset

The training data include image text annotations and map text annotations. Image text annotations were derived from the state-of-the-art OCR benchmark dataset. Table 1 lists the name and the accessing ways of these OCR benchmark datasets, and the deep learning models being fine-tuned by these OCR benchmark datasets. Moreover, we manually labeled a number of map text as map text annotations.

Table 1. Details of benchmark dataset used in the experiment.

Benchmark Dataset	Links	Used by OCR Models
ICDAR 2013	https://rrc.cvc.uab.es/?ch=2&com=downloads (accessed on 7 February 2023)	CTPN
ICDAR 2015	https://rrc.cvc.uab.es/?ch=4&com=downloads (accessed on 7 February 2023)	CPTN, EAST, FCENet, DB++, TextSnake
CTW1500	https://ctwdataset.github.io/downloads.html (accessed on 7 February 2023)	DRRG, FCENet, DB++, TextSnake
MSRA-TD500	http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_%28MSRA-TD500%29 (accessed on 7 February 2023)	DRRG, EAST, DB++, TextSnake
Total-Text	https://drive.google.com/file/d/1bC68CzsSVTusZVvOkk7imSZSbgD1MqK2/view?usp=sharing (accessed on 7 February 2023)	DRRG, FCENet, DB++, TextSnake
Synth Text	https://www.robots.ox.ac.uk/~vgg/data/scenetext/ (accessed on 7 February 2023)	TextSnake
COCO-Text	https://bgshih.github.io/cocotext/ (accessed on 7 February 2023)	EAST
MLT-2017	https://rrc.cvc.uab.es/?ch=8&com=downloads (accessed on 7 February 2023)	DB++
MLT-2019	https://rrc.cvc.uab.es/?ch=15&com=downloads (accessed on 7 February 2023)	DB++

Besides the OCR benchmark dataset, we provide a limited amount of map text labeled data for data augmentation and model optimization from the deepMap V1 [16]. These map text characters vary in size, color, texture, background, and so on. Figure 5 illustrates the selected map text used for testing in the experiment. There are 68 test data groups in total. These images illustrate the considerable variation of unstructured map text and characters, which are derived from volunteered maps.

Figure 5. General architecture of a transfer learning-enhanced CNN for OCR.

(2) Implementations

We have implemented the classical OCR-powered CNN listed in Section 4.1 by the Pytorch package. The codes for implementing these models are listed as follows:

- CTPN: https://github.com/CrazySummerday/ctpn.pytorch (accessed on 7 February 2023)
- DRRG: https://github.com/GXYM/DRRG (accessed on 7 February 2023)
- EAST: https://github.com/songdejia/EAST (accessed on 7 February 2023)
- FCENet: https://github.com/tamerthamoqa/facenet-pytorch-glint360k (accessed on 7 February 2023)
- DBNet (DB++): https://github.com/WenmuZhou/DBNet.pytorch (accessed on 7 February 2023)
- PSENet: https://github.com/youngguncho/PoseNet-Pytorch (accessed on 7 February 2023)
- TextSnake: https://github.com/princewang1994/TextSnake.pytorch (accessed on 7 February 2023)

We trained these models with training data as mentioned above and employed the pre-train models to detect map text. The results generated by these models are listed in Table 2.

Then, we implemented the data-based and model-based transfer learning on each classical OCR-powered CNN. In detail, we implemented the data augmentation mentioned in Section 3.2 to extend the diversity and heterogeneity of training data. Moreover, based on the architecture of each classical OCR-powered CNN, we exploited the adaptive fine-tuning technique (mentioned in Section 3.2) for the region proposal networks or feature pyramid

networks to enhance the scalability of CNN. The results generated by these models are listed in Table 2.

Table 2. Comparison of precision of map text detected by various approaches.

Map	CTPN	DRRG	EAST	FCENet	DB++	PSENet	TextSnake
1	0.6904	0.8095	0.6547	0.6667	0.8095	0.8095	0.9285
1+	0.7381	0.8333	0.7023	0.6904	0.8333	0.8333	0.9404
2	0.8571	0.8214	0.8214	0.9642	0.9285	0.9642	1
2+	0.9285	0.8571	0.8928	1	0.9642	0.9642	1
3	0.5357	0.7589	0.4642	0.2946	0.6964	0.7589	0.6517
3+	0.5714	0.7589	0.5	0.2946	0.7053	0.7767	0.6607
4	0.6315	0.7368	0.7105	0.728	0.7894	0.6842	0.8859
4+	0.6929	0.7543	0.7543	0.7456	0.7982	0.7192	0.9035
5	0.4869	0.8869	0.7217	0.8173	0.8260	0.9043	0.9304
5+	0.5391	0.9826	0.8	0.8782	0.8695	0.9304	0.9478
6	0.4302	0.7325	0.5465	0.3604	0.6511	0.6395	0.5697
6+	0.4767	0.7441	0.5930	0.4069	0.6860	0.6627	0.5697
7	0.7619	0.7142	0.7619	0.4761	0.7142	0.8571	0.9047
7+	0.8571	0.8095	0.8095	0.523	0.7619	1	0.9047
8	0.6769	0.8461	0.7692	0.8	0.8615	0.8769	0.9384
8+	0.7538	0.9076	0.8153	0.8461	0.9230	0.9076	0.9692
9	0.7457	0.7288	0.8135	0.8305	0.8474	0.8474	0.7796
9+	0.8644	0.8305	0.8813	0.8813	0.8983	0.8983	0.7966
10	0.4363	0.8787	0.8484	0.8848	0.8303	0.9393	0.8182
10+	0.4848	0.9151	0.8666	0.9030	0.8484	0.9636	0.8181
11	0.6019	0.9029	0.3398	0.5048	0.6213	0.8932	0.9514
11+	0.6699	0.9708	0.3592	0.5242	0.6310	0.9320	0.9611
12	0.9583	1	0.8541	0.9167	1	0.9792	0.9792
12+	1	1	0.875	0.9167	1	0.9792	0.9792
13	0.9411	0.9558	0.9705	0.9705	0.9852	1	1
13+	0.9852	1	1	0.9852	0.9852	1	1
14	0.6379	0.7931	0.6896	0.5689	0.8965	0.7413	0.8275
14+	0.6896	0.8620	0.7068	0.5689	0.9482	0.7931	0.8275
15	0.9629	0.9629	0.9629	0.8889	0.9629	1	0.9629
15+	0.9629	0.9629	0.9629	0.8888	0.9629	1	0.9629
avg *	0.7731	0.9179	0.8157	0.8071	0.8907	0.9086	0.9205
avg+ *	0.8056	0.9388	0.8396	0.8313	0.9048	0.9154	0.9221

* avg and avg+ refer to the average precision of all 68 map samples.

(3) Results and discussion

The experiment includes two groups: the first group detects volunteered map text using the state-of-the-art CNN for OCR, and the second group detects volunteered map text with our proposed approaches. The precision of map text detection is expressed as follows:

$$precision = \frac{T}{A} \tag{4}$$

where A refers to the number of all text samples in a volunteered map, and T refers to the number of text samples correctly detected in a volunteered map.

Table 2 lists the results of 15 selected maps. The column *Map* denotes the index of selected volunteered maps; 1/1+ respectively denotes the precision of map detection by various classical approaches and the precision of map detection by various classical approaches with our proposed data and model transfer learning.

We can make several conclusions from the results listed in Table 2.

a. The state-of-the-art CNN for OCR can support to detect of map text from a proportion of volunteered maps, and the precision of detection can reach greater than 85% for a number of maps. Since these state-of-the-art CNNs for OCR are pretrained by a variety of optical image datasets, this means that the text characters in the optical image and the

text characters in the volunteered map have a similar data feature space. Moreover, the results also prove that the architecture of CNN holds good transferability in data feature extraction and model structure optimization.

b. Considering that similarity between two different tasks is the precondition of implementing transfer learning, similar data features and the transferability of CNN architectures means that data-based and model-based transfer learning might be valid for map text detection.

c. However, the text characters of the volunteered maps have the characteristics of complex styles, varied shapes, and random arrangement. It means that large-scale well-labeled data would not be available for volunteered map text. Thus, we found that these classical CNN models seem to be challenging to deal with detecting the text with irregular shapes perform. Moreover, this means the state-of-the-art CNNs for OCR still perform not well on a majority of maps.

d. Generally speaking, the state-of-the-art CNNs for OCR enhanced by our proposed transfer learning operates better than those original CNNs. Without massive well-labeled data that are impossible to establish, deep learning methods might not deal with detecting volunteered map text. Thus, how to design few-shot learning is critical to support volunteered map text detection when large-scale annotated data sets are not available.

e. The techniques of data augmentation and adaptive fine-tuning seem competitive to address detecting map text with different styles, backgrounds, etc. Thus, this means that data and model-integrated transfer learning can improve the precision of volunteered map text detection.

f. We also found that the performance of data-based and model-based transferring learning still holds a distance to the perfect condition of map text detection. Thus, for the text in optical images and volunteered maps, the similarity in terms of data features and model structure could be extended. For example, the mapping of representative features might be helpful in enhancing data-based and model-based transfer learning.

5. Conclusions

Map text annotation is a key map element to represent the real world that a map visualizes and describes. The emergence of big data and the third information wave makes maps, not the products specially designed and generated by professional institutes yet. The volunteered maps created by the public provide plentiful volunteered geographical information regarding a place while posing a challenge to recognizing the unstructured text in these maps. How to effectively employ the power of deep learning with limited amount of training data might be an essential way to accurately and completely obtain the geographical information from the volunteered maps.

This paper proposes to integrate the thought of transfer learning into the volunteered map text detection. We design the data augmentation techniques and adaptive fine-tuning methods to develop a data-based and model-based transfer learning framework enhanced by data transferring and model transferring. The experiment proved that our proposed framework can effectively reinforce the state-of-the-art CNN in detecting unstructured map text.

We believe that future investigations might take two strategies into consideration: how to integrate the cartographical principles into the multi-layer processing of CNN and how to design feature-based transfer learning approaches that reinforce learning the feature space of map text.

Author Contributions: Conceptualization, Xiran Zhou; methodology, Xiran Zhou and Yanrui Zhai; validation, Yanrui Zhai and Honghao Li; investigation, Xiran Zhou; data curation, Yanrui Zhai, Honghao Li, and Xiran Zhou; writing—Xiran Zhou and Yanrui Zhai; writing—review and editing, Xiran Zhou; visualization, Yanrui Zhai and Xiran Zhou; funding acquisition, Xiran Zhou. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 42201473 and 41971370.

Data Availability Statement: The data that support the findings of this study are openly available in "Baidu Wangpan" at https://pan.baidu.com/s/1w6aHU2Uia6e_qrOvtSrjow (accessed on 7 February 2023), with a password: ibs9.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ballatore, A. Defacing the Map: Cartographic Vandalism in the Digital Commons. *Cartogr. J.* **2014**, *51*, 214–224. [CrossRef]
2. Clarke, K.C.; Johnson, J.M.; Trainor, T. Contemporary American cartographic research: A review and prospective. *Cartogr. Geogr. Inf. Sci.* **2019**, *46*, 196–209. [CrossRef]
3. Chiang, Y.Y.; Knoblock, C.A. Recognizing text in raster maps. *Geoinformatica* **2015**, *19*, 1–27. [CrossRef]
4. Li, H.L.; Liu, J.; Zhou, X.R. Intelligent Map Reader: A Framework for Topographic Map Understanding with Deep Learning and Gazetteer. *IEEE Access* **2018**, *6*, 25363–25376. [CrossRef]
5. Chiang, Y.; Duan, W.; Leyk, S.; Uhl, J.H.; Knoblock, C.A. Creating structured, linked geographic data from historical maps: Challenges and trends. In *Using Historical Maps in Scientific Studies*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 37–63.
6. Chiang, Y.Y.; Leyk, S.; Knoblock, C.A. A Survey of Digital Map Processing Techniques. *ACM Comput. Surv.* **2014**, *47*, 1–44. [CrossRef]
7. Miao, Q.G.; Liu, T.G.; Song, J.F.; Gong, M.G.; Yang, Y. Guided Superpixel Method for Topographic Map Processing. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6265–6279. [CrossRef]
8. Long, S.B.; He, X.; Yao, C. Scene Text Detection and Recognition: The Deep Learning Era. *Int. J. Comput. Vis.* **2021**, *129*, 161–184. [CrossRef]
9. Chiang, Y.Y.; Leyk, S.; Nazari, N.H.; Moghaddam, S.; Tan, T.X. Assessing the impact of graphical quality on automatic text recognition in digital maps. *Comput. Geosci.* **2016**, *93*, 21–35. [CrossRef]
10. Liu, T.E.; Xu, P.F.; Zhang, S.H. A review of recent advances in scanned topographic map processing. *Neurocomputing* **2019**, *328*, 75–87. [CrossRef]
11. Armstrong, M.P. Active symbolism: Toward a new theoretical paradigm for statistical cartography. *Cartogr. Geogr. Inf. Sci.* **2019**, *46*, 72–81. [CrossRef]
12. He, Y.F.; Sheng, Y.H.; Jing, Y.Q.; Yin, Y.; Hasnain, A. Uncorrelated Geo-Text Inhibition Method Based on Voronoi K-Order and Spatial Correlations in Web Maps. *ISPRS Int. J. Geo.-Inf.* **2020**, *9*, 381. [CrossRef]
13. Uhl, J.H.; Leyk, S.; Chiang, Y.Y.; Duan, W.W.; Knoblock, C.A. Automated Extraction of Human Settlement Patterns from Historical Topographic Map Series Using Weakly Supervised Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 6978–6996. [CrossRef]
14. Hu, Y.J.; Gui, Z.P.; Wang, J.M.; Li, M.X. Enriching the metadata of map images: A deep learning approach with GIS-based data augmentation. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 799–821. [CrossRef]
15. Ory, J.; Christophe, S.; Fabrikant, S.I.; Bucher, B. How Do Map Readers Recognize a Topographic Mapping Style? *Cartogr. J.* **2015**, *52*, 193–203. [CrossRef]
16. Zhou, X. GeoAI-Enhanced Techniques to Support Geographical Knowledge Discovery from Big Geospatial Data. Ph.D. Thesis, Arizona State University, Tempe, AZ, USA, 2019.
17. Zhuang, F.Z.; Qi, Z.Y.; Duan, K.Y.; Xi, D.B.; Zhu, Y.C.; Zhu, H.S.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
18. Zhou, X.; Li, D.; Xue, Y.; Wan, Y.; Shao, Z. Intelligent Map Image Recognition and Understanding: Representative Features, Methodology and Prospects. *Geomat. Inf. Sci. Wuhan Univ.* **2022**, *47*, 641–650. [CrossRef]
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
22. Tian, Z.; Huang, W.L.; He, T.; He, P.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9912, pp. 56–72.
23. Zhang, S.; Zhu, X.; Hou, J.; Liu, C.; Yang, C.; Wang, H.; Yin, X. Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9696–9705.
24. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2642–2651.

25. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3122–3130.
26. Aroudi, A.; Braun, S. DBnet: Doa-Driven Beamforming Network for end-to-end Reverberant Sound Source Separation. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP–2021), Toronto, ON, Canada, 6–11 June 2021; pp. 211–215.
27. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape Robust Text Detection with Progressive Scale Expansion Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 9328–9337.
28. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. Textsnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 20–36.