Tabular Context-aware Optical Character Recognition and Tabular Data Reconstruction for Historical Records

Loitongbam Gyanendro Singh, Stuart E. Middleton

School of Electronics and Computer Science, University of Southampton, United Kingdom.

Contributing authors: (gyanendro.loitongbam, sem03)@soton.ac.uk;

Abstract

Digitizing historical tabular records is essential for preserving and analyzing valuable data across various fields, but it presents challenges due to complex layouts, mixed text types, and degraded document quality. This paper introduces a comprehensive framework to address these issues through three key contributions. First, it presents `UoS_Data_Rescue`, a novel dataset of 1,113 historical logbooks with over 594,000 annotated text cells, designed to handle the complexities of handwritten entries, aging artifacts, and intricate layouts. Second, it proposes a novel context-aware text extraction approach (`TrOCR-ctx`) to reduce cascading errors during table digitization. Third, it proposes an enhanced end-to-end OCR pipeline that integrates `TrOCR-ctx` with ByT5 for real-time post-OCR correction, providing improved multilingual support. This pipeline reduces errors encountered in table digitization tasks by correcting OCR outputs in real time during training. The model achieves superior performance with a 0.049 word error rate and 0.035 character error rate, outperforming existing methods by up to 41% in OCR tasks and 10.74% in table reconstruction tasks. This framework offers a robust solution for large-scale digitization of tabular documents, extending its applications beyond climate records to other domains requiring structured document preservation. The dataset and implementation are available as open-source resources.*

Keywords: Optical Character Recognition, Tabular Structure Recognition, Semi-Supervised Learning, Historical Document Analysis, Data Annotation

* URI to Zenodo (dataset and model pre-trained checkpoints) and GitHub (source code) will be added if paper accepted

1 Introduction

Digitizing historical tabular records, including climate data, agricultural logs, and financial ledgers, is essential for advancing research across various fields. These records contain valuable long-term data that help researchers identify historical patterns and trends. However, many records exist in analog formats, typically stored as tables in logbooks, ledgers, and archival documents. Extracting structured information from these sources poses unique challenges, especially for conventional Optical Character Recognition (OCR) systems, which are mainly designed for continuous text. These systems often struggle with the complex layouts of tables, leading to inaccuracies in capturing spatial relationships among cells, rows, and columns. This can result in fragmented or misaligned data, significantly reducing the quality and usability of the digitized information. Additionally, the scarcity of annotated historical logbook images further complicates the development of robust models for such tasks.

Recent advancements in transfer learning have shown substantial promise in addressing these challenges. Transfer learning allows models trained on large datasets to adapt to new, specific tasks with smaller datasets, thereby leveraging existing knowledge and features. Pre-trained models such as AlexNet [1] and Inception [2] have been successfully fine-tuned for OCR tasks in scenarios like script recognition and historical document digitization [3, 4]. Transformer-based models such as TrOCR [5] have also demonstrated effective text recognition capabilities for handwritten entries, making them particularly suitable for digitizing historical climate records. Similarly, deep learning models like DETR [6] and CascadeTabNet [7] have been applied for table structure recognition, enabling more accurate detection of cells, rows, and columns in complex tables [8–10]. These methods suggest that combining OCR advancements with structured data recognition can significantly improve the digitization of complex tabular data, even in resource-constrained environments.

Despite the technological advancements in the OCR model, building an end-to-end system for digitizing historical tabular logbooks remains expensive and resource-intensive, making it impractical for widespread use. While few studies have focused on smaller documents like receipts and business cards [11, 12], the challenge escalates when dealing with logbooks that contain over 1,000 densely packed cells. Transfer learning presents a potential solution by utilizing pre-trained models, yet current digitization pipelines are vulnerable to cascading errors. The Table Structure Recognition (TSR) model identifies and segments table regions in these pipelines, while the OCR model extracts text from these cells [8–10]. Failures in the TSR or OCR stage can propagate through the pipeline, compounding errors and reducing overall performance. Efforts to mitigate OCR errors often involve post-processing steps [13, 14]. Such a composite model, which integrates TrOCR with a language model such as ByT5 [15] for post-processing, showcases significant adaptability for handling historical documents that often contain degraded text perturbations.

In this paper, we address the challenges of digitizing historical tabular data through three key contributions. Firstly, we introduce `UoS_Data_Rescue`, a novel dataset comprising 1,113 historical logbooks with over 594,000 annotated text cells, specifically designed to capture the complexities of historical tabular data, including handwritten entries, aging artifacts, and intricate layouts. This dataset covers various text types (typed, mixed, handwritten), table layouts, and time periods (1860s to 1980s), providing a valuable resource for OCR and table structure recognition research. Secondly, we address cascading errors in the digitization process by proposing an enhanced training strategy for the TrOCR model pipeline, named `TrOCR-ctx`. This approach utilizes contextual information from neighboring cells to enhance text extraction. By doing so, `TrOCR-ctx` significantly reduces extraction errors and minimizes cascading failures, improving the accuracy of table reconstruction tasks. Finally, we incorporate ByT5 as an end-to-end model for post-OCR correction within the pipeline, enhancing the recognition of diverse languages, archaic terminology, and complex character sets. This setup significantly improves transcription accuracy across various table layouts, providing robust digitization for historical documents while effectively handling visual text perturbations [14].

By incorporating context awareness and addressing cascading errors through transfer learning, our model, `TrOCR-ctx`, consistently outperforms baseline OCR systems across diverse datasets, effectively handling complex table structures and mixed text formats (refer to Section 5). The key findings highlight the importance of incorporating neighboring cell information to reduce cascading errors and accurately capture spatial relationships within tables. While primarily focused on climate records, this methodology is adaptable to various fields requiring structured document digitization, such as financial archives, medical records, and historical census data. The research not only offers a practical framework for large-scale digitization of tabular documents but also enhances the accessibility of valuable historical records across diverse domains, identifying areas for future improvement in handling multi-cell layouts and multi-line text entries. By sharing our code and model weights[1], we provide a practical framework for large-scale digitization efforts, enhancing the accessibility of valuable historical records and offering tools for researchers to advance data rescue initiatives across diverse fields.

The contributions of the paper are threefold:

i A novel dataset (`UoS_Data_Rescue`) containing 1,113 historical logbooks with over 594,000 annotated text cells, covering various text types, table layouts, and time periods from the 1860s to the 1980s, offering a valuable resource for OCR and table structure recognition research.

ii A novel fine-tuning approach (`TrOCR-ctx`) that utilizes contextual information from neighboring cells, significantly reducing cascading failures and thereby enhancing the accuracy of table reconstruction tasks.

iii We incorporate ByT5 as an end-to-end model for post-OCR correction within the pipeline, enhancing the recognition of diverse languages, archaic terminology, and complex character sets. This approach significantly improves transcription accuracy and robustness for historical document digitization.

[1] URL placeholder

2 Related studies

The digitization of tabular documents from images has evolved significantly from traditional rule-based methods to advanced deep-learning models. Early approaches relied on predefined heuristics to identify tables based on visual layout features, effectively handling structured formats but struggling with irregular or complex layouts. As document diversity increased, the limitations of these rule-based systems became apparent, leading to adopting more adaptable machine-learning techniques. This review outlines the progression of techniques in this domain, highlighting key approaches and models that address the challenges of diverse document formats and the capabilities of OCR systems.

2.1 Rule-Based Approaches

Optical Character Recognition (OCR) has been a foundational technology in digitizing tabular documents. Early approaches to table detection and extraction primarily relied on rule-based systems, utilizing predefined heuristics to identify tables based on visual layout features such as grid lines, alignment, and consistent spacing [16, 17]. These methods were effective for structured tables with regular formats, leveraging techniques like grid line detection, pattern recognition, and bounding box analysis in controlled scenarios. However, they often struggled with irregular or complex layouts and were inflexible when confronted with diverse or unstructured data.

While rule-based systems offer advantages in interpretability and precision for consistent formats, they are constrained by the complexity of rule creation and their inability to adapt to varying table structures. As the diversity of documents increased, the limitations of these systems became more pronounced, necessitating the adoption of more flexible machine learning (ML) techniques. These advanced approaches provide improved scalability and robustness for extracting tabular data from complex or unstructured documents, thereby enhancing the efficacy of OCR technologies in contemporary applications [18].

2.2 Machine Learning Approaches

Machine learning techniques have been introduced to overcome the limitations of rule-based systems for table extraction, offering greater adaptability and precision. By combining OCR with statistical models, these methods automate detection and recognition, enhancing the ability to handle diverse table types through accurate whitespace identification and data extraction. Supervised learning techniques, including Convolutional Neural Networks (CNNs) [19–21] and Support Vector Machines (SVMs) [22], have significantly improved the identification of tables within complex layouts, with CNNs particularly adept at recognizing spatial structures in images.

The advent of deep learning has marked a significant leap forward in table extraction capabilities. End-to-end models like TableNet [23] and TC-OCR [9] exemplify this progress by integrating table detection and structure recognition into unified frameworks. TableNet enhances efficiency and accuracy by treating these tasks as interdependent sub-problems within a single neural network, while TC-OCR improves table recognition by combining state-of-the-art models such as DETR [6], CascadeTabNet [7, 8, 24], and PP-OCR v2 [25]. This approach effectively addresses variations in table styles and image distortions, facilitating simultaneous table detection, structure recognition, and content extraction. Transformer-based models, including DeepDeSRT [26] and TableFormer [27], further enhance extraction capabilities. DeepDeSRT employs a pre-trained ResNet-18 backbone to generate structured representations of tables with high accuracy, while TableFormer predicts bounding boxes for individual cells, facilitating precise content extraction from PDF documents. The integration of transfer learning allows these architectures to recognize printed and handwritten text, making them particularly suitable for digitizing historical documents with diverse writing styles.

Recent studies on post-OCR correction have increasingly leveraged transformer-based models to enhance accuracy across various domains [13, 14, 28, 29]. A common trend is using TrOCR as the base OCR model, with researchers exploring different post-processing approaches. Chen et al. [13] integrate TrOCR with CharBERT [30], improving accuracy and reducing over-correction, particularly for historical documents. Seth et al. [14] pair TrOCR with ByT5, a byte-level transformer model, focusing on visual text perturbations and introducing the LEGIT dataset to study legibility. Rakshit et al. [28] present a comprehensive pipeline combining OCR (including TrOCR) with transformer-based NLP tools like ByT5 [15] and BART [31], refining outputs for printed and handwritten text. Karthikeyan et al. [29] use Roberta [32] to post-correct medical reports, leveraging masked word prediction to handle domain-specific terminology. These studies highlight the growing trend of using transformer-based models for post-OCR correction, demonstrating their ability to capture visual and linguistic information. Notably, ByT5's success in handling perturbed text scenes suggests it could be effective for improving OCR accuracy for historical document extraction in such challenging contexts, where degraded or unfamiliar characters often appear.

Despite these advancements, challenges remain in handling diverse document formats, densely packed cells, nested cells, and noisy images. These complexities often necessitate considerable computational resources and specific fine-tuning for different datasets, limiting existing solutions' practical application and scalability. To address these challenges, our work focuses on developing a specialized dataset focusing on historical climate logbook images and implementing transfer learning strategies, particularly fine-tuning the TrOCR model to navigate the intricacies of historical climate logbooks. By enhancing the context-awareness of TrOCR and integrating it with ByT5 for improved multilingual support, we aim to create more resilient pipelines for table extraction, paving the way for broader adoption in resource-constrained environments and other fields that require structured document extraction.

2.3 Datasets for Tabular Data Extraction

Several datasets have been developed to support research in Optical Character Recognition (OCR) and Tabular Data Extraction, each addressing different document types and challenges. While benchmarks such as PubTabNet, CORD, SROIE, and LayoutLM-based datasets primarily focus on modern documents, datasets specifically designed for historical tabular data remain scarce.

PubTabNet, developed by IBM Research Australia, consists of scientific tables extracted from academic publications, annotated with HTML representations for ground-truth validation [33]. While useful for OCR and table extraction tasks, it primarily focuses on structured, typed text, making it less suitable for historical documents that often contain handwritten entries, aging artifacts, and irregular layouts. Datasets such as CORD (Consolidated Receipt Dataset) [34] and SROIE (Scanned Receipt OCR and Information Extraction) [35] focus on receipt documents with relatively simple layouts and limited structural variability. CORD provides multilingual named entity annotations, whereas SROIE consists mostly of English-language receipts. While these datasets are useful for evaluating OCR models on structured, modern documents, they do not address the complexities of historical tabular data, which often involve irregular layouts, handwritten text, and document degradation.

LayoutLM-based [20] datasets leverage multimodal learning by incorporating both textual content and spatial layout information, enabling models to better understand document structures. These datasets, often derived from existing OCR benchmarks, are primarily used for pre-training and fine-tuning LayoutLM models on tasks such as key information extraction, entity recognition, and document classification. They are particularly effective for modern documents with well-defined layouts, such as invoices, forms, and reports. However, they are not specifically designed for historical table extraction, as they lack handwritten text variations, irregular table structures, and document degradation, which are common in archival records.

To address the gap in historical tabular datasets, we introduce `UoS_Data_Rescue`, a large-scale collection of 1,113 historical logbooks spanning diverse text types (typed, mixed, handwritten), intricate table structures, and aging artifacts across different periods (1860s–1980s). Unlike modern datasets such as PubTabNet and LayoutLM-based datasets, which focus on structured, printed documents, or receipt-based datasets like CORD and SROIE, which contain relatively simple layouts, `UoS_Data_Rescue` explicitly captures the unique challenges of historical documents. The dataset features dense, compact tabular images with tightly packed handwritten and printed text, reflecting the formatting constraints of archival records. By preserving both table structures and diverse text content, this dataset enables a more rigorous evaluation of OCR models, particularly in handling handwritten text, degraded documents, and complex archival layouts.

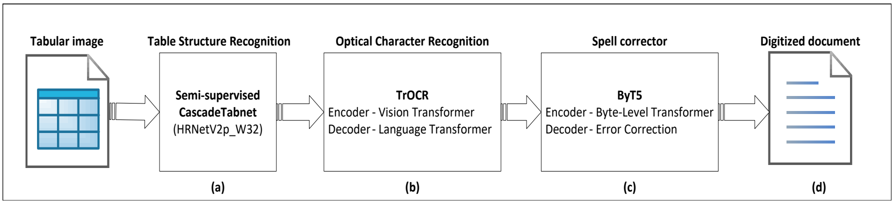


Fig. 1: Block Diagram of the Tabular Data Extraction Pipeline. (a) Table and cell regions are detected using a semi-supervised Table Structure Recognition (TSR) model. (b) Each identified table cell and neighboring cell image is processed by TrOCR's encoder-decoder architecture to generate OCR output text. (c) The OCR output is refined using ByT5, a byte-level transformer model, for post-OCR correction, ensuring accurate text extraction. (d) Finally, the tabular document is reconstructed based on spatial information from TSR and digitized text from OCR and spell correction, enabling precise tabular data representation.

3 Research methodologies

This section outlines the dataset and methodologies used to digitize historical tabular records, which are essential for preserving valuable data. These records often present challenges due to densely packed cells, handwritten entries, and complex layouts. To address these issues, we implement a systematic approach that integrates transfer learning for model fine-tuning and develops a robust tabular data reconstruction pipeline consisting of three components: (i) Table Structure Recognition (TSR), (ii) a customized tabular context-aware OCR model based on TrOCR, and (iii) a reconstruction module. Figure 1 presents the tabular data reconstruction pipeline. This integrated pipeline improves text extraction from noisy, aged records, enabling more effective digitization of tabular data.

3.1 `UoS_Data_Rescue` dataset

The dataset used in this study, `UoS_Data_Rescue`, consists of 1,113 scanned historical climate logbook images and includes over 594,000 human annotations for cell boundaries and transcribed text. This dataset significantly contributes to OCR and table structure recognition, as it captures various text types (typed, mixed, handwritten) and table layouts and spans from the 1860s to the 1980s. To ensure that the dataset reflects the complexities of historical documents, we employed a maximum variance sampling strategy, selecting logbook regions that maximize variance in document format, handwriting styles, and time periods. This approach resulted in broad coverage across low-density regions worldwide and high-density coverage in specific regions like Africa, aligning with the needs of climate scientists interested in these areas. Table 1 provides an overview of the distribution of annotated and unlabeled logbook images, categorized by year, region, and source. The source documents come from prestigious institutions

Location	Year	# Labelled images	Average cells/image	Average cells/image (hard-to-annotate)	# Unlabelled images
Sources: https://digital.nmla.metoffice.gov.uk/*					
UK✓	1830-1930	97	208.959	3.804	–
Natal, Africa	1870	26	99.429	0.048	46
Artics✓	1880	82	477.122	30.220	–
Devon, UK✓	1890-1940	33	229.545	1.758	–
Ben Nevis, UK	1890	97	1511.247	11.557	–
UK and World✓	1900	93	622.793	31.141	1330
Philippines	1900	24	740.292	5.458	6077
India (NOAA)	1930	24	2197.429	7.476	380
India (MO)	1970	24	1971.667	17.208	276
Sources: https://digital.nmla.metoffice.gov.uk/					
Zanzibar✓	1881-1882	8	133.500	9.250	12
Blantyre✓	1882	–	–	–	6
Egypt✓	1885-1886	6	699.500	17.667	9
Morocco✓	1891	–	–	–	2
Sources: https://libguides.library.noaa.gov/weather-climate/foreign-climate					
Mauritius	1862-1972	34	227.559	0.235	13887
Algeria✓	1877-1968	113	90.947	7.292	22356
Madagascar	1889-1968	40	148.775	2.925	10035
Egypt	1900-1966	51	148.137	6.412	44199
Tunisia✓	1907-1932	29	119.207	4.034	2531
Uganda	1909-1937	5	188.000	0.000	456
Mozambique	1909-1968	44	289.364	2.909	19547
South Africa✓	1920-1982	74	109.946	0.757	36502
Libya	1922-1931	3	221.000	12.000	501
Kenya	1936-1937	–	–	–	31
Angola	1937-1952	–	–	–	1840
Namibia	1941-1948	–	–	–	52
Djibouti	1950-1974	–	–	–	1695
Cameroon	1950-1975	–	–	–	1830
Morocco	1954-1978	58	223.914	9.724	10575
Guinea-Bissau	1957-1972	–	–	–	3331
Sources: https://catalog.archives.gov					
Bear	1940	12	792.583	1.917	21
Tennessee	1946	36	880.611	1.472	17
Sources: http://archives-climat.fr					
Ambanja Août-décembre✓	1904	5	559.200	95.800	–
Diego-Suarez✓	1949	47	640.447	30.255	1
Tromelin✓	1956	48	651.396	33.021	–
Total	–	1113	–	–	177545

*Original images sourced with permission from UK Met Office (MO), US NOAA and weatheerrescue.org (University of Reading) for the https://glosat.org/ project.
✓ The logbooks contain a mix of handwritten and typed text.

Table 1: Overview of the `UoS_Data_Rescue` dataset, including the distribution of annotated and unlabeled logbook images across 34 regions.

such as the UK Met Office[2], US NOAA[3], US naval ship logs[4], and Meteo-France[5]. The dataset also captures challenges like handwritten entries, faded artifacts, and complex layouts. To ensure high-quality annotations, we used a crowdsourced approach via the Appen platform[6], where annotators followed strict guidelines (See Appendix B) to simplify transcription and minimize errors.

[2] https://digital.nmla.metoffice.gov.uk
[3] https://libguides.library.noaa.gov
[4] https://catalog.archives.gov
[5] http://archives-climat.fr
[6] https://www.appen.com

Dataset	Table Structure Recognition		Optical Character Recognition		Average cells per image in test set
	#Training Images	#Testing Images	#Train text lines	#Test text lines	
UoS_Data_Rescue	1113	112	497045	97150	867.41
SROIE	1426	273	33626	18704	68.51
CORD	800	100	19367	2355	23.55
PubTabNet	6000*	15115	26000+	606719	40.14
ICDAR15	–	–	4468	2077	–

* The original PubTabNet dataset was released with 510K training samples.
* Randomly selected 26000 text lines from the 6000 training samples.

Table 2: Distribution of training and testing data for fine-tuning TSR and OCR models, highlighting the unique characteristics of each dataset.

To detect table structures and define cell boundaries within each scanned image to support an efficient and accurate annotation process, we used a pre-trained semi-supervised TSR model based on CascadeTabNet [8]. This model automatically identified table layouts and segmented text into individual cells, allowing annotators to focus on transcribing a manageable number of cells ($k = 40$) per image. When the TSR model produced incorrect predictions, annotators manually corrected the errors to ensure high-quality data. For images containing more than k cells, additional copies were created, each containing no more than k cells, making the annotation process more precise and manageable.

We employed six annotators, all proficient in English and familiar with Latin alphabetical characters. The annotation process was conducted in 18 batches, each containing approximately 1,500 images (40 cells or fewer per image). After each batch, a quality evaluation was performed to ensure accuracy. Incorrectly annotated batches were re-run after evaluation. Recognizing that crowdsourced annotators may lack expertise in transcribing historical handwriting styles, we instructed them to flag hard-to-annotate cells for correction by the domain experts, ensuring high-quality data for developing a reliable tabular data reconstruction model. This rigorous annotation process, combined with the dataset's diverse geographic coverage across 34 regions, ensures that the dataset supports the development of models capable of generalizing across various logbook types and formats, contributing to the digitization of historical climate records.

3.1.1 Dataset Characteristics for OCR Model Training

To comprehensively evaluate the robustness of the OCR models, we utilized a diverse range of datasets that represent various text formats and layouts. Notably, we employed the in-house curated dataset, `UoS_Data_Rescue`, which stands out due to its exceptionally dense tables—approximately 10 times denser than those found in other datasets considered for evaluation. This density poses a unique challenge and fills a critical gap in existing datasets. Additionally, we employed the ICDAR 2015 Scene Text Recognition dataset [36, 37], and tabular-structured datasets such as CORD [19, 34, 38], SROIE [12, 20, 35], and PubTabNet [33, 39]. The ICDAR 2015 dataset serves as a benchmark for scene text detection and recognition, featuring images with text embedded in natural environments. The CORD and SROIE datasets focus on receipt images, primarily in English, and were used to fine-tune and assess the model's performance in recognizing and extracting text from diverse receipt layouts. The PubTabNet dataset, consisting of scientific tables, was incorporated to fine-tune and test the model's ability to manage complex tabular structures. Table 2 provides an overview of the data distribution for training and testing sets used to fine-tune the OCR model. These datasets collectively provide a robust set of scenarios for fine-tuning and evaluating OCR and tabular data reconstruction (TDR) tasks, ensuring the models are thoroughly tested on various text formats and layouts.

3.2 Table Structure Recognition

To improve the accuracy of table structure recognition in historical climate logbooks, we fine-tuned a pre-trained Table Structure Recognition (TSR) model based on CascadeTabNet [8] using the annotated `UoS_Data_Rescue` dataset. The CascadeTabNet model employs a Cascade Mask R-CNN architecture [40] with a High-Resolution Network (HRNetV2p_W32) backbone [41], which extracts multi-scale features from document images and refines table detection through multiple stages. These stages predict the presence of tables and the precise boundaries of individual cells, making the model particularly effective for handling complex layouts and noisy, degraded images. The hyperparameters used to train CascadeTabNet are detailed in Appendix Table 8.

Since the trained model generates a limited number of table cells (up to 2000, including both positive and negative predictions), we implemented a method to infer missing cells based on the horizontal and vertical alignment of detected positive cells. This approach ensures complete table reconstruction by generating candidate cells where gaps are identified and aligning them with existing cells. This method is crucial in preprocessing data before OCR, as it accurately identifies table regions and defines cell boundaries. This robust table structure recognition provides a solid foundation for the downstream OCR model to perform precise text extraction. Ultimately, this improves the accuracy and reliability of tabular data reconstruction, supporting the successful digitization of historical climate records.

3.3 Tabular Context-aware Optical Character Recognition

Building on the robust table structure recognition provided by the fine-tuned TSR model, our methodology adapts the TrOCR model specifically for extracting tabular data. Originally designed for continuous text recognition, TrOCR employs a Transformer-based architecture, utilizing a Vision Transformer (ViT) encoder to process images into visual embeddings, along with an autoregressive text decoder that generates text from these embeddings. For a comprehensive understanding of TrOCR's architecture, readers are encouraged to refer to the original paper by Li et al. [5].

In this study, we enhanced TrOCR by incorporating context-awareness of neighboring table cells to improve its accuracy in digitizing historical tabular documents. Typically, TrOCR is fine-tuned on individual table cells or text line images. However, this approach can struggle with densely packed cells or handwritten text that crosses cell boundaries, leading to misalignment even when the TSR model accurately detects table layouts. To address this, we introduced a fine-tuning strategy that includes information from neighboring cells during training. Specifically, two additional images were generated for each table cell: one including the neighboring cell to the right and another including the cell below. The texts from neighboring cells were separated by boundary identifiers token `[SEP]`, enriching the training dataset with contextual information and improving the model's ability to handle irregular layouts and merged cells. We refer to this context-aware fine-tuning of TrOCR as `TrOCR-ctx`.

During digitization, each detected table cell is expanded into two configurations: one with the neighboring cell to the right and another with the one below. The common text before `[SEP]` is extracted as the final output for the target cell. This approach significantly reduces cascading errors caused by isolated text lines or ambiguous boundaries by leveraging contextual cues during text extraction. By incorporating neighboring cell information, `TrOCR-ctx` develops a more comprehensive understanding of adjacent cells, leading to improved accuracy in recognizing text from challenging tabular configurations commonly found in historical documents.

3.4 Post-processing OCR using a ByT5 Model

TrOCR, while effective for many OCR tasks, struggles with multilingual text in historical documents due to irregular fonts, inconsistent spacing, and image degradation [13, 14]. These challenges often lead to tokenization errors or misrecognition of characters, particularly in archaic languages and non-standard character sets typical of historical records. To address this, we integrate ByT5, a byte-level Transformer model known for handling perturbed text, into the pipeline for post-OCR correction [14]. ByT5 processes text at the byte level, bypassing traditional tokenization, which allows it to handle diverse languages, archaic terminology, and complex character sets more effectively.

In our pipeline, the output of TrOCR (`TrOCR-ctx`) is fed into ByT5, which corrects recognition errors at the byte level. This enables ByT5 to refine text with non-standard characters and spelling variations, significantly improving transcription accuracy across various table layouts. For instance, as shown in Figure 2, ByT5 corrects `TrOCR-ctx` output by accurately transforming complex historical text such as "Température," "Méchéria," "Géryville," and "11.2" into their correct digital forms, handling archaic characters and diacritical marks with precision. This byte-level approach significantly enhances the accuracy of digitizing complex multilingual historical documents, making ByT5 particularly well-suited for challenging OCR tasks involving nuanced text recognition and correction.

`TrOCR`:	(a) Temperature.	(b) Moeckiria.	(c) Geryville.	(d) 11, v.
`TrOCR+ByT5`:	(a) Température	(b) Méchéria	(c) Géryville	(d) 11.2

Fig. 2: Examples of ByT5 post-OCR corrections on `TrOCR` outputs for historical text images, accurately recognizing complex characters in examples like 'Température,' 'Méchéria,' 'Géryville,' and numerical data '11.2.'

Algorithm 1 Tabular Data Reconstruction

Input: TSR and `TrOCR-ctx` outputs
Output: Reconstructed table with preserved spatial and contextual relationships
Create horizontal and vertical centroid lists
for each cell in the table structure **do**
    Retrieve text output from `TrOCR-ctx`
    Compute centroid for the current cell
    if centroid lists are empty **then**
        Add centroid to both horizontal and vertical centroid lists
    else
        Compute distance to the last centroid in the lists
        if distance > k **then**
            Add centroid to lists
        end if
    end if
    Align text to centroid index horizontally and vertically
end for
Initialize table layout using horizontal and vertical centroid lists
Reconstruct table by aligning text using centroid lists
Return: Reconstructed table where each cell is separated by TAB space.

3.5 Tabular Data Reconstruction Module

After extracting text from individual cells, the final step is reconstructing the digitized text to the original tabular format to preserve spatial and contextual relationships. Algorithm 1 outlines the reconstruction process. This step is essential for maintaining the integrity of the digitized data. The reconstruction module combines outputs from the TSR and fine-tuned `TrOCR-ctx` models to accurately recreate the table layout, ensuring alignment with the original structure. This alignment improves the usability and accuracy of the digitized data, making it more valuable for research and analysis. Ultimately, the module enhances the fidelity of the digitization process, preserving historical data in its true form.

4 Experimental Setup

To thoroughly assess the robustness and effectiveness of our tabular data reconstruction pipeline, we evaluate and compare the performance of `TrOCR-ctx` model with three fine-tuned OCR models, namely TrOCR [5], Abinet [36], and PP-OCRv2 (PaddleOCR) [25], across a diverse set of datasets. This approach comprehensively evaluates each model's ability to generalize across varied document types, focusing on the fine-tuned `TrOCR-ctx` model's adaptability and effectiveness in real-world applications.

4.1 Evaluation metrics

To comprehensively evaluate the performance of the OCR models and the overall TDR pipeline, we employed a range of evaluation metrics tailored to both TSR and OCR tasks. These metrics offer insights into the accuracy, precision, and robustness of the models across various aspects of table structure detection, text extraction, and reconstruction.

4.1.1 Evaluation Metrics for Table Structure Recognition

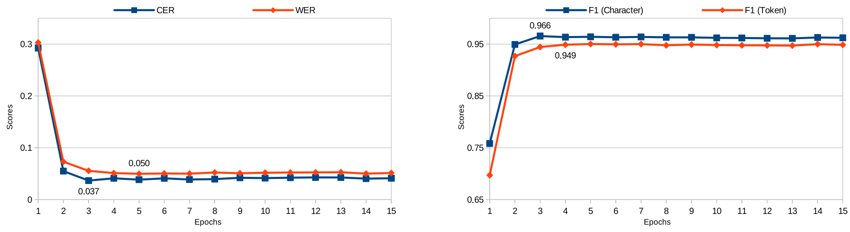
For TSR, the evaluation focuses on how accurately the model detects table structures, including cell boundaries and overall layout. In this study, we use the Weighted Average F1 (wF1) score as the primary metric [42]:

$$\mathrm{wF1} = \sum_i w_i \cdot \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{1}$$

Here, w_i represents the weight for each Intersection over Union (IoU) threshold i, and $Precision_i$ and $Recall_i$ are the precision and recall at the i^{th} IoU threshold. The IoU thresholds are set to 0.6, 0.7, 0.8, and 0.9. A prediction is considered correct if it meets or exceeds these thresholds, ensuring that precision and recall are balanced across varying levels of overlap between predicted and actual table structures.

4.1.2 Evaluation Metrics for Optical Character Recognition

To evaluate the robustness of the OCR models, we used a diverse set of metrics tailored to assess different aspects of performance, particularly in handling tabular structured documents. These metrics include ROUGE-L [43], Word Error Rate (WER) [44], Character Error Rate (CER) [44], Exact Match (EM) [45], and F1-scores at both character and token levels [45]. Each metric was selected for its ability to provide unique insights into OCR model performance. ROUGE-L measures sequence-level accuracy by comparing the longest common subsequence between predicted text and ground truth, making it useful for evaluating longer text sequences. WER and CER are standard OCR evaluation metrics that quantify word and character level errors, respectively, offering a granular view of text recognition accuracy. Exact Match (EM) provides a



(a) WER and CER (b) F1-score (Character and Token level)

Fig. 3: Performance evaluation of `TrOCR-ctx` per epoch, showing the progression of WER, CER, and F1-scores at both character and token levels. The labeled epoch indicates the peak performance among the 15 epochs, with the model converging after approximately 3 epochs.

strict evaluation by checking if the predicted text exactly matches the ground truth, which is critical for assessing perfect OCR output. Finally, F1-scores at both character and token levels balance precision and recall, capturing partial matches where minor errors occur. The F1-score is calculated as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision $= \frac{\mathtt{M}}{\mathtt{P}}$ and Recall $= \frac{\mathtt{M}}{\mathtt{G}}$. This formula applies to both evaluation levels. For the Character-Level F1-Score, `M`, `P`, and `G` represent matched, predicted, and ground truth characters, respectively. For the Token-Level F1-Score, these variables denote matched, predicted, and ground truth tokens. These metrics comprehensively evaluate the OCR model's robustness in handling complex historical tabular data.

4.1.3 Overall Performance Metrics

By integrating evaluation metrics for both TSR and OCR, we obtain a comprehensive assessment of the entire tabular data reconstruction pipeline. The weighted F1-score allows us to evaluate the accuracy of table structure detection, including cell boundaries and overall layout. Meanwhile, OCR metrics like Rouge-L, Word Error Rate (WER), Character Error Rate (CER), and F1-scores at both character and token levels provide detailed insights into text extraction accuracy. These metrics capture exact matches and account for partial correctness, which is crucial for handling complex historical documents often plagued by noisy or degraded data. This combination ensures a holistic evaluation of the digitization process, reflecting the high-level performance of table detection and fine-grained text recognition accuracy. Ultimately, this multi-faceted evaluation approach allows us to pinpoint specific areas for improvement while preserving the integrity of tabular documents.

4.2 Experiment Runtime and Hardware Specifications

The experiments were conducted on a multi-GPU system with two NVIDIA A100 GPUs, each with 80 GB of memory. This setup was essential to handle the extensive datasets and complex computations required for training `TrOCR-ctx` (refer to Table 2). Training `TrOCR-ctx` for 15 epochs took approximately two weeks, reflecting the computational demands of incorporating context-aware text extraction. In comparison, the baseline TrOCR model completed 15 epochs in about one week, demonstrating that introducing additional context-aware samples in `TrOCR-ctx` requires more processing time but improves performance across various datasets. Figure 3 illustrates the performance of `TrOCR-ctx` per epoch, showing improvements in word error rate (WER), character error rate (CER), and F1-scores at both character and token levels. Notably, the model converges approximately after the third epoch, indicating efficient learning despite the computational intensity.

5 Results and Discussion

5.1 OCR Performance Analysis

Table 3 presents the evaluation of OCR models—`TrOCR-ctx`, TrOCR, Abinet, and PP-OCRv2—across diverse datasets including `UoS_Data_Rescue`, ICDAR 2015, CORD, SROIE, and PubTabNet. These datasets were used to assess the robustness and accuracy in handling various text formats and layouts, with the evaluation conducted on properly segmented text lines from the test sets (refer Table 2). `TrOCR-ctx`, fine-tuned with context-aware patches, consistently achieved the highest performance across all datasets. Its superior F1-scores, ranging from 0.755 to 0.986, highlight its ability to accurately recognize text in challenging scenarios, such as mixed handwritten and typed text or complex table structures. This model excelled in handling historical tabular data, achieving F1-scores of 0.951 for `UoS_Data_Rescue`, 0.986 for CORD, 0.967 for SROIE, 0.909 for PubTabNet, and 0.755 for ICDAR 2015. Additionally, it demonstrated low CER, ranging from 0.014 (SROIE) to 0.102 (ICDAR 2015), and high Rouge-L scores from 0.776 (ICDAR 2015) to 0.957 (CORD). Among the other OCR models, TrOCR performed better than Abinet and PP-OCRv2 but lagged behind `TrOCR-ctx`. TrOCR achieved F1-scores of 0.945 on `UoS_Data_Rescue`, 0.834 on CORD, 0.947 on SROIE, 0.859 on PubTabNet, and 0.750 on ICDAR 2015. Its CER ranged from 0.044 to 0.102, and its Rouge-L scores from 0.777 to 0.919, indicating solid performance but with room for improvement compared to `TrOCR-ctx`. The lower performance of Abinet and PP-OCRv2 on datasets with complex historical data emphasizes context awareness, underscoring the importance of robust models like `TrOCR-ctx` for such tasks.

OCR Model	Rouge-L	WER	CER	EM	F1-score (Char)	F1-score (Token)
			UoS_Data_Rescue			
TrOCR	0.849	0.055	0.047	0.825	0.963	0.945
TrOCR-ctx	**0.857**	**0.049**	**0.035**	**0.847**	**0.966**	**0.951**
Abinet	0.545	0.557	0.346	0.432	0.681	0.449
PP-OCRv2	0.812	0.348	0.178	0.646	0.825	0.666
			CORD			
TrOCR	0.898	0.168	0.056	0.802	0.946	0.834
TrOCR-ctx	**0.957**	**0.034**	**0.016**	**0.833**	**0.985**	**0.986**
Abinet	0.520	0.356	0.304	0.574	0.710	0.644
PP-OCRv2	0.789	0.114	0.144	0.746	0.897	0.886
			SROIE			
TrOCR	0.919	0.053	0.044	0.830	0.984	0.947
TrOCR-ctx	**0.940**	**0.033**	**0.014**	**0.849**	**0.988**	**0.967**
Abinet	0.872	0.629	0.497	0.301	0.507	0.381
PP-OCRv2	0.882	0.432	0.235	0.493	0.777	0.577
			PubTabNet			
TrOCR	0.878	0.141	0.069	0.748	0.940	0.859
TrOCR-ctx	**0.913**	**0.091**	**0.067**	**0.789**	**0.965**	**0.909**
Abinet	0.315	0.813	0.450	0.153	0.598	0.197
PP-OCRv2	0.833	0.131	0.097	0.700	0.915	0.877
			ICDAR 2015			
TrOCR	**0.777**	**0.245**	0.102	0.744	0.904	0.750
TrOCR-ctx	0.776	0.250	0.102	**0.749**	**0.905**	**0.755**
Abinet	0.151	0.741	0.624	0.259	0.436	0.259
PP-OCRv2	0.665	0.374	0.178	0.626	0.831	0.627

Table 3: Performance comparison of TrOCR-ctx, TrOCR, Abinet, and PP-OCRv2 models on UoS_Data_Rescue, CORD, SROIE, PubTabNet, and ICDAR 2015 datasets using segmented text lines. Evaluation metrics include Rouge-L, Word Error Rate (WER), Character Error Rate (CER), Exact Match, and F1-scores at both character and token levels.

5.2 Tabular Data Reconstruction Performance

Following the superior OCR performance of TrOCR-ctx, we conducted a detailed evaluation of its tabular data reconstruction capabilities in multiple datasets, including UoS_Data_Rescue, CORD, SROIE, and PubTabNet. The results presented in Table 4, demonstrate a clear performance advantage of TrOCR-ctx over the baseline TrOCR model, primarily due to its context-aware fine-tuning. By incorporating contextual information during training, TrOCR-ctx consistently outperformed the baseline in all data sets. First, we evaluated the performance of TrOCR-ctx without post-correction from ByT5. The inclusion of contextual information in text extraction significantly improved performance compared to the non-contextual TrOCR model. Specifically, TrOCR-ctx achieved 0.61% and 3.20% improvement in the F1 scores at the character level and token level, respectively, on UoS_Data_Rescue. Similarly, it outperformed TrOCR in CORD by 2.58% and 3.71%, in SROIE by 4.60%

	Table structure recognition			Tabular data reconstruction					
Dataset	P	R	wF1	Rouge-L	WER	CER	EM	F1 (Char)	F1 (Token)
Without contextual information contextual information (TrOCR)									
`UoS_Data_Rescue`	0.742	0.919	0.805	0.771	0.281	0.254	0.719	0.819	0.719
CORD	0.970	0.715	0.798	0.890	0.043	0.031	0.863	0.890	0.863
SROIE	0.805	0.796	0.785	0.847	0.046	0.039	0.819	0.869	0.819
PubTabNet	0.959	0.814	0.869	0.618	0.584	0.593	0.408	0.525	0.408
With contextual information contextual information (`TrOCR-ctx` without ByT5 model)									
`UoS_Data_Rescue`	0.742	0.919	0.805	0.778	0.258	0.232	0.742	0.824 (Δ0.61%)	0.742 (Δ3.20%)
CORD	0.970	0.715	0.798	0.917	0.035	0.023	0.895	0.913 (Δ2.58%)	0.895 (Δ3.71%)
SROIE	0.805	0.796	0.785	0.872	0.025	0.023	0.875	0.909 (Δ4.60%)	0.875 (Δ6.84%)
PubTabNet	0.959	0.814	0.869	0.636	0.584	0.593	0.416	0.527 (Δ0.38%)	0.416 (Δ1.96%)
With contextual information contextual information (`TrOCR-ctx`)									
`UoS_Data_Rescue`	0.742	0.919	0.805	0.809	0.245	0.213	0.755	0.850 (Δ3.79%)	0.755 (Δ5.01%)
CORD	0.970	0.715	0.798	0.917	0.023	0.025	0.914	0.921 (Δ3.48%)	0.914 (Δ5.91%)
SROIE	0.805	0.796	0.785	0.908	0.023	0.022	0.907	0.914 (Δ5.18%)	0.907 (Δ10.74%)
PubTabNet	0.959	0.814	0.869	0.640	0.592	0.594	0.426	0.536 (Δ2.10%)	0.426 (Δ4.41%)

Table 4: Performance evaluation of `TrOCR-ctx` model on Tabular Data Reconstruction across `UoS_Data_Rescue`, CORD, SROIE, and PubTabNet datasets. Precision and Recall for table structure recognition are calculated based on an IoU threshold ≥ 0.6.

and 6.84%, and in PubTabNet by 0.38% and 1.96%, at the character level and the token level, respectively. Next, we evaluated the impact of adding a post-OCR correction using ByT5 to further refine the output of TrOCR-ctx. This additional step resulted in significant performance improvements in all datasets. Specifically, with ByT5 post-correction applied, `TrOCR-ctx` achieved a 3.79% and 5.01% improvement in F1-scores at the character-level and token-level on `UoS_Data_Rescue`, respectively. Similarly, it outperformed TrOCR on CORD by 3.48% and 5.91%, on SROIE by 5.18% and 10.74%, and on PubTabNet by 2.1% and 4.41%, at character-level and token-level. These results highlight the effectiveness of context-sensitive fine-tuning in `TrOCR-ctx` to improve OCR accuracy and demonstrate the additional benefits of integrating ByT5 for post-OCR correction in handling complex tabular data extraction tasks.

However, both OCR performances on the PubTabNet dataset were notably lower than others despite having good performance on properly segmented text lines. This lower performance can be attributed to two main factors. First, the Table Structure Recognition (TSR) model struggled with accurately aligning table layouts, even when the Intersection over Union (IoU) score exceeded 0.6. This misalignment significantly impacted the overall performance of the OCR system. For context, we used a randomly selected subset of 6,000 images for training and evaluated the model on a test set of 15,115 images from PubTabNet. Second, TrOCR faced difficulties processing longer multi-line text entries, which were abundant in PubTabNet. Similar issues were observed in some logbooks within the `UoS_Data_Rescue` dataset containing multi-line text entries. Additional challenges included handling complex table structures and irregular cell boundaries, which further affected performance on both datasets. These findings highlight the need to improve table layout alignment

	Table structure recognition			Tabular data reconstruction					
Dataset	P	R	F1	Rouge-L	WER	CER	EM	F1 (Char)	F1 (Token)
Full table	0.742	0.919	0.805	0.809	0.245	0.213	0.755	0.850	0.755
Table body (Mixed)	0.783	0.878	0.820	0.770	0.207	0.199	0.793	0.867	0.793
Table body (Typed)	0.827	0.964	0.885	0.897	0.108	0.121	0.892	0.937	0.892
Text (Mixed)	0.646	0.860	0.707	0.782	0.284	0.287	0.716	0.817	0.716
Text (Typed)	0.659	0.811	0.695	0.772	0.249	0.256	0.751	0.836	0.751
Number (Mixed)	0.774	0.863	0.805	0.943	0.076	0.056	0.924	0.954	0.924
Number (Typed)	0.820	0.920	0.858	0.969	0.047	0.039	0.953	0.974	0.953

Table 5: Performance breakdown of `TrOCR-ctx` on the `UoS_Data_Rescue` dataset, evaluating various aspects such as full tables, table body, individual text-only cells, and number-only cells. The table body, text-only, and number-only cells are further categorized based on a mix of handwritten (Mixed) and typed (Typed) text.

and multi-line text recognition to enhance OCR accuracy for complex tabular data reconstruction tasks.

To gain deeper insights into TrOCR-ctx's performance, we conducted a detailed evaluation by combining TrOCR's context-aware OCR capabilities with the post-OCR correction provided by the ByT5 model. The analysis focused on various table regions within historical logbook images from the `UoS_Data_Rescue` dataset. Table 5 presents a performance breakdown, highlighting the model's ability to handle dense tables and complex data, particularly those containing mixed handwritten content. In the full table evaluation, including the header and body, `TrOCR-ctx` achieved F1-scores of 0.850 at the character level and 0.755 at the token level. This demonstrates the model's ability to capture the overall table structure while maintaining content accuracy across the image. However, when focusing solely on the table body—where most of the critical information in historical logbooks resides—the model's performance improved, suggesting that alignment issues with header cells may impact overall reconstruction accuracy. For table bodies containing a mix of handwritten and typed text, `TrOCR-ctx` achieved an F1-score of 0.793 at the token level. When evaluating only typed text table bodies, the model reached an impressive F1-score of 0.892 at the token level, indicating its superior handling of typed content compared to handwritten-mixed entries.

Analysis of different types of cell content (text vs. numbers) extraction performance revealed significant disparities between text and numerical content processing. When evaluating different types of cell content, the model showed exceptional performance on numbered cells, achieving an F1-score of 0.924 for handwritten-mixed content and 0.953 for typed-only content at the token level. In contrast, text cells scored lower, with F1-scores of 0.719 for handwritten-mixed content and 0.751 for typed-only content at the token level. While this variation suggests that handwritten text and complex layouts in text cells present more difficulty, the model's strong performance on numbered cells demonstrates its potential. With further fine-tuning and targeted improvements, particularly in handling handwritten text, `TrOCR-ctx` can continue to advance in accuracy and robustness for historical document digitization.

	Table structure recognition			Tabular data reconstruction					
Regions	P	R	F1	Rouge-L	WER	CER	EM	F1 (Char)	F1 (Token)
Tromelin✓	0.903	0.903	0.897	0.928	0.088	0.081	0.912	0.949	0.912
Diego-Suarez✓	0.836	0.866	0.840	0.902	0.115	0.109	0.885	0.936	0.885
UK✓	0.666	0.998	0.782	0.939	0.131	0.121	0.869	0.942	0.869
Ambanja Août-décembre✓	0.588	0.944	0.719	0.922	0.131	0.068	0.869	0.945	0.869
South Africa✓	0.640	0.954	0.764	0.909	0.133	0.178	0.867	0.895	0.867
Natal, Africa	0.757	0.697	0.720	0.951	0.135	0.117	0.865	0.931	0.865
Tunisia✓	0.853	0.935	0.890	0.906	0.153	0.215	0.847	0.904	0.847
Zanzibar✓	0.912	0.946	0.928	0.905	0.156	0.162	0.844	0.884	0.844
Algeria✓	0.787	0.904	0.832	0.913	0.175	0.135	0.825	0.905	0.825
Tennessee	0.414	0.914	0.530	0.822	0.188	0.218	0.812	0.860	0.812
Libya	0.782	0.881	0.819	0.898	0.199	0.205	0.801	0.858	0.801
Bear	0.352	0.875	0.488	0.834	0.201	0.177	0.799	0.869	0.799
Devon, UK	0.907	0.986	0.943	0.839	0.202	0.182	0.798	0.861	0.798
Arctic✓	0.737	0.969	0.834	0.663	0.313	0.327	0.687	0.727	0.687
Egypt*✓	0.420	0.896	0.552	0.791	0.339	0.260	0.661	0.800	0.661
India*	0.781	0.915	0.841	0.663	0.345	0.297	0.655	0.789	0.655
Uganda	0.838	0.893	0.864	0.856	0.362	0.460	0.638	0.693	0.638
Morocco	0.508	0.722	0.589	0.706	0.377	0.481	0.623	0.710	0.623
Egypt+	0.779	0.822	0.783	0.854	0.423	0.381	0.577	0.744	0.577
Madagascar	0.819	0.923	0.864	0.778	0.426	0.598	0.574	0.627	0.574
Mozambique	0.729	0.765	0.730	0.745	0.469	0.770	0.531	0.603	0.531
India+	0.884	0.876	0.879	0.566	0.473	0.576	0.527	0.665	0.527
UK and World✓	0.821	0.863	0.834	0.547	0.489	0.812	0.511	0.633	0.511
Mauritius	0.846	0.831	0.834	0.786	0.506	0.725	0.494	0.583	0.494
Ben Nevis, UK	0.981	0.845	0.907	0.627	0.520	0.510	0.480	0.668	0.480
Philippines	0.869	0.858	0.842	0.406	0.711	0.652	0.289	0.482	0.289

* The source of these logbooks is from the UK Met Office.
+ The source of these logbooks are from the US NOAA.
✓ The logbooks contain a mix of handwritten and typed text.

Table 6: Performance of `TrOCR-ctx` across different regions in the `UoS_Data_Rescue` dataset, providing a logbook-wise analysis to evaluate how the model performs on various logbook types and regions.

To gain deeper insights into regional variations in logbook layouts, we conducted a logbook-wise evaluation of `TrOCR-ctx` performance across different sources and layouts. Table 6 provides a detailed breakdown of performance, sorted by F1-scores at the token level, revealing significant differences between logbooks. Logbooks with simpler layouts and clearer handwriting consistently achieved higher F1-scores, while those with more complex layouts or degraded handwriting presented greater challenges for the model. Factors contributing to these variations include densely packed table cells (e.g., India^{+}, Ben Nevis), irregular layout complexity (e.g., UK and World, Mozambique), mixed content with varying handwriting quality (e.g., UK and World), and dense multi-line text entries (e.g., Philippines). These factors made it more difficult for the model to accurately reconstruct tables in certain cases.

To better understand these performance variations, we conducted a detailed error analysis of the OCR output. This analysis revealed several key challenges. First, a notable issue arose from the crowdsourced annotation process, particularly in the representation of numerical data common in climate logbooks. For example, annotators frequently misinterpreted decimal points or periods (.) as interpuncts (·) and periods as degree symbols (°) due to historical handwriting styles. While these inconsistencies reflect the authentic appearance of historical records, they introduced additional complexity in evaluating the model's performance on numerical data. Second, we performed an error analysis focusing on character-level substitution errors. Figure 4 illustrates frequent character substitutions encountered during digitizing the `UoS_Data_Rescue` dataset. These substitutions offer valuable insights into common misrecognition patterns. High-frequency errors, such as (., ·), (°, ·), and (I, 1), indicate

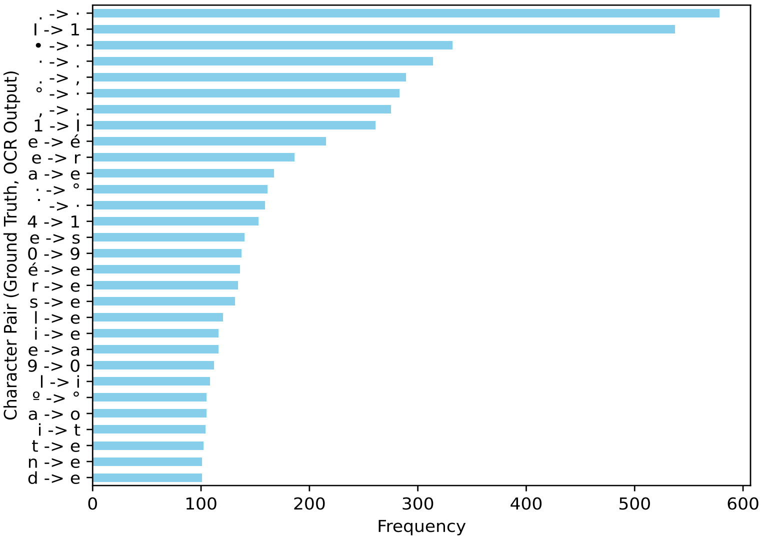


Fig. 4: Bar chart illustrating the frequency of `TrOCR-ctx` substitution errors, showcasing common character misrecognition between the ground truth and the OCR-predicted output. This breakdown highlights the top 30 frequently substituted character pairs, providing insights into recurring OCR inaccuracies and potential areas for model improvement.

challenges in distinguishing visually similar characters, particularly those with fine distinctions in handwritten dots, strokes, and numerals. Additionally, substitutions like (4, 1) and (0, 9) suggest difficulties recognizing certain numeric characters, likely due to overlapping or similarly shaped glyphs in cursive or non-standard handwriting styles. Character pairs like (e, r) and (e, é) further highlight issues with recognizing subtle handwriting variations, diacritical marks, and capitalization—common challenges in historical texts with irregular handwriting and faded ink.

These patterns emphasize the need to further enhance the `TrOCR-ctx` fine-tuning to improve accuracy in recognizing frequently misinterpreted characters in handwritten documents. Understanding these variations and annotation challenges will guide future improvements in both OCR and TSR models, particularly in addressing the unique challenges posed by historical documents with intricate layouts, poor handwriting quality, and specialized numerical notation.

5.3 Discussion

The performance analysis provides several key insights into the performance of `TrOCR-ctx` for tabular data reconstruction. One notable finding is the alignment issue between full tables and table bodies, where discrepancies in header alignment negatively impact digitization accuracy. While `TrOCR-ctx` performs well on properly segmented text lines (refer to Table 3), it struggles to maintain spatial relationships when headers are involved, leading to misaligned data during reconstruction. Incorporating contextual information from surrounding cells significantly enhances the model's ability to capture spatial relationships, particularly in historical documents with intricate layouts.

The error analysis, as illustrated in Figure 4, reveals common character substitution errors made by `TrOCR-ctx`. High-frequency errors, such as confusing visually similar characters (e.g., '.' and '·', 'I' and '1'), indicate difficulty distinguishing fine details in handwritten text. Numeric character recognition also presents difficulties, with substitutions like '4' for '1' and '0' for '9' suggesting issues with overlapping or similarly shaped glyphs in cursive or non-standard handwriting styles.

Performance varies significantly between numbered and text cells. Numbered cells consistently achieved higher F1-scores than text cells, especially when dealing with handwritten entries (refer to Table 5). For instance, numbered cells achieved F1-scores of 0.924 for handwritten-mixed content and 0.953 for typed-only content, compared to 0.719 and 0.751 for text cells, respectively. This disparity highlights the ongoing challenges in recognizing complex handwritten text and layouts. Additionally, `TrOCR-ctx` performed better on typed text than handwritten-mixed entries, which often cross cell boundaries and complicate alignment. This emphasizes the importance of refining table structure recognition (TSR) to better handle handwritten content. Lastly, the logbook-wise analysis (refer to Table 6) revealed performance variations based on layout complexity and handwriting quality, offering further opportunities for improvement.

In summary, while `TrOCR-ctx` demonstrates significant advancements in handling complex tabular data through context-aware fine-tuning, challenges related to alignment, handwritten text recognition, long multi-line text entries, and character-level distinctions still need to be addressed for further optimization. The error analysis provides valuable insights for future improvements, particularly in enhancing the model's ability to distinguish visually similar characters and handle the intricacies of handwritten text in historical documents.

6 Conclusion and Future work

This study on digitizing historical tabular records using the context-aware TrOCR model, particularly `TrOCR-ctx`, has demonstrated promising results. By introducing the specialized `UoS_Data_Rescue` historical climate logbook dataset, we provided a robust foundation for training and evaluating OCR models tailored to the complexities of historical tabular data. Through comprehensive evaluations across multiple datasets, `TrOCR-ctx` consistently outperformed baseline models, proving its effectiveness in recognizing text within complex table structures and diverse formats, including mixed handwritten and typed entries. Key findings highlight the importance of context-awareness in OCR and table reconstruction. By incorporating information from neighboring cells, `TrOCR-ctx` reduced recognition errors and more accurately captured spatial relationships within tables. However, challenges remain, particularly in aligning header cells and recognizing handwritten text that crosses cell boundaries. The model's strong performance on typed text underscores its potential for digitizing historical records with well-formatted text while also pointing to areas for improvement in handling handwritten entries.

Moving forward, future work will focus on refining table cell alignment, particularly addressing issues with header cells and improving the recognition of handwritten text—areas where `TrOCR-ctx` still faces challenges. This could involve fine-tuning models on more diverse handwritten datasets and developing advanced preprocessing techniques to better handle complex layouts. Expanding the `UoS_Data_Rescue` dataset with more intricate layouts and varied text styles will provide a broader training ground for OCR models. Additionally, re-correcting the ground truth based on identified crowdsourced annotation errors, such as misinterpretations of numerical data (e.g., decimal points misread as interpuncts or degree symbols), could enhance the accuracy of future evaluations. Efforts will also be made to improve model robustness against noise and distortions, optimize scalability for large-scale digitization projects, and incorporate feedback from domain experts to further refine the model's performance. Finally, rigorous cross-validation will ensure the model's generalization across diverse datasets and real-world scenarios, ensuring continued advancements in historical document digitization.

7 Acknowledgement

This work is funded through the Natural Environment Research Council (grant NE/S015604/1) and WCSSP South Africa project, a collaborative initiative between the Met Office, South African, and UK partners, supported by the International Science Partnership Fund (ISPF) from the UK's Department for Science, Innovation and Technology (DSIT). It is also supported by the Centre for Machine Intelligence (CMI) and Web Science Institute (WSI). The authors acknowledge the IRIDIS High-Performance Computing Facility at the University of Southampton.

References

[1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[2] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[3] Jalali, A., Kavuri, S., Lee, M.: Low-shot transfer with attention for highly imbalanced cursive character recognition. Neural Networks **143**, 489–499 (2021)

[4] KO, M.A., Poruran, S.: Ocr-nets: variants of pre-trained cnn for urdu handwritten character recognition via transfer learning. Procedia computer science **171**, 2294–2301 (2020)

[5] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 13094–13102 (2023)

[6] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer

[7] Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 572–573 (2020)

[8] Singh, L.G., Middleton, S.E.: Data rescue for historical document tables using semi-supervised learning (2024). https://doi.org/10.21203/rs.3.rs-4391424/v1

[9] Anand, A., Jaiswal, R., Bhuyan, P., Gupta, M., Bangar, S., Imam, M.M., Shah, R.R., Satoh, S.: Tc-ocr: Tablecraft ocr for efficient detection & recognition of table structure & content. In: Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval, pp. 11–18 (2023)

[10] Lehenmeier, C., Burghardt, M., Mischka, B.: Layout detection and table recognition–recent challenges in digitizing historical documents and handwritten tabular data. In: Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings 24, pp. 229–242 (2020). Springer

[11] Kim, G., Yokoo, S., Seo, S., Osanai, A., Okamoto, Y., Baek, Y.: On text localization in end-to-end ocr-free document understanding transformer without text localization supervision. In: International Conference on Document Analysis and Recognition, pp. 215–232 (2023). Springer

[12] Zhang, H., Whittaker, E., Kitagishi, I.: Extending trocr for text localization-free ocr of full-page scanned receipt images. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision, pp. 1479–1485 (2023)

[13] Chen, Y.-H., Ströbel, P.B.: Trocr meets language models: An end-to-end post-correction approach. In: International Conference on Document Analysis and Recognition, pp. 12–26 (2024). Springer

[14] Seth, D., Stureborg, R., Pruthi, D., Dhingra, B.: Learning the legibility of visual text perturbations. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3260–3273 (2023)

[15] Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: Byt5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics **10**, 291–306 (2022)

[16] Klink, S., Dengel, A., Kieninger, T.: Document structure analysis based on layout and textual features. In: Proc. of International Workshop on Document Analysis Systems, DAS2000, pp. 99–111 (2000)

[17] Oro, E., Ruffolo, M.: Trex: An approach for recognizing and extracting tables from pdf documents. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 906–910 (2009). IEEE

[18] Zanibbi, R., Blostein, D., Cordy, J.R.: A survey of table recognition: Models, observations, transformations, and inferences. Document Analysis and Recognition **7**, 1–16 (2004)

[19] Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (2022)

[20] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192–1200 (2020)

[21] Oliveira, D.A.B., Viana, M.P.: Fast cnn-based document layout analysis. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1173–1180 (2017). IEEE

[22] Fang, J., Mitra, P., Tang, Z., Giles, C.L.: Table header detection and classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 26, pp. 599–605 (2012)

[23] Paliwal, S.S., Vishwanath, D., Rahul, R., Sharma, M., Vig, L.: Tablenet:

Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 128–133 (2019). IEEE

[24] Ziomek, J., Middleton, S.E.: Glosat historical measurement table dataset: enhanced table structure recognition annotation for downstream historical data rescue. In: Proceedings of the 6th International Workshop on Historical Document Imaging and Processing, pp. 49–54 (2021)

[25] Du, Y., Li, C., Guo, R., Cui, C., Liu, W., Zhou, J., Lu, B., Yang, Y., Liu, Q., Hu, X., et al.: Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. arXiv preprint arXiv:2109.03144 (2021)

[26] Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1162–1167 (2017). IEEE

[27] Nassar, A., Livathinos, N., Lysak, M., Staar, P.: Tableformer: Table structure understanding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4614–4623 (2022)

[28] Rakshit, A., Mehta, S., Dasgupta, A.: A novel pipeline for improving optical character recognition through post-processing using natural language processing. In: 2023 IEEE Guwahati Subsection Conference (GCON), pp. 01–06 (2023). IEEE

[29] Karthikeyan, S., de Herrera, A.G.S., Doctor, F., Mirza, A.: An ocr post-correction approach using deep learning for processing medical reports. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2574–2581 (2021)

[30] Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., Hu, G.: CharBERT: Character-aware pre-trained language model. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics, pp. 39–50 (2020). https://doi.org/10.18653/v1/2020.coling-main.4

[31] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020). https://doi.org/10.18653/v1/2020.acl-main.703

[32] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis,

M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)

[33] Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European Conference on Computer Vision, pp. 564–580 (2020). Springer

[34] Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: Cord: a consolidated receipt dataset for post-ocr parsing. In: Workshop on Document Intelligence at NeurIPS 2019 (2019)

[35] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. In: International Conference on Document Analysis Recognition (2019)

[36] Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7098–7107 (2021)

[37] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., *et al.*: Icdar 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160 (2015). IEEE

[38] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision, pp. 498–517 (2022). Springer

[39] Li, X.-H., Yin, F., Dai, H.-S., Liu, C.-L.: Table structure recognition and form parsing by end-to-end object detection and relation parsing. Pattern Recognition **132**, 108946 (2022)

[40] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)

[41] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., *et al.*: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)

[42] Gao, L., Huang, Y., Déjean, H., Meunier, J.-L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: Icdar 2019 competition on table detection and recognition (ctdar). In: 2019 International Conference on Document Analysis and

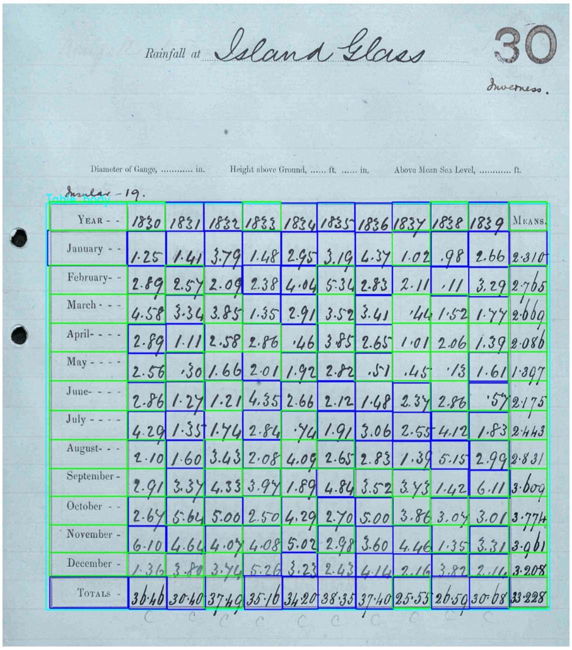
Recognition (ICDAR), pp. 1510–1515 (2019). IEEE

[43] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[44] Carrasco, R.C.: An open-source ocr evaluation tool. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 179–184 (2014)

[45] Rajpurkar, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)

A Appendix

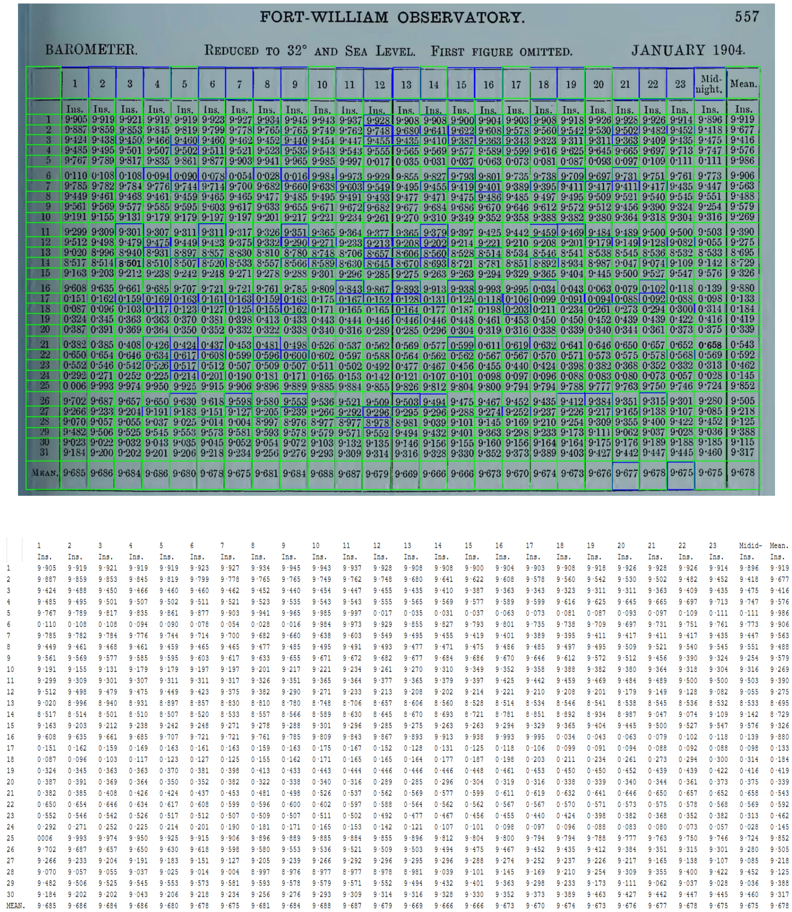


(a)

(b)

YEAR - - -	1830	1831	1832	1833	1834	1835	1836	1837	1838	1839	MEANS.
January - -	1.25	1.41	3.79	1.48	2.95	3.19	4.37	1.02	.98	2.66	2.310
February- -	2.89	2.57	2.09	2.38	4.04	5.34	2.83	2.11	.11	3.29	2.765
March - - - -	4.58	3.34	3.85	1.35	2.91	3.52	3.41	.44	1.52	1.77	2.669
April- - - -	2.89	1.11	2.58	2.86	.46	3.85	2.65	1.01	2.06	1.39	2.086
May - - - - -	2.56	.30	1.66	2.01	1.92	2.82	.51	.45	.13	1.61	1.397
June- - - - -	2.86	1.27	1.21	4.35	2.66	2.12	1.48	2.37	2.86	.51	2.175
July - - - - -	4.29	1.35	1.74	2.84	.74	1.91	3.06	2.55	4.12	1.83	2.43
August- - -	2.10	1.60	3.43	2.08	4.09	2.65	2.83	1.39	5.15	2.99	2.831
September -	2.91	3.37	4.33	3.97	1.89	4.84	3.52	3.73	1.42	6.11	3.609
October -	2.67	S.62	5.00	2.50	4.29	2.70	S.00	3.86	3.05	3.01	3.774
November -	6.10	4.64	4.02	4.08	5.02	2.98	3.60	4.46	1.35	3.31	3.061
December -	1.36	3.80	3.76	5.26	3.23	2.43	4.14	2.16	3.82	2.16	3.208
TOTALS -	36.46	30.40	37.49	35.16	34.20	38.35	37·40	25.55	26.59	30.68	33.228

Fig. 5: Visualization of the tabular data reconstruction process using a *UK* logbook image. (a) Table Structure Recognition output: The cells highlighted in blue represent the predictions made by CascadeTabnet, while those in green denote the newly generated cells derived from the coordinates of the blue-highlighted cells. (b) Tabular data reconstruction output: Content extraction using `TrOCR-ctx` and coordinate-based alignment for final table reconstruction. The performance of the TSR model on this image is 0.997 wF1-score, while the text extraction performances are 0.969 and 0.874 F1-scores at character and token levels, respectively.



(a)

(b)

Fig. 6: Visualization of the tabular data reconstruction process using a *Ben Nevis, UK* logbook image. (a) Table Structure Recognition output: The cells highlighted in blue represent the predictions made by CascadeTabnet, while those in green denote the newly generated cells derived from the coordinates of the blue-highlighted cells. (b) Tabular data reconstruction output: Content extraction using `TrOCR-ctx` and coordinate-based alignment for final table reconstruction. The performance of the TSR model on this image is 0.989 wF1-score, while the text extraction performances are 0.974 and 0.927 F1-scores at character and token levels, respectively.

B Annotation Guidelines

The guidelines for crowd annotators in this task were designed to provide clear instructions for working with table images containing highlighted cells. The primary objective for annotators is to accurately correct cell boundaries and transcribe the highlighted content. For consistency and precision in annotating various cell types, detailed instructions, as outlined in Table 7, are provided. The task encompasses several possible scenarios, each with specific actions to guide annotators in handling different cell boundaries and content configurations.

1. When the highlighted cell boundaries accurately enclose the text (i.e., all text is contained within the cell boundary lines), transcribe the text directly within the highlighted cell.
2. When the highlighted cell boundaries are inaccurate (e.g., the text extends beyond the boundary lines), adjust the cell boundaries to fully contain the text and then transcribe the content.
3. If a highlighted area merges multiple cells into a single box, remove the highlighted box, create individual boxes for each separate cell, adjust boundaries as necessary, and transcribe the content within each cell.

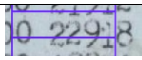
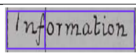
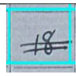
Type of cells Cells	Examples	Transcription
If the text in the cell is **easy to read and transcribe**, simply transcribe the content as it appears.		`1923`
If the highlighted cell **covers multiple distinct text regions**, adjust the cell boundary by adding new cells according to the table structure and transcribe each distinct text region within its own cell.		Cell 1: `1`, Cell 2: `9.16`
If the highlighted cell **only partially covers a text region**, correct the cell boundaries in line with the table structure, then transcribe the text within each corrected cell.		`229.18`
If a single word or group of words spans across multiple highlighted cells, **combine these cells** by adjusting boundaries so that each cell contains a single, complete text region. Then, transcribe the text accordingly.		`Information`
If any part of the highlighted text **cannot be easily transcribed**, transcribe the cell as ‘`@@@`‘. This will alert an expert to review the cell later for clarification.		`@@@`

Table 7: Guidelines on the type of cells to transcribe.

Description	**Value**
Input image size (height x width)	1024x1024
Backbone model used for feature extraction	ResNet-50
Number of output channels in the last layer of the backbone	256
Number of input channels	256
Number of fully connected (FC) layers	2
Number of output channels for each FC layer	1024
Number of stages in the cascade	3
Region Proposal Network (RPN) output threshold	2000
RPN minimum positive IoU threshold	0.3
Number of object classes (table or background)	2
Learning rate of the optimizer	0.005
Loss function for classification	Cross Entropy
Loss function for bounding box regression	Smooth L1

Table 8: Parameters for training CascadeTabNet Model