mathematics

MDPI

Article

HTTD: A Hierarchical Transformer for Accurate Table Detection in Document Images

Mahmoud SalahEldin Kasem [1,2], Mohamed Mahmoud [1,3], Bilel Yagoub [1], Mostafa Farouk Senussi [1,3], Mahmoud Abdalla [1] and Hyun-Soo Kang [1,*]

[1] Department of Information and Communication Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, Republic of Korea; mahmoud.salah@aun.edu.eg (M.S.K.)
[2] Multimedia Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
[3] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
* Correspondence: hskang@cbnu.ac.kr

Abstract: Table detection in document images is a challenging problem due to diverse layouts, irregular structures, and embedded graphical elements. In this study, we present HTTD (Hierarchical Transformer for Table Detection), a cutting-edge model that combines a Swin-L Transformer backbone with advanced Transformer-based mechanisms to achieve superior performance. HTTD addresses three key challenges: handling diverse document layouts, including historical and modern structures; improving computational efficiency and training convergence; and demonstrating adaptability to non-standard tasks like medical imaging and receipt key detection. Evaluated on benchmark datasets, HTTD achieves state-of-the-art results, with precision rates of 96.98% on ICDAR-2019 cTDaR, 96.43% on TNCR, and 93.14% on TabRecSet. These results validate its effectiveness and efficiency, paving the way for advanced document analysis and data digitization tasks.

Keywords: table detection; vision transformer; document processing; multiscale feature extraction; deformable attention; document image analysis

MSC: 68T07

Academic Editor: Konstantin Kozlov

Received: 6 December 2024
Revised: 2 January 2025
Accepted: 14 January 2025
Published: 15 January 2025

Citation: Kasem, M.S.; Mahmoud, M.; Yagoub, B.; Senussi, M.F.; Abdalla, M.; Kang, H.S. HTTD: A Hierarchical Transformer for Accurate Table Detection in Document Images. *Mathematics* **2025**, *13*, 266. https://doi.org/10.3390/math13020266

1. Introduction

Tables serve as a fundamental method for organizing and visualizing multidimensional relationships within data, making them indispensable across domains such as business, science, and government. While digital tables facilitate efficient data analysis and processing, the persistence of paper-based tables in scenarios requiring reliability and security, such as handwritten signatures, poses significant challenges. Converting table images or scanned documents into machine-readable formats requires preserving the semantic relationships and structural integrity of the original data. However, this task is complicated by the diversity in table layouts, varying element sizes, changing background colors, fonts, and borders [1,2].

In recent years, the task of detecting and analyzing tabular data in digital documents has garnered substantial interest within the computer vision and document analysis communities. Tables are a ubiquitous method of data presentation, widely used across diverse fields such as academia, business, and healthcare due to their efficiency in organizing and displaying information. Automatically detecting and extracting tables from digital documents not only aids in data digitization but also facilitates the downstream tasks of data analysis and machine learning model training [3–5].

Deep learning, a pivotal component of machine learning, is increasingly applied across various industries. In computer vision, it excels at tasks such as object detection, image classification, video analysis, and software engineering [6,7]. In Natural Language Processing (NLP), deep learning contributes to question answering and sentence similarity [8,9], as well as recognizing handwritten text in multiple languages [10,11]. The finance sector leverages deep learning for fraud detection, trading, and risk management. In gaming, it enhances decision-making, customer segmentation [12,13], and personalized recommendations. These examples underscore the expanding role of deep learning across diverse fields, highlighting its vast potential for future advancements [14].

Traditional approaches to table detection primarily relied on handcrafted features and heuristic methods that exploited the geometric and textual cues present in tables. While effective to a degree, these methods often struggled with the diversity of table formats and complex layouts seen in real-world documents. The advent of deep learning has significantly transformed this landscape, offering robust alternatives that leverage convolutional neural networks (CNNs) to capture spatial hierarchies and features directly from data. However, despite their effectiveness, CNN-based methods are predominantly local in their receptive fields and may not adequately capture the long-range dependencies crucial for recognizing tables, which are inherently structured and spatially diverse objects.

Transformers in Images [15–19] have recently demonstrated remarkable success in natural image understanding tasks, including classification, detection, and segmentation. Whether pre-trained in a supervised manner on ImageNet or through self-supervised learning, these models have achieved performance that is comparable to and often surpasses, that of CNN-based pre-trained models with a similar number of parameters.

The advent of Vision Transformers (ViTs) has transformed image processing by leveraging the self-attention mechanism to model global relationships within an image, making them highly effective for tasks like table detection. Unlike convolution-based methods, ViTs process images as sequences of tokens (patches), enabling them to dynamically focus on relevant regions while capturing complex layouts and structures. Their scalability, ability to generalize across diverse data, and superior performance over CNNs in object detection make ViTs particularly suited for detecting tables embedded in graphical elements or with irregular layouts, where traditional approaches often struggle.

In contrast to traditional object detection algorithms, DETR [20] introduces a transformative approach using Transformer-based architecture. By eliminating the need for handcrafted components, DETR matches the performance of well-optimized classical detectors. DETR treats object detection as a set prediction problem, utilizing bipartite graph matching for label assignment and learnable queries to identify objects.

Although DETR introduces an innovative approach to object detection, it encounters challenges such as slow training convergence and ambiguous query interpretability. To overcome these limitations, several enhancements have been proposed, including deformable attention mechanisms [21] and the decoupling of positional and content information [22]. Recent developments, such as DN-DETR [23] and DAB-DETR [24], further advance the framework by redefining queries as dynamic anchor boxes and incorporating denoising techniques to improve the stability of bipartite matching. These advancements have markedly improved DETR-like models, making them competitive with classical detectors in both training efficiency and inference performance.

Building on these advances, Hierarchical Transformer for Table Detection (HTTD) introduces a transformative approach to table detection by combining a hierarchical vision backbone with advanced Transformer-based mechanisms. Our model comprises a backbone, a multilayer Transformer encoder, a multilayer Transformer decoder, and multiple prediction heads. Similar to DAB-DETR, the decoder queries are formulated as dynamic

anchor boxes and refined iteratively across decoder layers. Following DN-DETR, ground-truth (GT) labels and boxes with added noise are introduced into the Transformer decoder layers during training to stabilize bipartite matching. Additionally, deformable attention is incorporated for computational efficiency. We use three methods to enhance performance. Contrastive denoising improves one-to-one matching by adding both positive and negative samples derived from the same ground truth simultaneously. By applying different noise levels to the same ground-truth box, the box with smaller noise is labeled positive, while the other is negative. This contrastive approach helps the model avoid duplicate predictions for the same target. Mixed query selection bridges DETR-like models and classical two-stage models by optimizing query initialization. Initial anchor boxes are selected as positional queries from the encoder outputs while the content queries remain learnable, allowing the first decoder layer to focus on spatial priors. The look-forward-twice scheme leverages refined box information from later layers for optimizing adjacent earlier layers. This method updates parameters using gradients informed by later-layer refinements, ensuring more robust optimization.

Although significant progress has been made in table detection, several key challenges remain in the field: (1) Traditional methods often struggle with the vast diversity in table formats, particularly in complex documents such as legal texts or scientific papers. This limitation reduces their applicability across varied document types. (2) Existing approaches frequently fail to handle complex table structures or degraded document images, leading to inaccuracies in detection and recognition. This is particularly problematic for historical or scanned documents with quality issues. (3) Many current models do not fully utilize the combination of visual and textual cues present in documents, which is crucial for precise table detection and structure recognition.

To address these gaps, this paper introduces the Hierarchical Transformer for Table Detection (HTTD), which incorporates several innovative features: (1) HTTD leverages a hierarchical vision backbone to capture features at multiple scales, enabling robust handling of diverse table layouts, ranging from simple to highly complex structures. (2) Techniques such as contrastive denoising and the look-forward-twice scheme enhance detection accuracy and stabilize the model during training, particularly in challenging scenarios involving complex or degraded tables. (3) By incorporating both visual and textual information through Transformer-based mechanisms, HTTD improves the detection and recognition of tables, even in documents with intricate layouts or varying quality.

By addressing these limitations, HTTD offers a robust and generalizable solution capable of handling diverse and real-world document scenarios, thereby advancing the state of the art in table detection. The contributions of this paper include the following:

- Proposing a novel Transformer-based architecture, HTTD, to address table detection challenges in both historical and modern documents with diverse and irregular layouts.
- Introducing innovative techniques such as contrastive denoising, mixed query selection, and look-forward-twice refinement, which enhance detection accuracy, improve training speed, and stabilize the model during convergence.
- Demonstrating the generalizability of HTTD through experiments on non-standard table types and other detection tasks, such as breast cancer detection and receipt key detection, highlighting its potential for diverse applications.
- Providing comprehensive experimental results, including ablation studies and comparisons with state-of-the-art methods, to validate the effectiveness and efficiency of the HTTD model.

The remainder of this paper is structured as follows. Section 2 reviews related work, discussing heuristic, machine learning-based, and deep learning-based approaches to table

detection, highlighting their limitations, and situating our proposed HTTD model within state-of-the-art research. Section 3 describes the HTTD methodology, detailing its architecture, the use of Swin Transformers for hierarchical feature extraction, and innovations like contrastive denoising, mixed query selection, and look-forward-twice refinement. Section 4 outlines the experimental setup, datasets, and evaluation metrics, presenting quantitative results and ablation studies to validate the contributions of HTTD. Section 5 discusses the findings, providing qualitative examples, addressing the challenges of irregular table layouts, and highlighting HTTD's strengths and limitations. Finally, Section 6 concludes this paper by summarizing key contributions and proposing future research directions, including table structure recognition and semi-supervised learning for broader applicability.

2. Related Work

Table detection has been a subject of research for an extensive period. Researchers have employed various approaches, which can be broadly categorized into heuristic-based methods, machine learning-based methods, and deep learning-based methods.

With the emergence of deep learning (DL), advanced object detection algorithms, and the availability of publicly accessible datasets, the development of fully data-driven approaches has significantly increased. A Gilani et al. [25] were among the first to propose a DL-based approach for table detection using Faster R-CNN [26]. In their method, document images are initially pre-processed before being passed through a Region Proposal Network (RPN) for table detection followed by a fully connected neural network for classification. Their approach demonstrates high precision across a variety of document images, including documents, research papers, and periodicals, accommodating diverse layouts. Then D Prasad et al. [5] presented an innovative method for automatic table detection in document images, tackling the dual challenges of table detection and table structure recognition. This method employs a unified convolutional neural network (CNN) model to deliver a comprehensive deep learning-based end-to-end solution for both tasks. The proposed model, CascadeTabNet, utilizes a Cascade mask Region-based CNN High-Resolution Network (Cascade mask R-CNN HRNet) to concurrently identify table regions and recognize the structural cells within those tables.

After that, A. Samari et al. [27] developed an innovative approach for detecting tables in digitized historical prints, addressing challenges posed by varied table characteristics and their visual similarity to other elements. The study introduced the NAS dataset, enhancing the diversity of evaluation. The method employed the Gabor filter for dataset preparation and Faster-RCNN for detection, effectively overcoming the limitations of labeled data through weakly supervised bounding box extraction and pseudo-labeling, thereby improving model generalization. Furthermore, M Agarwal et al. [28] introduced the Composite Deformable Cascade Network (CDeC-Net), an advanced deep learning framework for detecting tables in document images. This network enhances Mask R-CNN by incorporating a dual backbone structure with deformable convolutional layers, allowing for effective handling of tables across various scales and improving detection accuracy at higher Intersection over Union (IoU) thresholds. The model was rigorously evaluated on multiple benchmark datasets. Furthermore, X Zheng et al. [29] introduced the Global Table Extractor (GTE), a method for jointly detecting tables and recognizing cell structures that can be implemented on any object detection model. To enhance their table network using cell placement predictions, the authors developed GTE-Table, which introduces a new penalty based on the inherent cell confinement limitations of tables. Additionally, a novel hierarchical cell identification network, GTE-Cell, leverages table styles. To efficiently and cost-effectively build a large corpus of training and test data, the authors devised a method to automatically classify table and cell structures in existing texts.

In 2022, D.D. Nguyen [30] introduced TableSegNet, a fully convolutional network designed to simultaneously separate and detect tables with a streamlined architecture. TableSegNet utilizes a shallower path to pinpoint table locations at high resolution and a deeper path to detect table regions at low resolution, subsequently dividing these regions into individual tables. Throughout the feature extraction process, TableSegNet employs convolutional blocks with large kernel sizes and incorporates an additional table-border class in the main output to enhance detection and separation capabilities. Furthermore, J Li et al. [31] proposed DiT, a self-supervised pre-trained Document Image Transformer model designed for Document AI tasks such as classification, layout analysis, table detection, and OCR text detection. By using large-scale unlabeled document images and a Masked Image Modeling (MIM) strategy, DiT eliminates the need for human-labeled data and achieves state-of-the-art results across multiple benchmarks. The model significantly improves performance in key tasks and provides a strong, adaptable backbone for various Document AI applications, addressing the lack of large-scale labeled datasets in the field. Furthermore, M. Haloi et al. [32] addressed the limitations of existing table detection benchmarks by introducing a comprehensive, large-scale dataset comprising over seven thousand samples with diverse table structures from various sources. The study employed convolutional neural network-based methods, demonstrating their superiority over traditional computer vision techniques in detecting table structures within documents. This dataset serves as a valuable resource for advancing the development of efficient deep learning methods for document layout understanding and tabular data processing.

In 2023, Q Ren et al. [33] proposed a table detection method based on YOLOv5, incorporating deformable convolutional networks (DCNs), a new residual module (ResDCN), a Global Attention Mechanism (GAM), and Adaptive Spatial Feature Fusion (ASFF) to improve detection accuracy for complex tables with diverse layouts. Evaluated on the TNCR and ICDAR-2017POD datasets, the model showed significant improvements, particularly on TNCR, where it excelled in detecting challenging wireless tables with a 2% increase in F1-score. On ICDAR-2017POD, it achieved an F1-score of 96.7% and a 98% recall rate, demonstrating robust performance across various table formats. In addition, T. Shehzadi et al. [34] introduced a novel semi-supervised table detection method that employs the deformable Transformer, a sophisticated deep learning technique. Unlike traditional deep learning approaches that require substantial amounts of labeled data, this method significantly minimizes the need for labeled samples. By utilizing the deformable Transformer, the proposed method achieves remarkable results across various datasets, including PubLayNet, DocBank, ICADR-19, and TableBank. It outperforms both fully supervised and previous semi-supervised methods, demonstrating superior performance with a limited amount of labeled data.

Y Ni et al. [35] proposed a Transformer-based model for table detection in document images, addressing the challenge of detecting small or irregularly shaped tables. By fine-tuning a pre-trained Transformer and integrating a Dual-branch Dilated Context Convolutional (DCC) module, their approach enhances feature extraction and prediction accuracy for tables of varying sizes and shapes. Additionally, multilevel residual convolutional layers are employed for improved multiscale feature fusion. Evaluated on public datasets, their model achieved advanced performance in table detection, demonstrating robustness and precision in various document layouts.

3. Methodology

In our HTTD model, we structure detection queries into two components: a positional component, represented as a 4D anchor box (x, y, w, h), and a content component that remains learnable. The positional component includes the center coordinates (x, y) and

dimensions (w, h) of an anchor box, enabling dynamic refinement of anchor boxes layer by layer within the decoder.

To address the slow training convergence typical in Transformer-based detection models, we introduce a denoising training technique. This approach involves feeding ground-truth labels and boxes with controlled noise into the Transformer decoder. The model is then trained to reconstruct the original ground-truth boxes, which stabilizes training and accelerates convergence. Specifically, we add noise $(\Delta x, \Delta y, \Delta w, \Delta h)$ to each ground-truth box, constrained within a small range relative to the box's dimensions, ensuring the noisy anchor remains close to the original box. An auxiliary loss term, based on this denoising process, further enhances the stability and speed of model training.

We also incorporate a deformable attention mechanism, which introduces reference points to focus attention on key sampling locations surrounding each reference. This targeted approach to attention allows the model to prioritize relevant regions within each image. Additionally, we implement *query selection*, where we directly use encoder features and reference boxes to initialize decoder queries. To further refine bounding box predictions, we employ an iterative refinement strategy with a unique "look-forward-twice" technique, which updates gradient paths effectively between layers.

By combining these elements—deformable attention, denoising anchor boxes, and iterative query refinement—our model provides a robust approach for table detection. Each component is designed to enhance detection accuracy, accelerate convergence, and maintain stable training, making our model well suited for high-precision table localization tasks.

Model Overview

Our HTTD model is a Transformer-based architecture for table detection, consisting of a backbone, a multilayer Transformer encoder, a multilayer Transformer decoder, and prediction heads. Given an input image, we first extract multiscale features using backbones like ResNet [36] or a Swin Transformer [18]. These features, along with positional embeddings, are passed into the Transformer encoder. After enhancing the features through the encoder layers, we initialize anchor boxes as positional queries for the decoder while leaving content queries learnable. This initialization process allows for the iterative refinement of bounding box predictions as the decoder updates queries layer by layer using deformable attention [21]. The final outputs are refined anchor boxes and classification scores.

Contrastive denoising (CDN): This technique is added to stabilize training and improve detection accuracy by differentiating between high-quality anchors (positive examples) and less relevant anchors (negative examples). Positive queries are generated with minimal noise, constrained within an inner region close to the ground-truth (GT) box, and are trained to reconstruct the GT box. Negative queries are sampled from an outer region with larger noise, specifically designed to predict "no object". By explicitly teaching the model to reject anchors farther from GT boxes, CDN prevents confusion caused by multiple anchors near a single GT box and reduces duplicate predictions.

In the HTTD model, CDN plays a crucial role in handling diverse table layouts by ensuring precise localization of table boundaries. For each GT box, CDN generates one positive and one negative query, forming balanced training pairs that guide the model to focus on high-quality anchors while suppressing irrelevant ones. This approach enhances the model's ability to distinguish between valid and background regions, significantly improving detection performance in complex document layouts. Moreover, the reconstruction loss (L1 and GIoU for box regression, focal loss for classification) ensures robust predictions, making CDN effective for both simple and densely populated layouts, as shown in Figure 1.

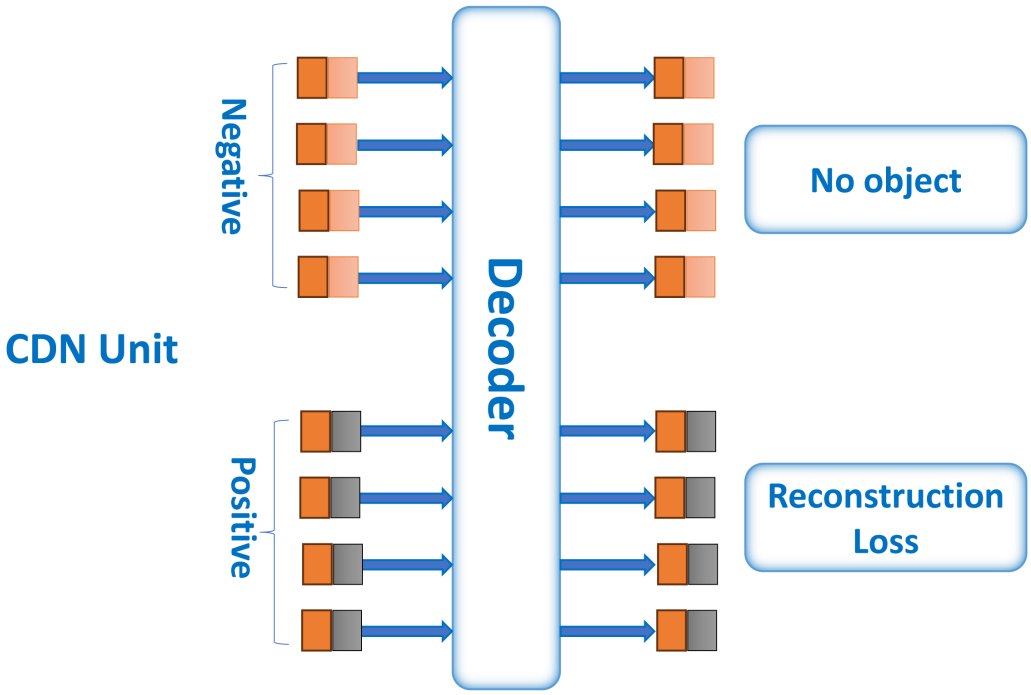


Figure 1. Illustration of the CDN structure, demonstrating the generation of positive and negative examples. The inner square represents the region near the ground-truth (GT) box, where anchors are classified as positive examples. Anchors in the region between the inner and outer squares are classified as negative examples.

Effectiveness metric: To assess anchor quality, we use *Average Top-K Distance* (ATD(k)), which measures the average distance of the k farthest anchors from matched ground-truth boxes. For a validation set with N ground-truth boxes $b_i = (x_i, y_i, w_i, h_i)$, and their corresponding anchors a_i, the ATD(k) is given by

$$\mathrm{ATD}(k) = \frac{1}{k}\sum\{\mathrm{topK}(\{\|b_0 - a_0\|_1, \|b_1 - a_1\|_1, \ldots, \|b_{N-1} - a_{N-1}\|_1\}, k)\} \quad (1)$$

where $\|b_i - a_i\|_1$ represents the L1 distance between each ground-truth box b_i and anchor a_i. This approach is particularly effective for small object detection, providing a consistent improvement in average precision on such targets.

Mixed query selection: In traditional Transformer-based detection models, decoder queries are typically static embeddings without input from specific encoder features. Our approach enhances decoder queries by introducing a *mixed query selection* strategy. In this approach, we initialize positional queries for the decoder using the positional information of the top-K features from the encoder, allowing the model to select contextually relevant anchors dynamically. Unlike conventional methods that initialize both positional and content queries, we retain the content queries as learnable parameters, allowing the model to pool more comprehensive content features from the encoder layers.

By incorporating encoder-selected features solely for positional queries, we avoid potential ambiguity in content queries caused by preliminary content features that may include multiple objects or partial object views. This mixed query selection provides a refined initialization for the positional component of decoder queries, enhancing the model's localization capabilities in complex scenes.

Look forward twice: We also add a *look-forward-twice* strategy for iterative box refinement, designed to improve both the accuracy and stability of box predictions. In typical refinement processes, gradient backpropagation is blocked between layers to stabilize

training. Our look-forward-twice method, however, updates the parameters of layer i based on losses from both layer i and layer $(i+1)$, allowing improved box information from a later layer to influence earlier layers.

In this scheme, each box prediction at layer i incorporates both its initial box from the preceding layer and an additional refinement step based on the subsequent layer's output. Specifically, given an input box b_{i-1} at the i-th layer, the final prediction $b_i^{(\text{pred})}$ is defined as follows:

$$\begin{aligned} \Delta b_i &= \text{Layer}_i(b_{i-1}), & b_i' &= \text{Update}(b_{i-1}, \Delta b_i), \\ b_i &= \text{Detach}(b_i'), & b_i^{(\text{pred})} &= \text{Update}(b_{i-1}', \Delta b_i), \end{aligned} \tag{2}$$

where b_i' is the undetached version of b_i. The function $\text{Update}(\cdot, \cdot)$ refines the box prediction by applying the predicted offset Δb_i. This iterative refinement helps achieve more precise localization by leveraging information across adjacent layers.

This is detailed in our HTTD model architecture shown in Figure 2, illustrating the application of advanced deep learning techniques for enhanced model precision. To clarify the methodology, we present a pseudo-code in Algorithm 1, which outlines the sequential steps involved in iteratively refining bounding boxes across layers. The algorithm highlights the core components, including the computation of deltas, intermediate updates, and the look-forward mechanism that incorporates feedback from subsequent layers for improved optimization.

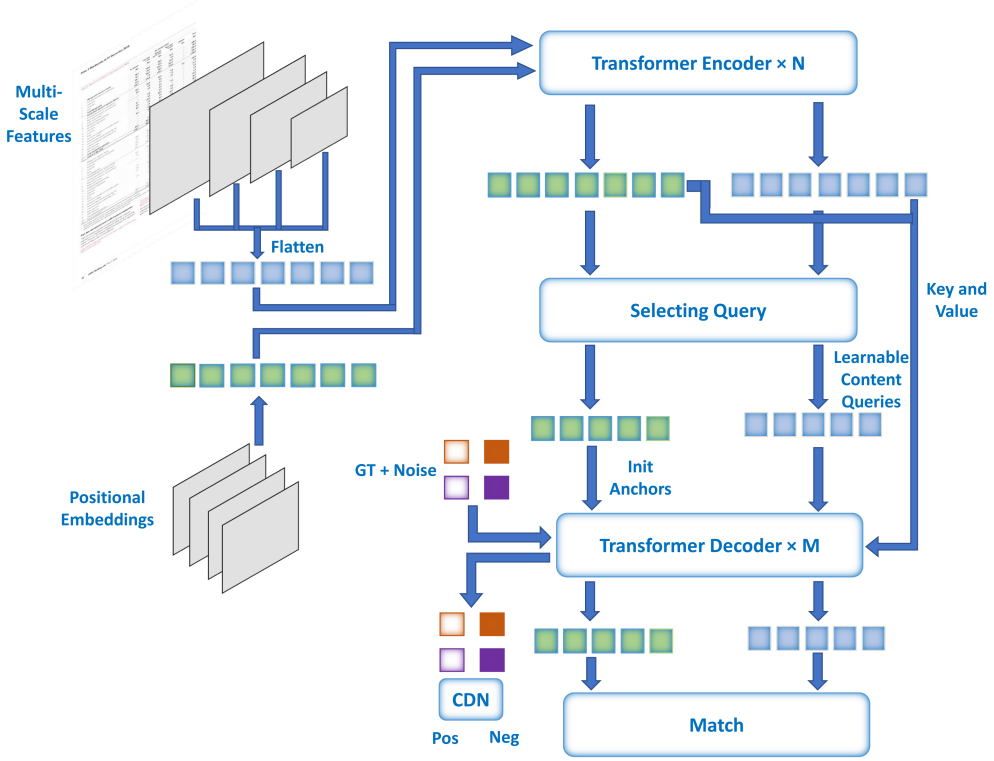


Figure 2. The architecture of the proposed model for table detection

Algorithm 1 Look-forward-twice refinement strategy.

Input:

- Initial bounding box predictions: $B_{\text{init}} = \{b_0, b_1, \ldots, b_n\}$
- Encoder features: F_{encoder}
- Number of refinement layers: L
- Prediction function: Predict(box, features)
- Update function: Update(box, Δ)

Output: Final refined bounding boxes: B_{refined}

Initialize: $B_{\text{current}} \leftarrow B_{\text{init}}$
for $i = 1$ to L **do**
    Compute deltas using the current layer:

$$\Delta B_i \leftarrow \text{Predict}(B_{\text{current}}, F_{\text{encoder}})$$

    Update bounding boxes with layer-specific deltas:

$$B_{\text{intermediate}} \leftarrow \text{Update}(B_{\text{current}}, \Delta B_i)$$

    if $i < L$ **then**
        Look-forward mechanism: Incorporate feedback from the next layer:

$$\Delta B_{\text{next}} \leftarrow \text{Predict}(B_{\text{intermediate}}, F_{\text{encoder}})$$

$$B_{\text{current}} \leftarrow \text{Update}(B_{\text{intermediate}}, \Delta B_{\text{next}})$$

    else
        Final refinement:

$$B_{\text{current}} \leftarrow B_{\text{intermediate}}$$

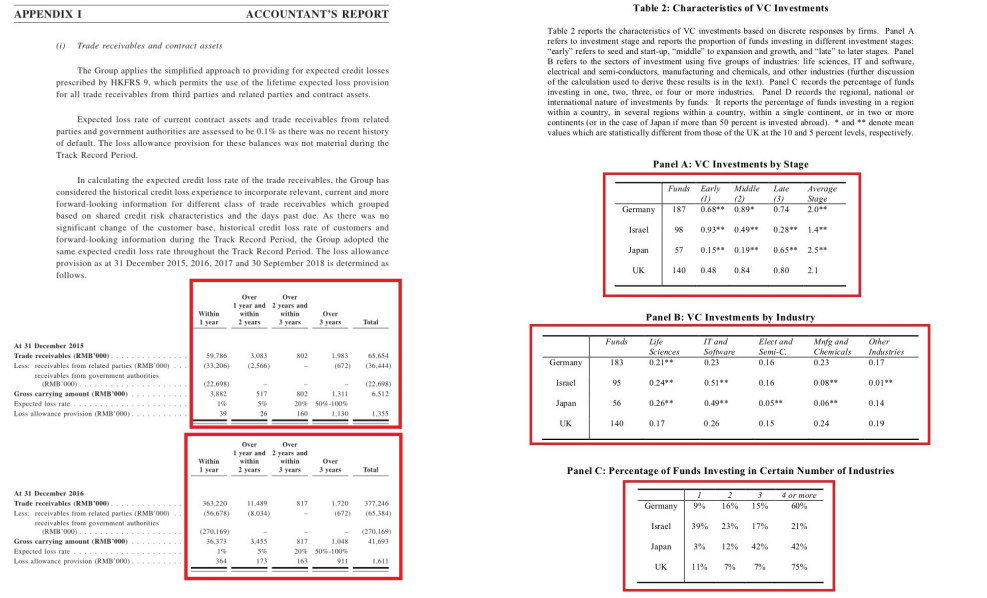
    end if
end for
Output: $B_{\text{refined}} \leftarrow B_{\text{current}}$

4. Experiment Result

4.1. Dataset

We try our HTTD model on different datasets such as the ICDAR2019 cTDaR dataset, as detailed by Gao [37], which is a specialized resource designed for table detection (Track A) and recognition (Track B). It categorizes data into historical and modern divisions. The modern segment includes a variety of formats such as scientific articles, forms, and financial records, while the historical segment comprises images from handwritten ledgers and ancient texts. The dataset consists of 1600 training images and 839 testing images. Track A focuses on images containing tables, whereas Track B is divided into two subtracks for table structure recognition, with or without prior knowledge. Examples of the dataset are shown in Figure 3.

Furthermore, The TNCR dataset, introduced by Abdallah [38], is a newly curated collection of table images of varying quality sourced from freely accessible websites. This dataset is designed for recognizing and classifying tables in scanned document images across five distinct categories. It includes approximately 6621 images featuring 9428 captioned tables. Utilizing advanced deep learning methodologies for table detection, the research established substantial baselines. Notably, the integration of Deformable DETR with a ResNet-50 Backbone Network achieved the highest performance metrics on the TNCR dataset, with an accuracy of 86.7%, a recall of 89.6%, and an F1-score of 88.1%. Examples of the dataset are shown in Figure 4.



APPENDIX I ACCOUNTANT'S REPORT

(i) Trade receivables and contract assets

The Group applies the simplified approach to providing for expected credit losses prescribed by HKFRS 9, which permits the use of the lifetime expected loss provision for all trade receivables from third parties and related parties and contract assets.

Expected loss rate of current contract assets and trade receivables from related parties and government authorities are assessed to be 0.1% as there was no recent history of default. The loss allowance provision for these balances was not material during the Track Record Period.

In calculating the expected credit loss rate of the trade receivables, the Group has considered the historical credit loss experience to incorporate relevant, current and more forward-looking information for different class of trade receivables which grouped based on shared credit risk characteristics and the days past due. As there was no significant change of the customer base, historical credit loss rate of customers and forward-looking information during the Track Record Period, the Group adopted the same expected credit loss rate throughout the Track Record Period. The loss allowance provision as at 31 December 2015, 2016, 2017 and 30 September 2018 is determined as follows.

	Within 1 year	Over 1 year and within 2 years	Over 2 years and within 3 years	Over 3 years	Total
At 31 December 2015					
Trade receivables (RMB'000)	59,786	3,083	802	1,983	65,654
Less: receivables from related parties (RMB'000)	(33,206)	(2,566)	–	(672)	(36,444)
receivables from government authorities (RMB'000)	(22,698)	–	–	–	(22,698)
Gross carrying amount (RMB'000)	3,882	517	802	1,311	6,512
Expected loss rate	1%	5%	20%	50%-100%	
Loss allowance provision (RMB'000)	39	26	160	1,130	1,355

	Within 1 year	Over 1 year and within 2 years	Over 2 years and within 3 years	Over 3 years	Total
At 31 December 2016					
Trade receivables (RMB'000)	363,220	11,489	817	1,720	377,246
Less: receivables from related parties (RMB'000)	(56,678)	(8,034)	–	(672)	(65,384)
receivables from government authorities (RMB'000)	(270,169)	–	–	–	(270,169)
Gross carrying amount (RMB'000)	36,373	3,455	817	1,048	41,693
Expected loss rate	1%	5%	20%	50%-100%	
Loss allowance provision (RMB'000)	364	173	163	911	1,611

– I-34 –

Table 2: Characteristics of VC Investments

Table 2 reports the characteristics of VC investments based on discrete responses by firms. Panel A refers to investment stage and reports the proportion of funds investing in different investment stages: "early" refers to seed and start-up, "middle" to expansion and growth, and "late" to later stages. Panel B refers to the sectors of investment using five groups of industries: life sciences, IT and software, electrical and semi-conductors, manufacturing and chemicals, and other industries (further discussion of the calculation used to derive these results is in the text). Panel C records the percentage of funds investing in one, two, three, or four or more industries. Panel D records the regional, national or international nature of investments by funds. It reports the percentage of funds investing in a region within a country, in several regions within a country, within a single continent, or in two or more continents (or in the case of Japan if more than 50 percent is invested abroad). * and ** denote mean values which are statistically different from those of the UK at the 10 and 5 percent levels, respectively.

Panel A: VC Investments by Stage

	Funds	*Early (1)*	*Middle (2)*	*Late (3)*	*Average Stage*
Germany	187	0.68**	0.89*	0.74	2.0**
Israel	98	0.93**	0.49**	0.28**	1.4**
Japan	57	0.15**	0.19**	0.65**	2.5**
UK	140	0.48	0.84	0.80	2.1

Panel B: VC Investments by Industry

	Funds	*Life Sciences*	*IT and Software*	*Elect and Semi-C.*	*Mnfg and Chemicals*	*Other Industries*
Germany	183	0.21**	0.23	0.16	0.23	0.17
Israel	95	0.24**	0.51**	0.16	0.08**	0.01**
Japan	56	0.26**	0.49**	0.05**	0.06**	0.14
UK	140	0.17	0.26	0.15	0.24	0.19

Panel C: Percentage of Funds Investing in Certain Number of Industries

	1	*2*	*3*	*4 or more*
Germany	9%	16%	15%	60%
Israel	39%	23%	17%	21%
Japan	3%	12%	42%	42%
UK	11%	7%	7%	75%

29

Figure 3. Examples of images in ICDAR 2019 cTDaR dataset.

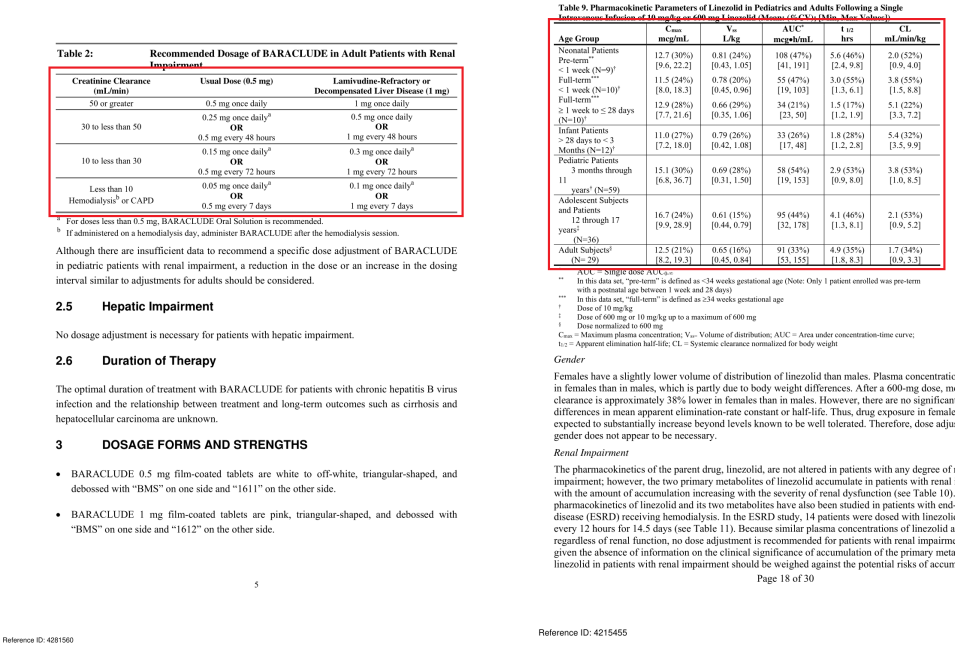


Table 2: Recommended Dosage of BARACLUDE in Adult Patients with Renal Impairment

Creatinine Clearance (mL/min)	Usual Dose (0.5 mg)	Lamivudine-Refractory or Decompensated Liver Disease (1 mg)
50 or greater	0.5 mg once daily	1 mg once daily
30 to less than 50	0.25 mg once daily[a] **OR** 0.5 mg every 48 hours	0.5 mg once daily **OR** 1 mg every 48 hours
10 to less than 30	0.15 mg once daily[a] **OR** 0.5 mg every 72 hours	0.3 mg once daily[a] **OR** 1 mg every 72 hours
Less than 10 Hemodialysis[b] or CAPD	0.05 mg once daily[a] **OR** 0.5 mg every 7 days	0.1 mg once daily[a] **OR** 1 mg every 7 days

[a] For doses less than 0.5 mg, BARACLUDE Oral Solution is recommended.
[b] If administered on a hemodialysis day, administer BARACLUDE after the hemodialysis session.

Although there are insufficient data to recommend a specific dose adjustment of BARACLUDE in pediatric patients with renal impairment, a reduction in the dose or an increase in the dosing interval similar to adjustments for adults should be considered.

2.5 Hepatic Impairment

No dosage adjustment is necessary for patients with hepatic impairment.

2.6 Duration of Therapy

The optimal duration of treatment with BARACLUDE for patients with chronic hepatitis B virus infection and the relationship between treatment and long-term outcomes such as cirrhosis and hepatocellular carcinoma are unknown.

3 DOSAGE FORMS AND STRENGTHS

- BARACLUDE 0.5 mg film-coated tablets are white to off-white, triangular-shaped, and debossed with "BMS" on one side and "1611" on the other side.
- BARACLUDE 1 mg film-coated tablets are pink, triangular-shaped, and debossed with "BMS" on one side and "1612" on the other side.

5

Reference ID: 4281560

Table 9. Pharmacokinetic Parameters of Linezolid in Pediatrics and Adults Following a Single Intravenous Infusion of 10 mg/kg or 600 mg Linezolid (Mean+ (% CV); [Min, Max Values])

Age Group	C_{max} mcg/mL	V_{ss} L/kg	AUC[*] mcg•h/mL	$t_{1/2}$ hrs	CL mL/min/kg
Neonatal Patients Pre-term[**] < 1 week (N=9)[†]	12.7 (30%) [9.6, 22.2]	0.81 (24%) [0.43, 1.05]	108 (47%) [41, 191]	5.6 (46%) [2.4, 9.8]	2.0 (52%) [0.9, 4.0]
Full-term[***] < 1 week (N=10)[†]	11.5 (24%) [8.0, 18.3]	0.78 (20%) [0.45, 0.96]	55 (47%) [19, 103]	3.0 (55%) [1.3, 6.1]	3.8 (55%) [1.5, 8.8]
Full-term[***] ≥ 1 week to ≤ 28 days (N=10)[†]	12.9 (28%) [7.7, 21.6]	0.66 (29%) [0.35, 1.06]	34 (21%) [23, 50]	1.5 (17%) [1.2, 1.9]	5.1 (22%) [3.3, 7.2]
Infant Patients > 28 days to < 3 Months (N=12)[†]	11.0 (27%) [7.2, 18.0]	0.79 (26%) [0.42, 1.08]	33 (26%) [17, 48]	1.8 (28%) [1.2, 2.8]	5.4 (32%) [3.5, 9.9]
Pediatric Patients 3 months through 11 years[‡] (N=59)	15.1 (30%) [6.8, 36.7]	0.69 (28%) [0.31, 1.50]	58 (54%) [19, 153]	2.9 (53%) [0.9, 8.0]	3.8 (53%) [1.0, 8.5]
Adolescent Subjects and Patients 12 through 17 years[‡] (N=36)	16.7 (24%) [9.9, 28.9]	0.61 (15%) [0.44, 0.79]	95 (44%) [32, 178]	4.1 (46%) [1.3, 8.1]	2.1 (53%) [0.9, 5.2]
Adult Subjects[§] (N= 29)	12.5 (21%) [8.2, 19.3]	0.65 (16%) [0.45, 0.84]	91 (33%) [53, 155]	4.9 (35%) [1.8, 8.3]	1.7 (34%) [0.9, 3.3]

[*] AUC = Single dose $AUC_{0-\infty}$
[**] In this data set, "pre-term" is defined as <34 weeks gestational age (Note: Only 1 patient enrolled was pre-term with a postnatal age between 1 week and 28 days)
[***] In this data set, "full-term" is defined as ≥34 weeks gestational age
[†] Dose of 10 mg/kg
[‡] Dose of 600 mg or 10 mg/kg up to a maximum of 600 mg
[§] Dose normalized to 600 mg
C_{max} = Maximum plasma concentration; V_{ss} = Volume of distribution; AUC = Area under concentration-time curve; $t_{1/2}$ = Apparent elimination half-life; CL = Systemic clearance normalized for body weight

Gender

Females have a slightly lower volume of distribution of linezolid than males. Plasma concentrations are higher in females than in males, which is partly due to body weight differences. After a 600-mg dose, mean oral clearance is approximately 38% lower in females than in males. However, there are no significant gender differences in mean apparent elimination-rate constant or half-life. Thus, drug exposure in females is not expected to substantially increase beyond levels known to be well tolerated. Therefore, dose adjustment by gender does not appear to be necessary.

Renal Impairment

The pharmacokinetics of the parent drug, linezolid, are not altered in patients with any degree of renal impairment; however, the two primary metabolites of linezolid accumulate in patients with renal impairment, with the amount of accumulation increasing with the severity of renal dysfunction (see Table 10). The pharmacokinetics of linezolid and its two metabolites have also been studied in patients with end-stage renal disease (ESRD) receiving hemodialysis. In the ESRD study, 14 patients were dosed with linezolid 600 mg every 12 hours for 14.5 days (see Table 11). Because similar plasma concentrations of linezolid are achieved regardless of renal function, no dose adjustment is recommended for patients with renal impairment. However, given the absence of information on the clinical significance of accumulation of the primary metabolites, use of linezolid in patients with renal impairment should be weighed against the potential risks of accumulation of

Page 18 of 30

Reference ID: 4215455

Figure 4. Examples of images in TNCR dataset.

Furthermore, TabRecSet, introduced by F Yang (2023) [39], investigates table recognition (TR) within the domain of pattern recognition, encompassing table detection (TD), table structure recognition (TSR), and table content recognition (TCR). The study introduces the Table Recognition Set (TabRecSet), a groundbreaking dataset that uniquely includes both English and Chinese languages, tailored to support comprehensive end-to-end TR research. TabRecSet consists of 38.1K tables, with 20.4K in English and 17.7K in Chinese, presented in various formats such as tables with complete and incomplete borders, with regular and irregular shapes, and sourced from scanned images, camera-captured images, documents, Excel sheets, educational materials, and financial invoices. Additionally, the study introduces TableMe, an innovative annotation tool designed to enhance anno-

tation efficiency and quality through features that promote visualization and interactive engagement. Examples of the dataset are shown in Figure 5.

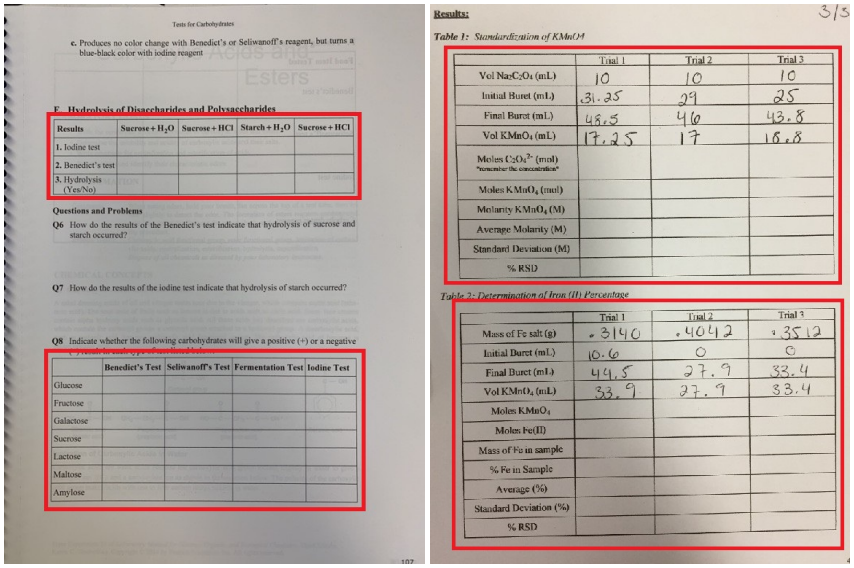


Figure 5. Examples of images in TabRecSet.

4.2. Experiment Setup

In this study, we organized our dataset into two subsets, 90% for training and 10% for validation, facilitating extensive training. During fine-tuning, we augmented image sizes by 1.5 times their original dimensions to improve feature detection. The model's training was optimized with an initial learning rate of 1×10^{-4} using the AdamW [40,41] optimizer and the same rate for weight decay. We employed L1 and Generalized Intersection over Union (GIOU) [42] losses for box regression and focal loss (with $\alpha = 0.25, \gamma = 2$) for classification, aiming to enhance bounding box precision and classification accuracy. Model performance was evaluated using the average precision (AP) metric across various Intersection over Union (IoU) thresholds and object scales, providing insights into its accuracy and capability to handle different object sizes. The implementation was conducted on a dual NVIDIA GeForce RTX 4090 with 32 GB memory, using PyTorch 2.4.1. Training was completed in approximately 5 h, outperforming CNN-based models like Cascade Mask R-CNN, which typically require 8–10 h under similar conditions. This efficiency is attributed to the optimized Swin-L backbone and advanced techniques such as contrastive denoising and look-forward-twice refinement. For inference, HTTD achieves an average processing speed of 25 images per second, exceeding the performance of Faster R-CNN (18 images/second). The use of deformable attention focuses computations on relevant regions, reducing latency without compromising accuracy.

4.3. Results

Our sophisticated model for advanced table detection, HTTD, was extensively evaluated using a meticulously curated dataset. Comparative analyses reveal that this model surpasses current state-of-the-art methods in terms of performance.

Table 1 compares our model with other models like HRNets Cascade Mask R-CNN, HRNets Mask R-CNN, HRNets HTC, HRNets Faster R-CNN, HRNets Cascade R-CNN, Mask R-CNN ResNeXt-101, and YOLOv5 + ResDCN GAM + ASFF. Our model significantly outperforms others, especially at higher IoU thresholds, which is crucial for accurate table detection in TNCR.

For instance, at an IoU threshold of 60%, our model scores 0.987 in precision, surpassing the HRNets Cascade Mask R-CNN model (0.884). At an IoU threshold of 80%,

our model scores 0.961 in precision, demonstrating its superior ability to pinpoint table locations accurately. This high performance across various IoU thresholds highlights our model's effectiveness in detecting both small and complex tables, making it valuable for efficient data processing.

Table 1. Table detection for TNCR (with the best values highlighted in bold).

Approach	Method	Metric	IoU										
			10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	Avg
Abdallah [3]	HRNets Cascade Mask R-CNN	Precision	0.888	0.887	0.887	0.886	0.885	0.884	0.872	0.858	0.828	0.732	0.810
		Recall	0.970	0.970	0.970	0.967	0.967	0.965	0.955	0.942	0.918	0.836	0.903
		F1-score	0.927	0.926	0.926	0.924	0.924	0.922	0.911	0.898	0.870	0.780	0.903
Abdallah [3]	HRNets - Mask R-CNN	Precision	0.859	0.857	0.857	0.857	0.852	0.848	0.833	0.816	0.764	0.585	0.816
		Recall	0.971	0.969	0.969	0.969	0.965	0.960	0.947	0.934	0.889	0.744	0.934
		F1-score	0.911	0.909	0.909	0.909	0.904	0.900	0.886	0.871	0.821	0.654	0.871
Abdallah [3]	HRNets - HTC	Precision	0.885	0.885	0.883	0.882	0.881	0.875	0.862	0.849	0.808	0.691	0.788
		Recall	0.987	0.987	0.984	0.984	0.982	0.976	0.966	0.954	0.915	0.816	0.901
		F1-score	0.933	0.933	0.930	0.930	0.928	0.922	0.911	0.898	0.858	0.748	0.840
Abdallah [3]	HRNets - Faster R-CNN	Precision	0.867	0.865	0.863	0.859	0.853	0.845	0.827	0.806	0.750	0.556	0.711
		Recall	0.972	0.970	0.968	0.964	0.959	0.952	0.940	0.915	0.869	0.711	0.842
		F1-score	0.916	0.914	0.912	0.908	0.902	0.895	0.879	0.857	0.805	0.624	0.770
Abdallah [3]	HRNets - Cascade R-CNN	Precision	0.893	0.891	0.891	0.891	0.888	0.880	0.871	0.854	0.831	0.705	0.799
		Recall	0.967	0.965	0.965	0.964	0.961	0.956	0.948	0.935	0.914	0.811	0.889
		F1-score	0.928	0.926	0.926	0.926	0.923	0.916	0.907	0.892	0.870	0.754	0.841
Abdallah [3]	Mask R-CNN - ResNeXt-101	Precision	0.778	0.777	0.774	0.769	0.759	0.749	0.713	0.651	0.477	0.407	0.434
		Recall	0.975	0.974	0.968	0.964	0.952	0.941	0.913	0.856	0.725	0.695	0.626
		F1-score	0.865	0.864	0.860	0.855	0.844	0.834	0.800	0.739	0.575	0.513	0.512
Abdallah [3]	Faster R-CNN - ResNeXt-101	Precision	0.884	0.884	0.880	0.879	0.876	0.871	0.856	0.833	0.780	0.581	0.733
		Recall	0.972	0.970	0.969	0.967	0.965	0.961	0.950	.931	0.884	0.724	0.848
		F1-score	0.925	0.925	0.922	0.920	0.918	0.913	0.900	0.879	0.828	0.644	0.786
Q Ren [33]	YOLOv5 + ResDCN GAM + ASFF	Precision	-	-	-	-	-	0.953	-	0.949	-	-	0.951
		Recall	-	-	-	-	-	0.915	-	0.913	-	-	0.914
		F1-score	-	-	-	-	-	0.934	-	0.931	-	-	0.9325
Our model	HTTD	Precision	**0.997**	**0.997**	**0.997**	**0.996**	**0.996**	**0.987**	**0.985**	**0.961**	**0.891**	**0.747**	**0.9554**
		Recall	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.996**	**0.994**	**0.981**	**0.933**	**0.830**	**0.9734**
		F1-score	**0.9985**	**0.9985**	**0.9985**	**0.9980**	**0.9980**	**0.9915**	**0.9895**	**0.9709**	**0.9115**	**0.7863**	**0.9643**

Table 2 compares our model which achieves an average precision of 0.9314, while the CDeC-Net model by F Yang has an average precision of 0.928. This indicates that our model performs slightly better in terms of average precision for table detection in TabRecSet.

Table 2. Table detection for TabRecSet (with the best values highlighted in bold).

Approach	Method	Metric	IoU										
			10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	Avg
F Yang [39]	CDeC-Net	Precision	-	-	-	-	-	-	-	-	-	-	0.928
		Recall	-	-	-	-	-	-	-	-	-	-	-
		F1-score	-	-	-	-	-	-	-	-	-	-	-
Our Model	HTTD	Precision	**0.969**	**0.967**	**0.965**	**0.961**	**0.953**	**0.948**	**0.940**	**0.921**	**0.874**	**0.816**	**0.9314**
		Recall	**0.999**	**0.999**	**0.998**	**0.997**	**0.996**	**0.994**	**0.992**	**0.981**	**0.947**	**0.899**	**0.9802**
		F1-score	**0.9838**	**0.9827**	**0.9812**	**0.9787**	**0.9740**	**0.9705**	**0.9653**	**0.9501**	**0.9090**	**0.8555**	**0.9554**

Table 3 extends the analysis to compare our model with other models like Cascade mask R-CNN HRNet, Cascade mask R-CNN, object detection networks, fully convolutional network, Vanilla Transformer architecture, Faster R-CNN, and Pre-trained Transformer + DCC, using various backbones. Notably, our model with a Swin-L backbone significantly outperforms other models, particularly at higher IoU thresholds, which is critical for accurate and reliable table detection. For example, at an IoU threshold of 60%, our model scores 0.972 in precision, surpassing the Cascade mask R-CNN model (97.7%), object detection networks (96.0%), and fully convolutional network (91.0%). We performed *t*-tests

to assess the statistical significance of our model comparisons, with p-values indicating significant differences.

Table 3. Table detection with ICDAR-2019 cTDaR Modern dataset (with the best values highlighted in bold).

Approach	Method	Metric	IoU										
			10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	Avg
D Prasad [5]	CascadeTabNet	Precision	-	-	-	-	-	-	-	-	-	-	-
		Recall	-	-	-	-	-	-	-	-	-	-	-
		F1-score	-	-	-	-	-	94.3	93.4	92.5	90.1	-	0.901
M Agarwal [28]	Cascade mask R-CNN	Precision	98.7	-	98.0	-	97.7	-	97.1	-	93.4	-	-
		Recall	94.6	-	93.9	-	93.6	-	93.0	-	89.5	-	-
		F1-score	96.6	-	95.9	-	95.6	-	95.0	-	91.5	-	-
X Zheng [29]	Object detection networks	Precision	-	-	-	-	-	-	96.0	-	90.0	-	-
		Recall	-	-	-	-	-	-	95.0	-	89.0	-	-
		F1-score	-	-	-	-	-	-	94.0	-	94.0	-	-
DD Nguyen [30]	Fully convolutional network	Precision	-	-	-	-	-	-	-	-	-	-	-
		Recall	-	-	-	-	-	-	-	-	-	-	-
		F1-score	-	-	92.8	-	91.7	-	91.0	-	87.4	-	-
C Ma [43]	Faster R-CNN	Precision	-	-	98.4	-	98.2	-	97.7	-	95.0	-	-
		Recall	-	-	94.0	-	93.9	-	93.3	-	90.8	-	-
		F1-score	-	-	96.1	-	96.0	-	95.4	-	92.9	-	-
J Li [31]	Dit	Precision	-	-	-	-	-	-	-	-	-	-	-
		Recall	-	-	-	-	-	-	-	-	-	-	-
		F1-score	-	-	-	-	-	95.7	95.0	94.4	91.4	-	93.9
Y Ni [35]	Pre-trained Transformer + (DCC)	Precision		-	-	-	-	-	-	-	-	-	-
		Recall	-	-	-	-	-	-	-	-	-	-	-
		F1	-	-	-	-	-	97.3	97.1	96.4	92.2	-	95.5
Our model	HTTD	Precision	**0.981**	**0.981**	**0.980**	**0.980**	**0.972**	**0.972**	**0.966**	**0.950**	**0.932**	**0.882**	**0.9596**
		Recall	**0.993**	**0.993**	**0.993**	**0.993**	**0.990**	**0.990**	**0.987**	**0.977**	**0.963**	**0.924**	**0.9803**
		F1-score	**0.9870**	**0.9870**	**0.9865**	**0.9865**	**0.9809**	**0.9809**	**0.9764**	**0.9633**	**0.9472**	**0.9025**	**0.9698**

When we look at how well different models perform at an IoU threshold of 80%, it is clear that our model is doing something right. It manages to score 0.950 in precision, while most other models do not even get on the scoreboard. This tells us that our model is getting better at spotting and pinpointing the location of tables. Our model shows impressive results at higher IoU thresholds. For instance, at an IoU of 60%, our model scores 0.972 in precision. This score tells us that our model is good at pinpointing the exact location of tables, which is super important for detecting them accurately. This is a big step up from the next best model, Faster R-CNN, which only scores 97.7% at the same IoU level. Furthermore, even when the IoU thresholds get tougher, like 80%, our model still scores 0.950, showing its precision.

4.4. Statistical Analysis of Performance Differences

To ensure that the observed differences between the HTTD model and other methods are statistically significant, paired t-tests were conducted on precision, recall, and F1-score results across various IoU thresholds. The significance level was set to $\alpha = 0.05$.

The results of the paired t-tests comparing HTTD with other models (e.g., HRNets Cascade Mask R-CNN, Mask R-CNN ResNeXt-101, YOLOv5 + ResDCN GAM + ASFF) are summarized in Table 4. Across all IoU thresholds, HTTD exhibited statistically significant improvements in precision, recall, and F1-score ($p < 0.05$).

Table 4. Paired *t*-tests results for HTTD vs. other methods.

Metric	IoU Threshold	HTTD Mean	Comparison Model	Model Mean	*p*-Value	Significance
Precision	60%	0.985	HRNets Cascade Mask R-CNN	0.882	1.52×10^{-5}	Yes
Precision	80%	0.960	YOLOv5 + ResDCN GAM + ASFF	0.945	2.56×10^{-5}	Yes
F1-score	80%	0.9704	Mask R-CNN ResNeXt-101	0.734	2.02×10^{-7}	Yes
Recall	90%	0.930	HRNets Faster R-CNN	0.881	2.40×10^{-5}	Yes

5. Ablation Study

To comprehensively evaluate the HTTD model, we performed two sets of ablation studies. The first focuses on the contribution of individual components to the performance of table detection on the ICDAR-2019 cTDaR Modern dataset. The second examines the generalizability of the methodology to other detection tasks, including receipt item detection and cancer region detection.

5.1. Ablative Analysis of Model Features

To assess the contributions of individual components in the HTTD model, we conducted an ablation study on the ICDAR-2019 cTDaR Modern dataset. This study focuses on evaluating the model's performance in terms of precision, recall, and F1-score across various IoU thresholds. The impact of each key feature—contrastive denoising, mixed query selection, look-forward-twice refinement, and the Swin-L backbone—is analyzed. The contrastive denoising (CDN) stabilizes training and enhances the model's ability to focus on high-quality anchors, reducing redundant predictions and improving overall detection accuracy. Adding CDN to the baseline model resulted in a precision increase from 0.885 to 0.921 (+3.6%), a recall increase from 0.910 to 0.944 (+3.4%), and an F1-score increase from 0.897 to 0.932 (+3.5%). These improvements highlight CDN's effectiveness in handling complex table layouts. Mixed query selection (MQS) refines the initialization of positional queries, leveraging spatial priors to enhance localization performance. Incorporating MQS further improved the model's precision to 0.938 (+1.7%), recall to 0.955 (+1.1%), and F1-score to 0.944 (+1.2%). These gains illustrate MQS's ability to optimize the model's focus on contextually relevant anchors. The look-forward-twice (LFT) refinement enhances bounding box predictions by leveraging gradient updates across consecutive layers. Adding LFT increased precision to 0.950 (+1.2%), recall to 0.969 (+1.4%), and F1-score to 0.959 (+1.5%), demonstrating its value in improving localization accuracy at higher IoU thresholds.

Replacing the baseline backbone with Swin-L provided significant improvements due to its ability to capture multiscale features and global context. This final addition raised precision to 0.960 (+1.0%), recall to 0.977 (+0.8%), and F1-score to 0.963 (+0.4%), achieving the highest overall performance. The full HTTD model, combining all components, achieved the highest performance on the ICDAR-2019 cTDaR Modern dataset, with an average F1-score of 0.9633, precision of 0.950, and recall of 0.977. These results confirm the complementary contributions of the individual components.

Table 5 provides a comparison of precision, recall, and F1-score for the full model (with all components) and when specific components are removed. Results are reported at 80% IoU.

Table 5. Ablation study results on ICDAR-2019 cTDaR Modern dataset (IoU = 80%).

Configuration	Precision	Recall	F1-Score
Original model	0.885	0.910	0.897
Add contrastive denoising (CDN)	0.921 (+3.6%)	0.944 (+3.4%)	0.932 (+3.5%)
Add mixed query selection (MQS)	0.938 (+1.7%)	0.955 (+1.1%)	0.944 (+1.2%)
Add look forward twice (LFT)	0.950 (+1.2%)	0.969 (+1.4%)	0.959 (+1.5%)
With Swin-L backbone (full model)	0.960 (+1.0%)	0.977 (+0.8%)	0.963 (+0.4%)
Full HTTD model	0.950	0.977	0.9633

5.2. Experimental Results Across Diverse Detection Domains

To evaluate the versatility and generalization potential of the HTTD model, we conducted an additional ablation study on a different detection tasks (breast cancer detection and receipt key detection).

5.2.1. Datasets and Experiment Setup

For the breast cancer detection task, we used a dataset comprising 12,476 annotated mammographic images [44]. These images were meticulously labeled following BI-RADS standards and underwent expert-led data cleaning and region-of-interest (ROI) extraction. The dataset was split into training (90%; 11,228 images) and validation (10%; 1248 images) sets.

The CORU dataset [4], consisting of 20,000 annotated receipts in Arabic and English, was utilized for the receipt key detection task. It focused on detecting critical entities such as merchant names, dates, receipt numbers, items, and total prices. The dataset was divided into training (90%; 18,000 receipts) and validation (10%; 2000 receipts) sets.

To enhance model performance, images were resized and augmented during training, with their dimensions increased by 1.5 times. The training process utilized the AdamW optimizer [40,41] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. Loss functions included L1 and Generalized Intersection over Union (GIoU) [42] for box regression, as well as focal loss (parameters: $\alpha = 0.25, \gamma = 2$) for classification. This configuration aimed to refine bounding box predictions and boost classification accuracy. The localization performances of various models at different Intersection over Union (IoU) thresholds are shown in Tables 6 and 7.

5.2.2. Breast Cancer Detection Result

The HTTD model demonstrated remarkable performance in detecting regions of interest in mammographic images. The integration of hierarchical feature extraction, contrastive denoising, and mixed query selection contributed to its superior results. As shown in Table 6, our approach outperformed conventional methods, achieving an average localization accuracy of 56.38% across varying IoU thresholds. Notably, the use of the Swin-L backbone enabled the model to maintain high precision and recall even at stricter IoU thresholds, underscoring its robustness in complex medical imaging scenarios.

Table 6. Performance comparison of different models for breast cancer detection (BCD) across IoU thresholds(with the best values highlighted in bold).

Model	Backbone	IOU								
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**Avg**
HAS	Resnet50	24.00	18.00	10.66	9.00	6.39	3.79	1.18	0.23	9.41
	VGG-16	30.2	25.5	15.43	10.28	5.65	5.01	2.82	0.57	12.68
CAM	Resnet50	65.30	58.29	46.68	37.91	29.85	21.09	9.47	1.65	33.15
	VGG-16	61.40	54.97	40.75	29.62	21.09	11.84	4.73	1.42	28.23
ACOL	Resnet50	39.67	31.56	25.11	19.43	12.55	8.53	4.02	0.94	17.98
	VGG-16	36.78	30.46	20.68	13.34	10.04	6.18	2.13	0.56	15.40
SPG	Resnet50	35.20	28.87	21.72	14.25	9.45	5.40	1.23	0.0	14.27
	VGG-16	52.36	48.58	29.89	20.00	9.69	3.15	0.84	0.0	20.56
ADL	Resnet50	72.10	68.72	55.45	44.31	33.64	24.64	14.45	5.68	39.62
	VGG-16	48.40	43.60	24.17	12.08	5.92	2.13	0.94	0.0	17.78
Our approach	**Swin-L**	**78.6**	**63.2**	**62.4**	**61.1**	**59.5**	**56.4**	**46.3**	**22.6**	**56.38**

5.2.3. Receipt Key Detection Result

For the receipt key detection task, the HTTD model outperformed existing methods, including ACOL, HAS, SPG, and DINO, as detailed in Table 7. By leveraging its Transformer-based architecture, the model achieved an average IoU of 36.5%, significantly surpassing other approaches, particularly at higher IoU thresholds. This highlights its ability to accurately localize and classify key components within multilingual receipts, even in noisy and cluttered layouts.

Table 7. Performance(with the best values highlighted in bold). comparison of different methods for receipt key detection across IoU thresholds.

Method	Backbone	Avg	IoU								
			10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
ACOL	ResNet50	3.94	27.49	7.19	2.72	1.01	0.48	0.32	0.21	0.00	0.00
	VGG16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAS	ResNet50	7.14	43.26	16.84	6.55	2.72	1.17	0.59	0.21	0.11	0.00
	VGG16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SPG	ResNet50	6.53	42.14	13.00	5.06	2.66	1.33	0.64	0.27	0.11	0.05
Cutmix	ResNet50	6.32	41.61	13.37	5.01	1.97	0.69	0.37	0.16	0.05	0.00
	VGG16	5.54	38.47	11.45	3.25	1.28	0.59	0.27	0.11	0.00	0.00
ADL	ResNet50	6.57	39.74	15.13	6.34	2.56	1.17	0.48	0.16	0.11	0.05
	VGG16	6.97	47.52	15.34	4.69	1.60	0.37	0.16	0.05	0.00	0.00
CAM	ResNet50	6.17	41.08	11.93	4.48	2.24	1.07	0.59	0.21	0.05	0.05
	VGG16	5.86	45.23	10.12	2.24	0.64	0.16	0.11	0.05	0.00	0.00
DINO	Swin	32.2	45.4	44.6	43.3	41.9	39.9	35.9	27.5	10.6	0.7
	ResNet50	31.9	45.9	45.0	43.6	41.9	39.4	35.2	25.6	10.2	0.5
Our approach	**Swin-L**	**36.5**	**50.0**	**49.6**	**47.8**	**46.5**	**44.1**	**39.8**	**31.6**	**14.3**	**1.2**

5.3. Case Study

To provide a comprehensive analysis of the model's performance, we include qualitative examples showcasing both successful detections and failure cases. Figure 6 illustrates four examples: two successful cases (a and b), where the HTTD model accurately detected tables in clean and well-structured documents, and two failure cases (c and d), where the model struggled due to noisy or partially cropped input data.

In the failure cases, while the model successfully identified the presence of tables, it was unable to accurately delineate the boundaries due to significant noise, text occlusion, or incomplete information in the document. These results highlight the challenges faced by HTTD in handling degraded or incomplete inputs, emphasizing the need for further advancements in robust feature extraction under adverse conditions.

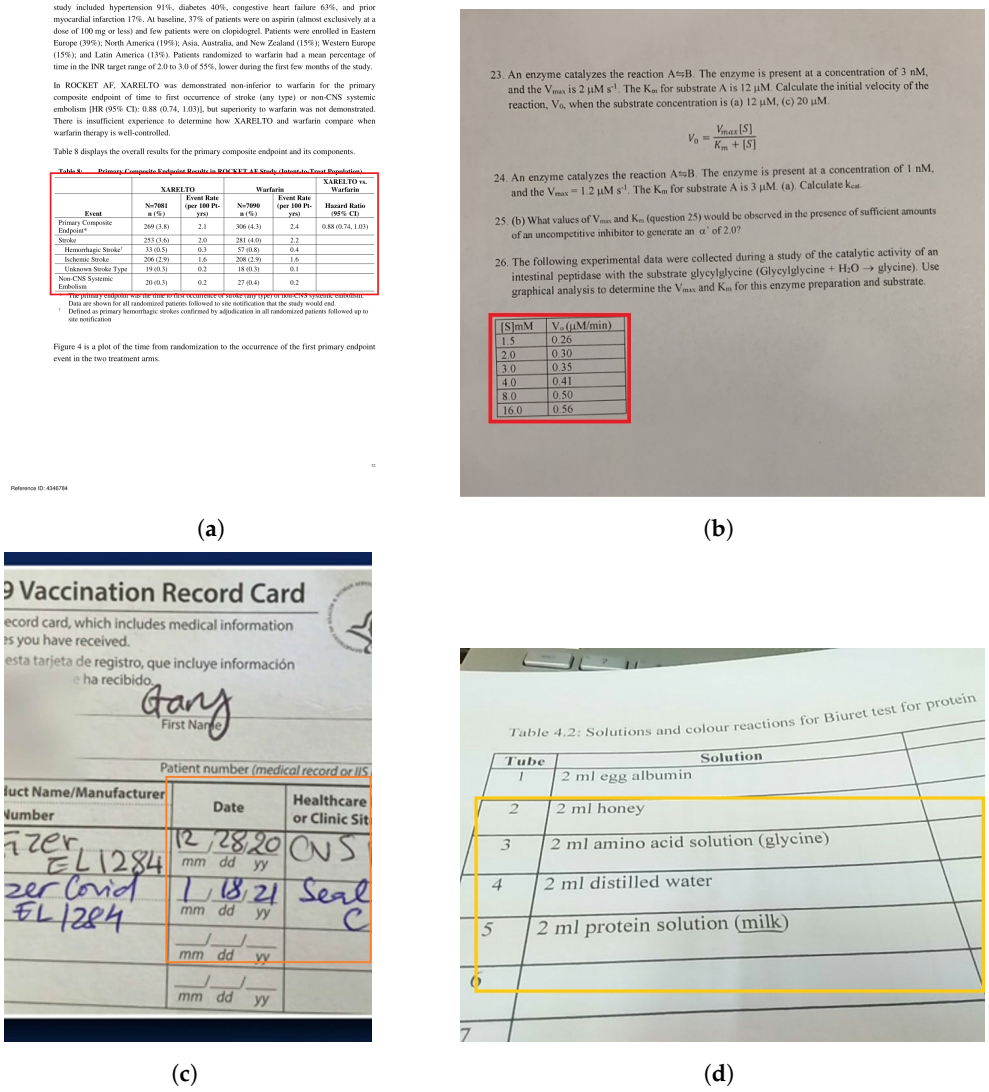


Figure 6. Qualitative results of our model: (**a**,**b**) Successful detections in clean and moderately complex layouts; (**c**,**d**) Failures due to noisy and cropped inputs.

6. Conclusions and Future Work

This study introduces HTTD, a Hierarchical Transformer designed to address the key challenges of table detection in diverse document layouts, including historical and modern documents. By improving computational efficiency, enhancing training speed, and generalizing to non-standard tasks such as breast cancer detection and receipt key detection, HTTD demonstrates its versatility and adaptability. Our innovations, including contrastive denoising and mixed query selection, establish a new standard for table detection and open pathways for broader applications in document analysis. Future work will focus on extending HTTD's capabilities to include semi-supervised learning and content recognition for even greater generalization.

In future work, we aim to explore hybrid architectures that further integrate CNN-based local feature extraction with Transformer-based global context modeling. Additionally, leveraging semi-supervised and unsupervised learning methods could unlock the

potential of unlabeled data, enhancing the model's generalization across diverse document types. Finally, expanding HTTD's scope to include table structure recognition and content extraction will position it as a comprehensive solution for document processing tasks.

Author Contributions: Conceptualization, M.S.K., M.M. and B.Y.; Methodology, M.S.K., M.M. and M.A.; Software, M.S.K., B.Y., M.F.S. and M.A.; Validation, M.S.K. and H.-S.K.; Formal analysis, M.S.K. and M.F.S.; Investigation, H.-S.K.; Resources, H.-S.K.; Data curation, M.S.K.; Writing—original draft, M.S.K.; Writing—review & editing, M.S.K. and H.-S.K.; Visualization, H.-S.K.; Supervision, H.-S.K.; Project administration, H.-S.K.; Funding acquisition, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (No. 2023R1A2C1006944, 50%) and partly by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201462, 50%).

Data Availability Statement: Data are contained within the article.

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

References

1. Riba, P.; Goldmann, L.; Terrades, O.R.; Rusticus, D.; Fornés, A.; Lladós, J. Table detection in business document images by message passing networks. *Pattern Recognit.* **2022**, *127*, 108641. [CrossRef]
2. Ngubane, T.; Tapamo, J.R. TableExtractNet: A Model of Automatic Detection and Recognition of Table Structures from Unstructured Documents. *Informatics* **2024**, *11*, 77. [CrossRef]
3. Salaheldin Kasem, M.; Abdallah, A.; Berendeyev, A.; Elkady, E.; Mahmoud, M.; Abdalla, M.; Hamada, M.; Vascon, S.; Nurseitov, D.; Taj-Eddin, I. Deep learning for table detection and structure recognition: A survey. *ACM Comput. Surv.* **2024**, *56*, 1–41. [CrossRef]
4. Abdallah, A.; Abdalla, M.; Kasem, M.S.; Mahmoud, M.; Abdelhalim, I.; Elkasaby, M.; ElBendary, Y.; Jatowt, A. CORU: Comprehensive Post-OCR Parsing and Receipt Understanding Dataset. *arXiv* **2024**, arXiv:2406.04493.
5. 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, Seattle, WA, USA, 14–19 June 2020; pp. 572–573.
6. Menghani, G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
7. Yang, Y.; Xia, X.; Lo, D.; Grundy, J. A survey on deep learning for software engineering. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–73. [CrossRef]
8. Farouk, M. Measuring text similarity based on structure and word embedding. *Cogn. Syst. Res.* **2020**, *63*, 1–10. [CrossRef]
9. Abdallah, A.; Kasem, M.; Abdalla, M.; Mahmoud, M.; Elkasaby, M.; Elbendary, Y.; Jatowt, A. Arabicaqa: A comprehensive dataset for arabic question answering. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 2049–2059.
10. Mahmoud, S.A.; Ahmad, I.; Al-Khatib, W.G.; Alshayeb, M.; Parvez, M.T.; Märgner, V.; Fink, G.A. KHATT: An open Arabic offline handwritten text database. *Pattern Recognit.* **2014**, *47*, 1096–1112. [CrossRef]
11. Toiganbayeva, N.; Kasem, M.; Abdimanap, G.; Bostanbekov, K.; Abdallah, A.; Alimova, A.; Nurseitov, D. Kohtd: Kazakh offline handwritten text dataset. *Signal Process. Image Commun.* **2022**, *108*, 116827. [CrossRef]
12. Kanellos, N.; Terzi, M.C.; Giannakopoulos, N.T.; Karountzos, P.; Sakas, D.P. The Economic Dynamics of Desktop and Mobile Customer Analytics in Advancing Digital Branding Strategies: Insights from the Agri-Food Industry. *Sustainability* **2024**, *16*, 5845. [CrossRef]
13. Kasem, M.S.; Hamada, M.; Taj-Eddin, I. Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Comput. Appl.* **2024**, *36*, 4995–5005. [CrossRef]
14. Kasem, M.S.; Mahmoud, M.; Kang, H.S. Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey. *arXiv* **2023**, arXiv:2312.11812.
15. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
16. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.

17. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9640–9649.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
19. Ali, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. Xcit: Cross-covariance image transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20014–20027.
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
21. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
22. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3651–3660.
23. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
24. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv* **2022**, arXiv:2201.12329.
25. Gilani, A.; Qasim, S.R.; Malik, I.; Shafait, F. Table detection using deep learning. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 771–776.
26. Ren, S. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497. [CrossRef]
27. Samari, A.; Piper, A.; Hedley, A.; Cheriet, M. Weakly supervised bounding box extraction for unlabeled data in table detection. In Proceedings of the Pattern Recognition, ICPR International Workshops and Challenges, Virtual, 10–15 January 2021; Proceedings, Part VII; Springer: Cham, Switzerland, 2021; pp. 339–352.
28. Agarwal, M.; Mondal, A.; Jawahar, C. Cdec-net: Composite deformable cascade network for table detection in document images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9491–9498.
29. Zheng, X.; Burdick, D.; Popa, L.; Zhong, X.; Wang, N.X.R. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 697–706.
30. Nguyen, D.D. TableSegNet: A fully convolutional network for table detection and segmentation in document images. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2022**, *25*, 1–14. [CrossRef]
31. Li, J.; Xu, Y.; Lv, T.; Cui, L.; Zhang, C.; Wei, F. DiT: Self-supervised Pre-training for Document Image Transformer. *arXiv* **2022**, arXiv:2203.02378.
32. Haloi, M.; Shekhar, S.; Fande, N.; Dash, S.S. Table Detection in the Wild: A Novel Diverse Table Detection Dataset and Method. *arXiv* **2022**, arXiv:2209.09207.
33. Ren, Q.; Ibrayim, M.; Hamdulla, A. Table Detection in Complex Layouts. In Proceedings of the 2023 China Automation Congress (CAC), Chongqing, China, 17–19 November 2023; pp. 9172–9177.
34. Shehzadi, T.; Azeem Hashmi, K.; Stricker, D.; Liwicki, M.; Zeshan Afzal, M. Towards End-to-End Semi-Supervised Table Detection with Deformable Transformer. In Proceedings of the International Conference on Document Analysis and Recognition, San Jose, CA, USA, 21–23 August 2023; Springer: Cham, Switzerland, 2023; pp. 51–76.
35. Ni, Y.; Wang, X.; Peng, H.; Li, Y.; Wang, J.; Li, H.; Huang, J. Dual-branch dilated context convolutional for table detection transformer in the document images. *Vis. Comput.* **2024**. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. 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 Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 1510–1515.
38. Abdallah, A.; Berendeyev, A.; Nuradin, I.; Nurseitov, D. Tncr: Table net detection and classification dataset. *Neurocomputing* **2022**, *473*, 79–97. [CrossRef]
39. Yang, F.; Hu, L.; Liu, X.; Huang, S.; Gu, Z. A large-scale dataset for end-to-end table recognition in the wild. *Sci. Data* **2023**, *10*, 110. [CrossRef]
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

41. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
42. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
43. Ma, C.; Lin, W.; Sun, L.; Huo, Q. Robust Table Detection and Structure Recognition from Heterogeneous Document Images. *Pattern Recognit.* **2023**, *133*, 109006. [CrossRef]
44. Abdelhalim, I.; Alksas, A.; Balaha, H.M.; Badawy, M.A.; Abou El-Ghar, M.; Alghamdi, N.S.; Ghazal, M.; Contractor, S.; Van Bogaert, E.; Gondim, D.; et al. Incorporating Imaging, Clinical, Pathology, and Demographic Markers for Hormonal Therapy Prediction in Prostate Cancer. *IEEE Access* **2024**, *12*, 195960–195973. [CrossRef]