Detecting Omissions in Geographic Maps through Computer Vision

Phuc Nguyen
VinAI Research, Vietnam
v.phucnda@vinai.io

Anh Do
Ministry of Information & Communications, Vietnam, dcanh@mic.gov.vn

Minh Hoai
VinAI Research, Vietnam
v.hoainm@vinai.io

***Abstract*—This paper explores the application of computer vision technologies to the analysis of maps, an area with substantial historical, cultural, and political significance. Our focus is on developing and evaluating a method for automatically identifying maps that depict specific regions and feature landmarks with designated names, a task that involves complex challenges due to the diverse styles and methods used in map creation. We address three main subtasks: differentiating maps from non-maps, verifying the accuracy of the region depicted, and confirming the presence or absence of particular landmark names through advanced text recognition techniques. Our approach utilizes a Convolutional Neural Network and transfer learning to differentiate maps from non-maps, verify the accuracy of depicted regions, and confirm landmark names through advanced text recognition. We also introduce the VinMap dataset, containing annotated map images of Vietnam, to train and test our method. Experiments on this dataset demonstrate that our technique achieves F1-score of 85.51% for identifying maps excluding specific territorial landmarks. This result suggests practical utility and indicates areas for future improvement. https://github.com/VinAIResearch/VinMap**

***Index Terms*—Map analysis, Vietnam map, landmark detection, Hoang Sa, Truong Sa**

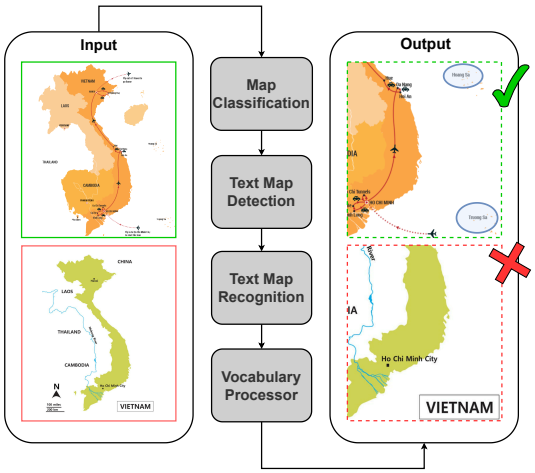

Fig. 1. Our proposed model, pre-trained on high-quality VinMap datasets, demonstrates the capability to recognize the Vietnam map and determine whether it includes the Hoang Sa or Truong Sa regions from multi-resolution input map images.

I. Introduction

Maps, representing one of the earliest forms of images, are deeply significant to our understanding of the environment and our place within it. They encapsulate more than just geographical information; maps are instilled with cultural, political, and historical meanings, making them a rich subject for analysis. Building a computer vision algorithm for map analysis is crucial as it enables the automatic extraction and interpretation of these layers of data embedded within maps. Such technology not only enhances our ability to understand historical changes and cultural insights but also aids in real-time detection and prevention of malicious activities. By leveraging computer vision for map analysis, we can unlock a more nuanced and comprehensive understanding of the multifaceted information that maps provide, ensuring that this valuable resource can be utilized to its fullest potential in various fields.

One particular computer vision technology that is potentially beneficial is the one capable of identifying maps that depict specific regions and feature landmarks with designated names. This technology is crucial in a variety of scenarios, especially when it involves navigating through vast archives of historical documents or continually scanning produced content. These capabilities are indispensable for historical research, such as determining when a city or region was first referred to by a specific name—a key to understanding geopolitical shifts, including the adoption of new political boundaries or name changes due to evolving political dynamics. For example, this technology can pinpoint the time when the ancient city of Alexandria was first labeled as such on maps, as well as when its inhabitants began to recognize and embrace this name. Moreover, this technology could be pivotal in demystifying mythical places such as Atlantis, analyzing cartographic records across different eras to trace the development of its legend and its impact on cultural narratives.

Having mentioned the potential use of computer vision for automatic map analysis, its efficacy for identifying maps that depict a specific region with landmarks bearing designated names remains uncertain. This complexity arises from the need to tackle three non-trivial subtasks: 1) distinguishing maps from non-map images; 2) verifying that the map in question actually represents the region of interest; and 3) confirming the presence or absence of a specific landmark name on the map. These subtasks are challenging due to the diverse styles and methods used in map creation. In the first subtask, although differentiating maps from other types of images might seem straightforward given the current advancement in computer

vision and transfer learning, achieving perfect accuracy is complicated by the fact that hand-drawn maps can closely resemble other artistic illustrations. The second subtask, determining whether or not a map contains the targeted area, is also challenging because maps of the same locale can vary significantly in appearance based on their intended content and creation process. Conversely, maps of different regions may share stylistic elements, making them misleadingly alike. The final subtask involves text spotting and recognition, which is rendered difficult by the variety of text presentations, ranging from neatly printed to cursive handwriting, and including text set in vertical, slanted, or curved orientations.

To assess the efficacy of computer vision for map analysis, this paper develops a method and evaluates its performance through a particular scenario. We consider the challenge of sifting through an extensive collection of images to identify maps that either depict Vietnam in its entirety or include segments of it. Furthermore, we consider the nuanced task of determining if such a map specifically excludes the contested islands Hoang Sa (Paracel) and Truong Sa (Spratly), which are at the center of international territorial disputes. While acknowledging the inherent political sensitivity of this issue, our study deliberately concentrates on the technical aspects. The selected problem serves as an exemplar for the actual demands of automated map analysis, encompassing all the subtasks previously outlined and providing a comprehensive test case for our developed method.

Our approach leverages state-of-the-art computer vision techniques presented in Fig. 1. Initially, using a Convolutional Neural Network [17, 13] and transfer learning [20], it determines whether an input image is a map of Vietnam. If the image passes this initial test, we proceed to detect and recognize all text present on the map. Subsequently, each instance of recognized text is cross-referenced with acceptable variations of the names Truong Sa and Hoang Sa to confirm whether these islands are depicted on the map.

To train and evaluate the performance of our method, we have assembled a dataset comprising a variety of map images of Vietnam, either in entirety or in sections, which will be referred to as the **Viet**nam Map or VinMap collection. This dataset includes a range of maps obtained from varied geographic sources, featuring inscriptions in either Vietnamese or English. The collection is organized with labels that confirm the map's depiction of significant territories, the Hoang Sa and Truong Sa islands. Additionally, it is enhanced with box annotations that mark the precise locations of text related to the Hoang Sa and Truong Sa, aiding in the fine-tuning of text recognition processes.

Experiments performed on the VinMap dataset reveal that our method can detect maps of Vietnam that exclude the Truong Sa and Hoang Sa islands with precision and recall values of 78.51% and 93.87%, respectively. These findings illustrate both the strengths and limitations of the approach. On the positive side, the algorithm performs well enough to be considered for practical applications, particularly in scenarios that involve large-scale map scanning where some level of human verification is acceptable. However, the results are not flawless, indicating a clear need for continued research and improvement in this field.

In short, the contributions of this paper are threefold. First, we introduce and explore a new avenue in map analysis, an area ripe for investigation with significant potential impact. Second, we have developed a complete program harnessing cutting-edge computer vision technologies, and we examine its effectiveness through a specific use case. Our experimental results reveal satisfactory performance by our method, underscoring the value of computer vision while also highlighting avenues for further enhancements. Lastly, we introduce the VinMap dataset, a comprehensive collection of thousands of annotated map images, which serves not only as the foundation for developing and testing our approach but may also be instrumental for future endeavors in map analysis tasks.

II. Related Works

A. Map Classification

Previous research [10, 4] addressed the map-matching problem and proposed solutions at a local image scale, primarily utilizing methods that aggregate observed elementary data similarities rather than deep learning frameworks. In contrast, deepMap dataset was [19] introduced to explore map classification using deep learning techniques. They employed a straightforward deep convolutional neural network architecture, which led to significant improvements compared to heuristic approaches. Subsequent studies [2, 8] have further advanced this field by leveraging deep learning models to extract deep-level features for multi-resolution maps. However, recent practice favors pre-training the model on datasets like [16] before fine-tuning it on a specific dataset, as it yields more promising results, becomes a standard approach.

B. Text Detection

Text detection is an important research problem and it has received much research attention. Earlier studies [1, 5] used machine learning clustering-based algorithms to extract text from background images. While these methods are relatively straightforward, they tend to achieve inferior performance. Recently, deep learning-based approaches [12, 11] have shown much better performance in text detection tasks. In our work, we propose to fine-tune a detection model that has been pre-trained on a public dataset such as [18, 15] to our dataset to direct its attention and adapt it to the task of detecting text on maps, especially text for the landmark names of the two sets of islands.

C. Text Recognition

Previous work on Text recognition or Optical Character Recognition (OCR) [6] typically employed simple Neural Networks to perform logistic regression for preset characters. However, due to their simplicity, these models can only recognize a single character at a time. In contrast, recent Transformer-based methods [9, 14] are fast, scalable, and patch-based, achieving promising results by processing a text

region as a whole query-able feature vector. Particularly noteworthy is the fine-tuning of OCR models on specific language datasets such as [15] from pre-trained ENG text recognition datasets like [7, 3], enabling them to become multilingual.

III. Task and Dataset

This section introduces a new challenge in map understanding and offers detailed statistics for the newly proposed VinMap dataset.

A. Task definition

We focus on the task of scanning images to identify Vietnam maps that do not include Hoang Sa (Paracel) and Truong Sa (Spratly). We consider both English and Vietnamese text descriptions of the two islands. We frame this as a detection problem, where the positive class comprises Vietnam maps that exclude both Hoang Sa (Paracel) and Truong Sa (Spratly). All other cases are considered negative, including non-map images, non-Vietnam maps, or Vietnam maps that contain either Hoang Sa (Paracel) or Truong Sa (Spratly).

B. VinMap Dataset

The VinMap dataset comprises a total of 6,858 images with diverse resolutions. Among these, 2,000 images are non-map images, 2,777 maps do not depict Vietnam, and 1,002 maps represent Vietnam and include either the Truong Sa or Hoang Sa islands (866 maps are in Vietnamese, and 136 maps are not in Vietnamese). There are 1,079 maps of Vietnam that do not contain both the Truong Sa and Hoang Sa islands (291 maps are in Vietnamese, and 788 maps are not in Vietnamese). Vietnam maps encompass various geographic regions, yet to instruct vision models to prioritize specific map areas such as Truong Sa and Hoang Sa, adhering to governmental regulations, VinMap offers box annotations for every Vietnam map containing both the Truong Sa and Hoang Sa islands. This meticulous annotation process establishes the groundwork for advancing map analysis research in Vietnam. Box annotations are depicted in Fig. 3. Table I summarizes the statistics of the VinMap dataset. Some images of VinMap are shown in Fig. 2. The dataset presents several advantages, introducing more challenging aspects than previous map datasets.

Type of images	Language	#Train	#Test	Total	Annotation?
Not maps	Mixed	1400	600	2,000	✗
Not Vietnam maps	Mixed	1944	833	2,777	✗
Vietnam maps containing (TS or HS)	Vietnamese	606	260	866	✓
	English	95	41	136	
	Sub-total	701	301	1,002	
Vietnam maps not containing (TS and HS)	Vietnamese	204	87	291	✗
	English	552	236	788	
	Sub-total	756	323	1,079	
Total		4,801	2,057	6,858	

TABLE I: Statistics of the VinMap dataset. There is a total of 6,858 images, divided into disjoint training and testing subsets of 4801 and 2057 images, respectively.

IV. Experiments

A. Proposed Method

This section describes the proposed method, which consists of several steps: map classification, text detection, text recognition, and vocab matching; depicted in Fig. 4

Map classification. Our objective is to categorize map images into two groups: those that depict Vietnam and those that do not. To accomplish this, we utilize the EfficientNet-B4 classification model [17], modifying the final classification layer to output only two categories instead of the original 1,000. The remaining layers are initialized with weights pre-trained on the ImageNet dataset [16]. Training is conducted on the training set of VinMap comprising 4,801 images, with 3,344 non-Vietnam maps (1,400 non-map images, 1,944 maps depicting regions other than Vietnam) and 1,457 maps of Vietnam. We employ Cross-Entropy Loss over 100 epochs, with a batch size of 4 and a learning rate of 0.1. Additionally, we apply random crop-flip augmentations to enhance training data diversity.

Text Detection. Our objective is to spot semantic text regions within map images depicting Vietnam, with a particular focus on the key regions of the two islands. Initially, we utilize DBNet [11], pre-trained on the English ICDAR2015 dataset [7]. We adopt a two-step training approach. In the first step, we aim to adapt the model to recognize Vietnamese text regions. To achieve this, we fine-tune the model on 33,000 Vietnamese text instances from 1,200 training images from the VinText dataset [15]. In the second step, we refine the model to specifically focus on the key regions (text regions of the two islands). For this stage, we fine-tune the model using the provided training set VinMap box annotations of the two islands within 701 maps. Throughout both stages, we utilize the ResNet50 backbone and employ combination probability, binary, and threshold map losses [11], with a batch size of 2 and a learning rate of 0.001. Additionally, we incorporate random crop and rotate augmentations to better accommodate map images.

Text Recognition. The objective is to comprehend the semantic information extracted from the detected text regions. To accomplish this, we utilize the open-source VietOCR [1], which is built upon the Transformer OCR architecture [9]. This VietOCR tool has been pre-trained on an extensive dataset comprising over 10 million synthetic, handwritten, and scanned images. Since the pre-trained OCR model performed effectively on both Vietnamese and English text, we did not finetune it further.

Vocab Matching. The objective is to align predicted text instances with a predefined vocabulary policy. This involves calculating the Levenshtein distance [2] between the known vocabulary and the text predicted by the OCR model described in Fig. 4. If the distance is smaller than a threshold value,

[1] https://github.com/pbcquoc/vietocr

[2] https://en.wikipedia.org/wiki/Levenshtein_distance

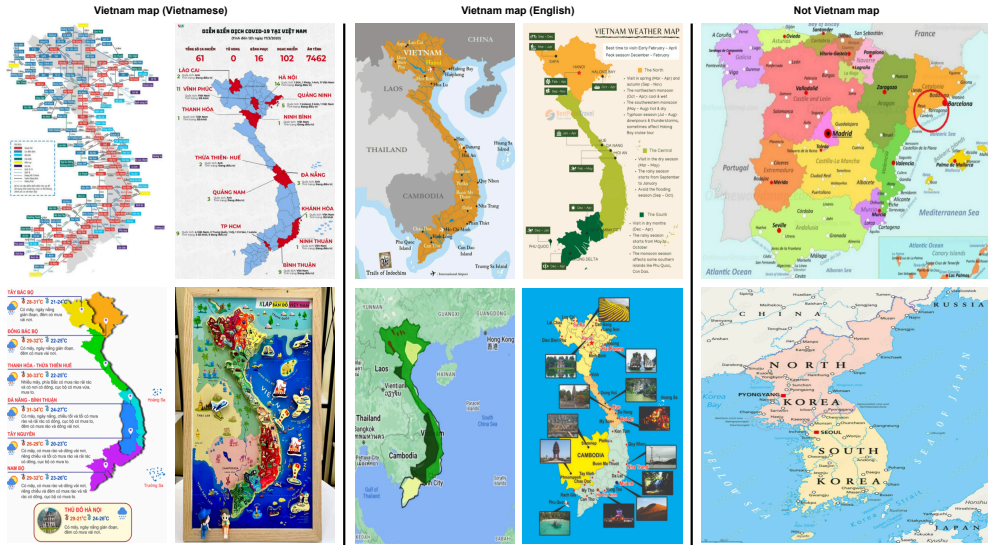

Fig. 2. The VinMap dataset comprises high-quality images in both English and Vietnamese. Tailored specifically for Vietnam, the Vietnam map set encompasses maps depicting various contexts of Vietnam, whereas the Not Vietnam map set comprises map images from diverse countries and regions.

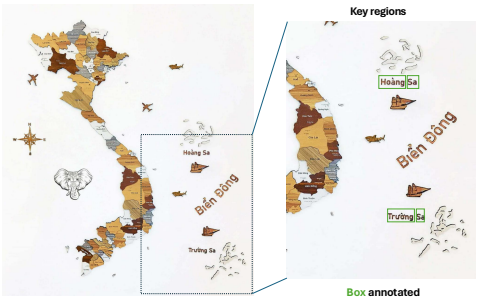

Fig. 3. Annotation visualization of the Vietnam map image and our provided box annotation for regions of interest.

denoted as $\lambda = 2$, the predicted text region is considered a match with the key text regions—in our case, the two islands.

B. Evaluation protocol

We regard the map's final prediction according to Section III-A. To evaluate the proposed pipeline on the VinMap dataset, we consider:

- **Precision**: measures the accuracy of the positive predictions. It answers the question: "Of all the images predicted as Vietnam maps that do not contain Truong Sa and Hoang Sa, how many were actually maps that do not contain Truong Sa and Hoang Sa?"
- **Recall**: measures the ability of the model to find all the Vietnam map that do not contain Truong Sa and Hoang Sa. It answers the question: "Of all the actual Vietnam maps that do not contain Truong Sa and Hoang Sa, how many were correctly detected?"
- **F1-Score**: The harmonic mean of precision and recall provides a balanced measure that considers both precision and recall equally: $F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$

C. Results Analysis

TABLE II: The comparison of the proposed pipeline and the naive image classification approach

Method	Test setting	Precision	Recall	F1-Score
Raw Image Classification	ENG-VN	39.25	53.84	45.40
VinMap	ENG	93.12	95.91	94.49
VinMap	VN	65.53	91.28	76.29
VinMap	ENG-VN	78.51	93.87	85.51

TABLE III: Results on the related tasks

Task	AP
detect map from ALL images	99.6
detect VN maps from ALL images	97.52
detect VN maps not containing (HS and TS) from ALL images	75.21

We present detailed quantitative results for VinMap using both the proposed pipeline and the Raw Image Classification pipeline, as illustrated in Table II. In the Raw Image Classification setting, we train a binary classification EfficientNet-B4 model on the VinMap training set following the policy outlined in Section III-A. The direct approach yields only a 45.40 % F1-Score on the English-Vietnamese test set, indicating significant challenges posed by the VinMap dataset. Conversely, our proposed pipeline demonstrates substantial improvement across three evaluation metrics for the dataset. Specifically, in the case of the Vietnamese maps test set, our method experiences a notable drop from the English maps test set by 18.2% F1-Score and 27.59% Precision score, while maintaining a Recall rate of over 90%. The decrease in performance can be ascribed to the challenges associated with identifying five distinct diacritics

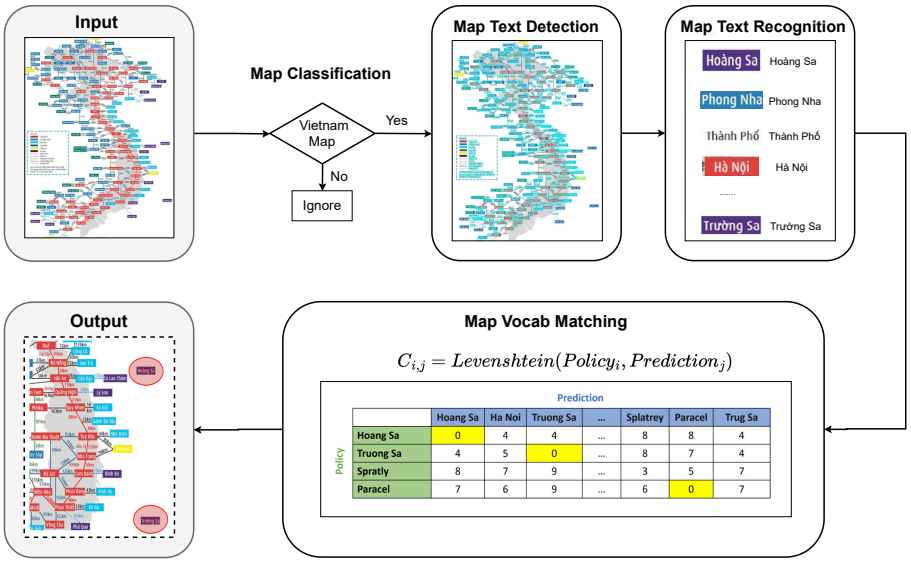

Fig. 4. The proposed pipeline for VinMap comprises four stages. Initially, the Map Classification module determines whether a given input image is a Vietnam map. If affirmative, the Map Text Detection module identifies all text regions within the map using computer vision techniques. Subsequently, the Map Text Recognition module scans these detected text regions and predicts the corresponding texts using an OCR model. Finally, the predicted texts are compared with a predefined policy using Levenshtein distance by the Map Vocab Matching to ascertain whether the input map image contains key regions.

TABLE IV: Ablation Study of the proposed pipeline

Classification	Pretrained ImageNet	x	x	
	Non-Map images included	x		x
F1-Score		85.51	62.34	55.62
Text Detection	Pretrained ICDAR2015	x	x	x
	Fine-tune VinText	x	x	
	Fine-tune VinMap box	x		
F1-Score		85.51	72.34	63.15
Vocab Matching	λ	5	2	1
F1-Score		63.44	85.51	77.23

in Vietnamese text images within the Map Text Recognition module. The detected text regions (see Fig. 4), cropped from map images, frequently display blurriness and noise owing to the low resolution of the maps.

We delve deeper into related tasks, presenting the recorded Average Precision (AP) scores outlined in Table III. The initial experiment entails utilizing our Map Classification model. In this experiment, the classification model is trained on a dataset comprising 1,400 non-map images and the remaining 3,401 map images, achieving an AP score of 99.6%. The last two experiments are based on the classification model and the final predictions derived from our method.

We explore various configurations of the proposed method, as outlined in Table IV. Each module is studied independently, with one module being investigated while the others remain at their default settings. Evaluation F1-Score is conducted on the overall pipeline's final prediction.

In the Map Classification module, we conduct ablations on the EfficientNet-B4 model by examining whether to utilize the pre-trained backbone on ImageNet1000 [16] and whether to include the Non-Map images set during training. Results indicate that utilizing the EfficientNet-B4 model pre-trained from ImageNet1000 significantly improves the understanding of map images, as evidenced by a performance drop to 55.62% F1-Score when not using it. Additionally, including 1,400 Non-Map images during the training process of the model pre-trained from ImageNet1000, boosts overall performance by 85.51 F1-Score, enabling the model's capability to differentiate between real-world and map images.

In the ablation of the Map Text Detection module, we study refining the model's performance across various datasets. Primarily, ensuring the model can effectively detect English text regions necessitates pretraining it on ICDAR2015 [7]. Moreover, enhancing the model's detection capabilities across both English and Vietnamese text entails fine-tuning the detection module on VinText, resulting in an improved overall F1-Score to 72.34 from only 63.15 when using only the pre-trained ICDAR20215. Additionally, to guide the detection model to focus on specific regions on the map, such as the two islands mentioned in this context, we further fine-tuned the model using the proposed VinMap box annotation set, leading to a notable surge in F1-Score to 85.51. This validates that our proposed annotation set on VinMap effectively directs the detection toward identifying regions of interest on map images.

We examine the impact of λ on the proposed Vocab Matching Algorithm. Decreasing λ entails a stricter adherence to matching the predicted text regions from our Map Text Recognition module with the specified policy terms (in our case, "Hoang Sa", "Truong Sa", "Spratly", "Paracel"). As

illustrated in Table IV, the highest F1-Score of 85.51 is achieved when $\lambda = 2$. However, adjusting $\lambda = 1$ inadvertently disregards near-perfect text predictions such as "Trung Sa", "Hoag Sa", "Spatly", "Parcl", etc., leading to misinterpretations of the map images. Conversely, relaxing $\lambda = 5$ significantly impacts performance with predicted regions like "Trung Son", "Ha Noi", etc.

V. Conclusion

Conclusively, we present a pioneering endeavor in geographic map comprehension through the introduction of the challenging dataset VinMap, comprising meticulously annotated map images. Furthermore, we establish a resilient method utilizing contemporary computer vision methodologies to scrutinize the dataset, thus laying the groundwork for forthcoming explorations in map analysis.

Acknowledgement. We sincerely thank the MIC-VN team for data crawling and labeling. We thank Mr. Que Nguyen and his team for their support in testing and deployment.

References

[1] Chandranath Adak. "Unsupervised text extraction from G-maps". In: *2013 International Conference on Human Computer Interactions (ICHCI)*. 2013.

[2] Samantha T Arundel, Wenwen Li, and Sizhe Wang. "GeoNat v1. 0: A dataset for natural feature mapping with artificial intelligence and supervised learning". In: *Transactions in GIS* 24.3 (2020), pages 556–572.

[3] Zhanzhan Cheng, Jing Lu, Baorui Zou, Shuigeng Zhou, and Fei Wu. *ICDAR 2021 Competition on Scene Video Text Spotting*. 2021. arXiv: 2107.11919 [cs.CV].

[4] Steffen Fritz and Linda See. "Comparison of land cover maps using fuzzy agreement". In: *International Journal of Geographical Information Science* 19.7 (2005), pages 787–807.

[5] Sachin Grover, Kushal Arora, and Suman K Mitra. "Text extraction from document images using edge information". In: *2009 Annual IEEE India Conference*. 2009.

[6] Rashad Al-Jawfi. "Handwriting Arabic character recognition LeNet using neural network." In: *Int. Arab J. Inf. Technol.* 6.3 (2009), pages 304–309.

[7] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. "ICDAR 2015 competition on robust reading". In: *2015 13th international conference on document analysis and recognition (ICDAR)*. 2015.

[8] Jialin Li and Ningchuan Xiao. "Computational cartographic recognition: Identifying maps, geographic regions, and projections from images using machine learning". In: *Annals of the American Association of Geographers* 113.5 (2023), pages 1243–1267.

[9] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. "Trocr: Transformer-based optical character recognition with pre-trained models". In: *Proc. AAAI*. 2023.

[10] Zhilin Li and Peizhi Huang. "Quantitative measures for spatial information of maps". In: *International Journal of Geographical Information Science* 16.7 (2002), pages 699–709.

[11] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. "Real-time scene text detection with differentiable binarization". In: *Proc. AAAI*. 2020.

[12] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. "Real-time scene text detection with differentiable binarization and adaptive scale fusion". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45.1 (2022), pages 919–931.

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s". In: *Proc. CVPR*. 2022.

[14] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. *MaskOCR: Text Recognition with Masked Encoder-Decoder Pretraining*. 2023. arXiv: 2206.00311 [cs.CV].

[15] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Triet Tran, Thanh Ngo, Thien Nguyen, and Minh Hoai. "Dictionary-guided Scene Text Recognition". In: *Proceedings of IEEE/CVF Computer Vision and Pattern Recognition*. 2021.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: 1409.0575 [cs.CV].

[17] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *Proc. ICML*. 2019.

[18] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Mike Zheng Shou, Umapada Pal, Dimosthenis Karatzas, and Xiang Bai. *ICDAR 2023 Video Text Reading Competition for Dense and Small Text*. 2023. arXiv: 2304.04376 [cs.CV].

[19] Xiran Zhou, Wenwen Li, Samantha T. Arundel, and Jun Liu. *Deep Convolutional Neural Networks for Map-Type Classification*. 2018. arXiv: 1805.10402 [stat.ML].

[20] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: 1911.02685 [cs.LG].