

Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction

Qintong Zhang^{1,2*}, Victor Shea-Jay Huang^{2*}, Bin Wang^{1*}, Junyuan Zhang¹,
Zhengren Wang², Hao Liang², Shawn Wang³, Matthieu Lin³, Conghui He^{1†}, Wentao Zhang²

¹Shanghai Artificial Intelligence Laboratory; ²Peking University; ³Tsinghua University;
zhangqintong@pjlab.org.cn, jeix782@gmail.com

Abstract

Document parsing is essential for converting unstructured and semi-structured documents—such as contracts, academic papers, and invoices—into structured, machine-readable data. Document parsing extract reliable structured data from unstructured inputs, providing huge convenience for numerous applications. Especially with recent achievements in Large Language Models, document parsing plays an indispensable role in both knowledge base construction and training data generation. This survey presents a comprehensive review of the current state of document parsing, covering key methodologies, from modular pipeline systems to end-to-end models driven by large vision-language models. Core components such as layout detection, content extraction (including text, tables, and mathematical expressions), and multi-modal data integration are examined in detail. Additionally, this paper discusses the challenges faced by modular document parsing systems and vision-language models in handling complex layouts, integrating multiple modules, and recognizing high-density text. It emphasizes the importance of developing larger and more diverse datasets and outlines future research directions.

1 Introduction

As digital transformation accelerates, electronic documents have increasingly supplanted paper documents as the primary medium for information exchange across various industries. This shift has significantly expanded the diversity and complexity of document types, including contracts, invoices, and academic papers. Consequently, there is an escalating need for efficient systems to manage and retrieve information [1, 2]. However, a substantial proportion of historical records, academic publications, and legal documents remain in scanned or image-based formats, posing considerable challenges to tasks such as information extraction, document comprehension, and enhanced retrieval [3–5].

To address these challenges, document parsing (DP), also known as document content extraction, has emerged as an essential tool for converting unstructured and semi-structured documents into structured information. Document parsing recognizes and extracts various elements such as text, equations, tables, and images from various document inputs while preserving their structural relationships. The extracted content is then transformed into structured formats like Markdown or JSON, enabling seamless integration into modern workflows [6].

Document parsing is critical for document-related tasks, reshaping how information is stored, shared, and applied across numerous applications. It provides a foundation for various downstream processes, including the development of Retrieval-Augmented Generation(RAG) systems in various practical

*Equal Contribution.

†Corresponding Authors.

fields, the automated construction of electronic storage and retrieval libraries for paper materials [7–10]. Besides that, there are plenty of potential information in document which remains largely underdeveloped. Document parsing technology can effectively extract and organize these rich knowledge, laying a solid foundation for the development of the next generation of intelligent systems. For example, training more professional and powerful multimodal models [5, 11].

However, recent years have seen significant advancements in document parsing technologies, particularly those based on deep learning, leading to the proliferation of document parsing tools and the emergence of promising document parsers. Nevertheless, research in this field still faces certain limitations. Many surveys on document parsing are outdated, resulting in pipelines that lack rigor and comprehensiveness, with technological descriptions failing to capture recent advancements and changes in application scenarios [3, 4]. Additionally, high-quality reviews often focus on specific sub-technologies within document parsing, such as layout analysis [12–14], mathematical expression recognition [15–17], table structure recognition [18–20], and chart-related work in documents [21], without providing a comprehensive overview of the entire document parsing process.

Given these limitations, a comprehensive review of document parsing is urgently needed. In this survey, we analyze advancements in document parsing from a holistic perspective, providing researchers and developers with a broad understanding of recent developments and future directions in the field. The key contributions of this survey are as follows:

- **Comprehensive Review of Document Parsing.** This paper systematically integrates and evaluates recent advancements in document parsing technologies across the stages of the document parsing pipeline.
- **Consolidation of Datasets and Evaluation Metrics.** We consolidate widely used datasets and evaluation metrics, addressing gaps in existing reviews within the document parsing field.
- **Holistic Insight for Researchers and Practitioners.** This work provides a holistic perspective on the current state and future directions of document parsing, bridging the gap between academic research and practical applications.
- **Introductory Guide for Newcomers.** It serves as a guide for newcomers to quickly understand the field’s landscape and identify promising research directions.

The organization of this article is as follows: Section 2 outlines two main approaches to document parsing. From Section 3 to Section 6.4 examines key algorithms used in modular document parsing systems. Section 7 presents visual language macromodels applicable to document-related tasks, with a focus on document parsing and OCR. Section 8 and Section 9 covers datasets and evaluation metrics in document parsing. In Section 11, we discuss current challenges in the field and highlight significant future directions. Finally, Section 12 provides a concise and insightful conclusion.

2 Methodology

Document parsing can be broadly categorized into two methodologies: the modular pipeline document parsing system and the end-to-end approach based on large vision-language models.

2.1 Document Parsing System

2.1.1 Layout Analysis

Layout detection identifies the structural elements of a document—such as text blocks, paragraphs, headings, images, tables, and mathematical expressions—along with their spatial coordinates and reading order. This foundational step is crucial for ensuring accurate content extraction. Notably, the detection of mathematical expressions, especially inline ones, is often handled separately due to their complexity.

2.1.2 Content Extraction

- **Text Extraction:** This process leverages Optical Character Recognition (OCR) technology to convert document images into machine-readable text. By analyzing the shapes and patterns of characters, OCR accurately recognizes and processes the text contained within the images.

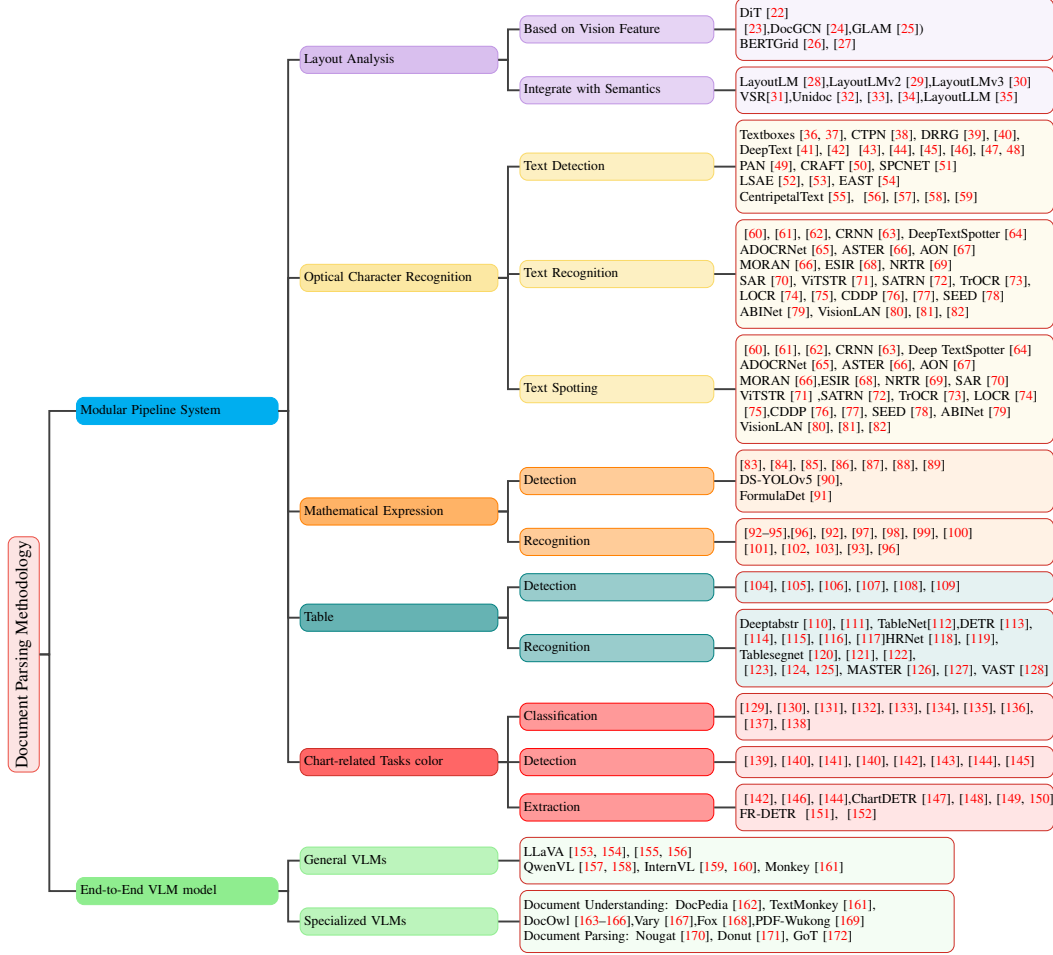


Figure 1: Overview of Document Parsing Methodology.

- **Mathematical Expression Extraction:** In this step, mathematical symbols and structures within document regions are detected and converted into standardized formats such as LaTeX or MathML. Due to the complexity of the symbols and their spatial arrangements, this task presents a unique challenge.
- **Table Data and Structure Extraction:** Table recognition involves detecting and interpreting table structures by identifying the layout of cells and the relationships between rows and columns in document images. The extracted table data is typically combined with OCR results and converted into formats like LaTeX for further use.
- **Chart Recognition:** This step focuses on identifying different types of charts and extracting the underlying data as well as their structural relationships. The visual information from charts is converted into raw data tables or structured formats like JSON.

2.1.3 Relation Integration

Each step builds upon the previous, ensuring a seamless flow from text to mathematical expressions, tables, and charts, all while leveraging advanced recognition technologies to convert document content into structured, machine-readable formats. Once individual content elements are extracted, relation integration combines them into a unified structure. This step uses the spatial coordinates identified during layout detection, ensuring that the spatial and semantic relationships between elements are preserved. Rule-based systems or specialized reading order models are commonly applied to maintain the logical flow of content.

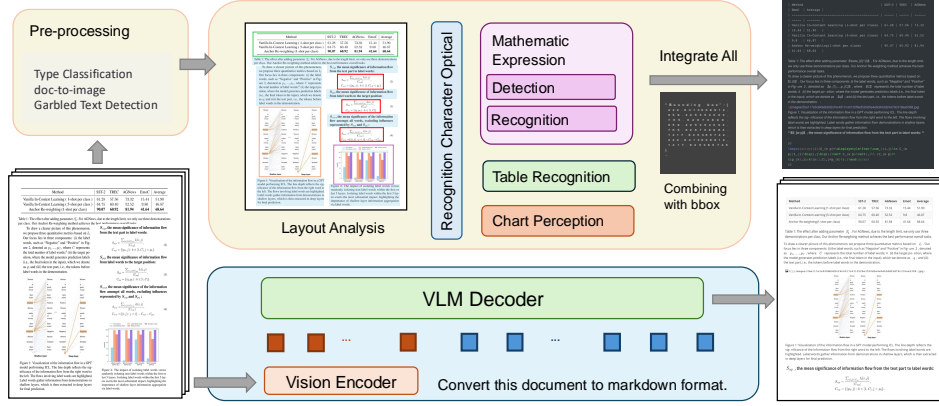


Figure 2: Two Methodology of Document Parsing

2.2 End-to-End Approaches and Multimodal Large Models

While traditional modular document parsing systems perform effectively within specific domains, their architecture often leads to limitations in joint optimization and generalization across diverse document types. Recent advances in multimodal large models, particularly vision-language models (VLMs), offer promising alternatives. Models such as GPT-4, Qwen, LLaMA, and InternVL can simultaneously process visual and textual data, facilitating end-to-end conversion of document images into structured outputs. Due to the unique challenges posed by document images—such as dense text, complex layouts, and high variability in visual elements—specialized large models like Nougat, Fox, and GOT have emerged. These models represent a significant leap forward in automating document parsing and comprehension.

3 Layout Analysis

3.1 Introduction to Layout Analysis Technology

Research on document layout analysis (DLA) for scanned images began in the 1990s. Early studies focused on simple document structures, often as a preprocessing step, and primarily used rule-based methods [173–186, 71, 187–196] or statistical techniques [197, 198, 178].

By the 2000s, DLA incorporated feature engineering and machine learning, framing the task as pixel-based semantic segmentation [199–201]. Since 2015, deep learning techniques, particularly convolutional neural networks (CNNs) and Transformers, have dominated the field, treating DLA as a pixel-wise segmentation problem and leveraging visual features to analyze physical layouts [202–207, 22].

Moreover, graph convolutional networks (GCNs) have been employed to model relational representations between document components [208, 23, 31, 209, 24, 25]. Grid-based approaches [210–212, 26, 27] have emphasized the importance of preserving spatial structures. Recent studies have also integrated multiple data sources into these models [28, 33, 31, 34, 32, 29, 30]. Around 2020, self-supervised pretraining in multimodal natural language processing (NLP) influenced DLA research, leading to models that jointly integrate text and visual layout information for end-to-end learning.

3.2 Based on Visual Feature

Early deep learning-based DLA primarily focused on analyzing physical layouts using visual features from document images. Documents were treated as images, with elements such as text blocks, images, and tables detected and extracted through neural network architectures [202].

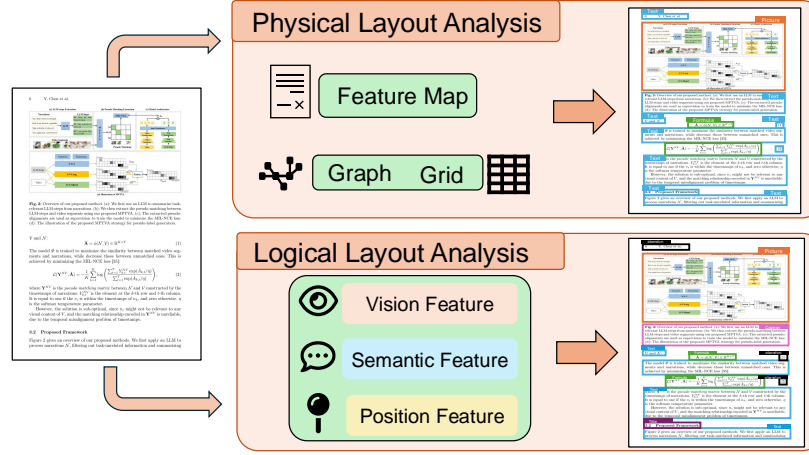


Figure 3: Overview of the DLA Algorithm

3.2.1 CNN-based Methods

The introduction of Convolutional Neural Networks (CNNs) marked a significant advancement in DLA. Originally designed for object detection, these models were adapted for tasks like page segmentation and layout detection. R-CNN, Fast R-CNN, and Mask R-CNN were especially influential for detecting components such as text blocks and tables [203]. Later studies improved the region proposal process and architecture for enhanced page object detection [204]. Models like Fully Convolutional Networks (FCN) and ARU-net were developed to handle more complex layouts [205, 206].

3.2.2 Transformer-based Methods

Recent advances in Transformer models have extended their application in DLA. BEiT (Bidirectional Encoder Representation from Image Transformers), inspired by BERT, employs self-supervised pretraining to learn robust image representations, excelling at extracting global document features such as titles, paragraphs, and tables [207]. The Document Image Transformer (DiT), with its Vision Transformer (ViT)-like architecture, splits document images into patches to enhance layout analysis. However, these models are computationally intensive and require extensive pretraining [22]. Recent work such as [213, 214] also focuses on using transformers to complete classification tasks based on document visual features.

3.2.3 Graph-based Methods

While image-based approaches have significantly advanced DLA, they often rely heavily on visual features, limiting their understanding of semantic structures. Graph Convolutional Networks (GCNs) address this issue by modeling relationships between document components, enhancing the semantic analysis of layouts [208, 23, 31]. For instance, Doc-GCN improves understanding of semantic and contextual relationships among layout components [24]. GLAM, another prominent model, represents a document page as a structured graph, combining visual features with embedded metadata for superior performance [25].

3.2.4 Grid-Based Methods

Grid-based methods preserve spatial information by representing document layouts as grids, which aids in retaining spatial details [210–212, 26, 27]. For instance, BERTGrid adapts BERT to represent layouts while maintaining spatial structures [26]. The VGT model integrates Vision Transformer (ViT) and Grid Transformer (GiT) modules to capture features at both token and paragraph levels. However, grid-based methods often face challenges such as large parameter sizes and slow inference speeds, limiting their practical application [27].

3.3 Integrate with Semantic Information

As document analysis becomes more complex, physical layout analysis alone is insufficient. Although there is work that proves that excellent object detection models such as YOLO v8 are still relatively leading in the layout analysis of documents in some small languages based on graphemes [215], and relevant improvements have been made [216], DLA methods that combine semantic information are still an important development direction. Logical layout analysis is needed to classify document elements by their semantic roles, such as titles, charts, or footers. With the rise of multimodal models, methods that combine visual, textual, and layout information have gained prominence in DLA research.

Logical layout analysis, driven by the need to classify document elements based on their semantic roles, has led to the development of multimodal models that integrate text and layout information for more comprehensive analysis. Studies have explored multimodal data integration by combining supervised learning with pre-trained natural language processing (NLP) or computer vision (CV) models. For example, LayoutLM was the first model to fuse text and layout information within a single framework, using the BERT architecture to capture document features through text, positional, and image embeddings [28].

[33] extended this by combining RoBERTa with GCNs to capture relational layout information from both text and images. [31] introduced a multi-scale adaptive aggregation module to fuse visual and semantic features, producing an attention map for more accurate feature alignment.

Self-supervised pretraining in multimodal NLP has also significantly advanced the field. During pretraining, models jointly process text, images, and layout information using a unified Transformer architecture, enabling them to learn cross-modal knowledge from various document types. This approach improves model versatility, requiring minimal supervision for fine-tuning across different document types and styles.

In 2020, [34] proposed a multimodal document pre-training framework that encodes information from multi-page documents end-to-end, incorporating tasks such as document topic modeling and random document prediction. This framework enables models to learn rich representations of images, text, and layout. Notable work such as UniDoc [32] uses a Transformer and ResNet-50 architecture to extract linguistic and visual features, aligned through a gated cross-modal attention mechanism.

Advancements include LayoutLMv2 and LayoutLMv3, which refine LayoutLM by optimizing the fusion of text, image, and layout information. These models improve feature extraction through deeper multimodal interactions and masking mechanisms, achieving more efficient and comprehensive document analysis [29, 30]. In addition, LayoutLLM [35] attempts to use a large language model to integrate certain semantic information to complete tasks related to document layout.

4 Optical Character Recognition

4.1 Introduction to Document OCR

Optical Character Recognition (OCR) has a long history, originating from the early development of computers. The concept was first introduced by Tausheck in 1929. Today, OCR is a critical area of research in computer vision and pattern recognition, aiming to identify text in visual data and convert it into editable digital formats for subsequent analysis and organization.

In the 1950s and 1960s, OCR research concentrated on handwritten document recognition, such as check processing and mail sorting. During this period, OCR systems primarily utilized preprocessing techniques and rule-based or template-matching methods. For instance, early versions of ABBYY OCR employed image binarization, noise reduction, and layout analysis to identify characters through template matching.

Before the advent of deep learning, OCR systems mainly relied on feature engineering and traditional machine learning techniques for character recognition. These approaches were commonly applied to tasks like postal code recognition, form processing, and banking. A notable example is Tesseract OCR, developed by HP Labs in 1984, which used such techniques in its earlier versions (prior to version 4.x).

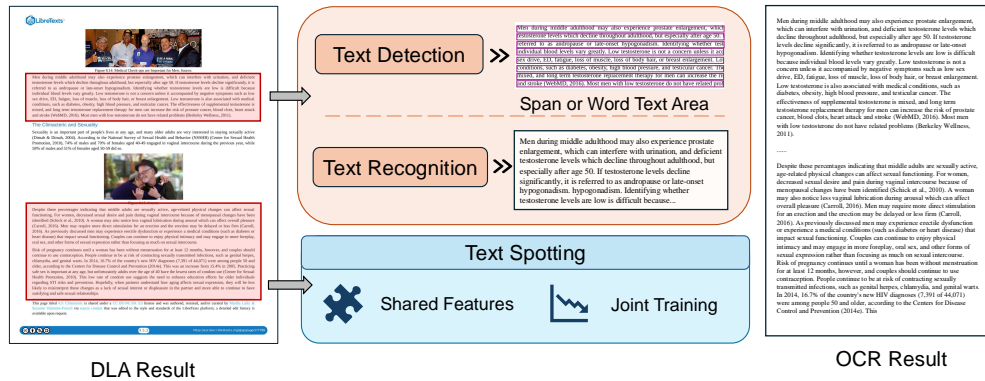


Figure 4: Overview of the OCR Algorithm

As OCR technology has been integrated into various industries, there has been increasing demand for higher accuracy and broader applicability. Researchers have since explored more advanced OCR applications, including scene text recognition, multi-language recognition, and document character recognition. Since 2010, the development of end-to-end deep learning algorithms has significantly transformed OCR, improving both its efficiency and its range of applications.

OCR generally involves two primary stages: text detection and text recognition. First, the text is localized within an image, and then recognition algorithms are applied to convert the identified text into computer-readable characters. When OCR integrates both text detection and recognition, it is referred to as text spotting. This section discusses these three crucial technical aspects of OCR.

4.2 Text Detection

Traditional, non-deep learning text detection algorithms are generally effective in simple scenes with high contrast backgrounds. However, they often require manual parameter adjustments to achieve optimal performance in different contexts, limiting their generalization capabilities. In contrast, deep learning-based text detection algorithms, which improve upon object detection and instance segmentation techniques, can be categorized into four main approaches: one-stage regression-based methods, two-stage region proposal methods, instance segmentation-based methods, and hybrid methods.

4.2.1 Regression-Based Single-Stage Methods

Regression-based methods, also called direct regression approaches, directly predict the corner coordinates or aspect ratios of text boxes from specific points in the image, bypassing the need for multi-stage candidate region generation and subsequent classification. Algorithms like YOLO and SSD have been adapted for text detection, with modifications to handle text-specific challenges, such as varied aspect ratios and orientations [36, 37]. For instance, CTPN [38] achieves precise text line localization through regression of vertical positions and lateral offsets. Methods such as SegLink [217] and DRRG [39] apply regression techniques to handle irregular text shapes, while Fourier transforms [40] enable compact representation of complex text contours. Although regression-based approaches are computationally efficient and integrate well with deep learning models, they can struggle with blurred edges and cluttered backgrounds.

4.2.2 Region Proposal-Based Two-Stage Methods

Region proposal-based methods treat text blocks as specific detection targets, processing them using two-stage object detection techniques like Fast R-CNN and Faster R-CNN. These methods aim to generate candidate boxes optimized for text and improve detection accuracy for arbitrarily oriented text. DenseBox [218], for example, introduced an end-to-end Fully Convolutional Network (FCN) framework that incorporates position and scale information through multi-task learning, enhancing

detection accuracy. Similarly, DeepText [41] introduced an Inception-RPN to generate more detailed text candidate boxes. While Faster R-CNN was primarily designed for horizontal text, several studies have enhanced its ability to detect irregular text regions [42–46]. Research by [47] extends these methods to handle text in any orientation, including curved text blocks, further increasing their robustness.

4.2.3 Segmentation-Based Methods

Text detection can also be approached as an image segmentation problem, where pixels are classified to identify text regions. This approach provides flexibility for handling various text shapes and orientations. Early methods [48] used Fully Convolutional Networks (FCNs) for detecting text lines, while later algorithms like PAN [49] improved efficiency and accuracy. CRAFT [50] represents a key advancement by employing character-level detection, eliminating the need for large receptive fields. Instance segmentation methods like [48] address challenges such as closely adjacent text blocks by treating each block as a distinct instance. Techniques like SPCNET [51] and LSAE [52] further improve this approach using pyramid attention modules and dual-branch architectures, respectively. Post-processing, such as binarization, is critical in segmentation-based methods, with Differentiable Binarization (DB) [53] improving both detection speed and accuracy by integrating binarization into the network.

4.2.4 Hybrid Methods

Hybrid methods combine the strengths of regression and segmentation approaches to capture both global and local text details, enhancing localization accuracy while reducing the need for extensive post-processing. Techniques like EAST [54] and MOST incorporate Position-Aware Non-Maximum Suppression (PA-NMS) to optimize detection across varying scales. Recent methods like Centripetal-Text [55] refine text localization using centripetal offsets. Additionally, innovations such as graph networks and transformer architectures [56, 57] further enhance detection by utilizing adaptive boundary proposals and attention mechanisms. Advances in transfer learning and multimodal integration [58, 59], particularly in transformer-based architectures, have addressed challenges in detecting small text areas and improved accuracy by integrating visual-textual representations.

In conclusion, text detection has advanced significantly, leveraging improvements in object detection, segmentation, and novel architectural innovations, making it a robust tool for various applications.

4.3 Text Recognition

Text recognition is a crucial component of Optical Character Recognition (OCR), referring to the automated process of converting images of written or printed text into machine-readable formats. The primary goal of a text recognition system is to interpret characters and words from visual data for subsequent computational tasks. Over time, various text recognition methods have emerged, primarily categorized into three groups: vision feature-based methods, connectionist temporal classification (CTC) loss-based methods, and sequence-to-sequence (seq2seq) techniques.

4.3.1 Vision Feature-Based OCR Technology

- **Image Feature-Based Methods:** Recent advancements in OCR technology leverage image processing, particularly Convolutional Neural Networks (CNNs), to capture spatial features from text images. These methods localize and recognize characters by eliminating the need for traditional feature engineering, directly deriving features from images. This simplifies model design and implementation while effectively capturing spatial structural information, making these techniques especially useful for regular or semi-structured text images. For instance, [219] proposed a model using CNNs for detecting text regions and classifying characters, effectively managing character spatial arrangements. Similarly, [60] introduced a synthetic data generator combined with a deep CNN architecture to improve adaptability across various text types. The CA-FAN model [61] enhances character recognition accuracy by employing a character attention mechanism. Additionally, TextScanner [62] combines CNNs with Recurrent Neural Networks (RNNs) to improve character segmentation and positioning accuracy. Despite their effectiveness, these methods face challenges with complex or irregular text, particularly in cases involving significant background noise or

intricate text structures, often requiring additional post-processing for enhanced recognition accuracy.

- **CTC Loss-Based Methods:** The Connectionist Temporal Classification (CTC) loss function addresses sequence alignment issues by enabling models to optimize without explicit alignment between input and output sequences during training. CTC computes probabilities over all possible alignment paths, making it particularly suited for processing texts of variable lengths.

A notable application of CTC is the CRNN model by [63], which integrates CNN and RNN architectures with CTC loss for sequence generation. Deep TextSpotter [64] combines CNN feature extraction with CTC to improve text detection and recognition accuracy. ADOCRNet [65] further applies CTC with CNN and Bidirectional Long Short-Term Memory (BLSTM) networks for Arabic document recognition. However, CTC struggles with extended text and contextual nuances, which can increase computational complexity and affect model training efficiency and real-time performance.

- **Sequence-to-Sequence Methods:** Sequence-to-sequence (seq2seq) techniques use an encoder-decoder architecture to encode input sequences and generate corresponding outputs. These methods manage long-distance dependencies between input and output sequences through attention mechanisms, facilitating end-to-end training. Traditional approaches often employ RNNs and CNNs to convert image features into one-dimensional sequences, which are then processed by attention-based decoders. Despite their effectiveness, converting images into one-dimensional sequences for Transformer-based architectures presents challenges, particularly with arbitrarily oriented and irregular texts.

To address these issues, models use strategies like input correction and two-dimensional feature maps. Spatial Transformer Networks (STNs), for instance, rectify text images into rectangular, horizontally aligned characters, as seen in ASTER [66], ESIR [68], and MORAN [66]. Other models avoid input modification by learning 2D feature maps to directly extract characters from 2D space, accommodating irregular and multi-directional text, as demonstrated by SAR [70], AON [67], and SATRN [72]. The rising use of Transformer architectures represents a shift from traditional CNN and RNN models toward attention-based encoder-decoder systems. NRTR [69], for example, employs a fully self-attention architecture, using convolutional layers to convert 2D input images into 1D sequences for the encoder-decoder framework. Vision Transformer models like ViTSTR [71] dispense with traditional backbone networks, using the Vision Transformer (ViT) architecture exclusively for encoding, while TrOCR [73] fully relies on Transformer architectures for both image processing and text generation, avoiding CNNs entirely.

Performance improvements for irregular or elongated text sequences focus on better handling two-dimensional geometric positional information. For example, the method proposed by [220] integrates a correction module akin to traditional R-CNN approaches, along with text grouping and arrangement modules. LOCR [74] enhances OCR performance on long texts in documents by incorporating positional information of document elements alongside positional encoding from image blocks. OCR research continues to evolve, especially in using Transformer architectures to improve performance for complex image texts [221, 81].

4.3.2 Incorporation of Semantic Information

Text recognition, traditionally approached as a visual classification task, benefits significantly from the integration of semantic information and contextual understanding, especially when dealing with irregular, blurred, or occluded text. Recent research emphasizes incorporating semantic understanding into text recognition systems, broadly classified into three approaches: character-level semantic integration, enhancements through dedicated semantic modules, and training refinements to improve contextual awareness.

- **Character-Level Semantic Integration:** Enhancing OCR performance with character-level semantic information involves leveraging character-related features, such as counts and orders. The RF-L (Reciprocal Feature Learning) framework proposed by [75] highlights the benefit of using implicit labels, such as text length, for improved recognition. RF-L incorporates a counting task (CNT) to predict character frequencies, aiding the recognition task. Similarly, [76] presents a context-aware dual-parallel encoder (CDDP), using cross-attention and specialized loss functions to integrate sorting and counting modules. Despite

improving performance by integrating character information, challenges remain in capturing diverse prior knowledge from large-scale, unlabeled texts for language model pre-training.

- **Enhancements Through Semantic Modules:** While character-level semantic integration is valuable, some approaches focus on independent semantic modules to capture higher-level semantic features. These strategies align visual and semantic data via contextual relationships within specialized modules. SRN [77], for instance, introduces a Parallel Visual Attention Module (PVAM) and a Global Semantic Reasoning Module (GSRM) to align 2D visual features with characters, transforming character features into semantic embeddings for global reasoning. Similarly, SEED [78] adds a semantic module between the encoder and decoder, enhancing feature sequences through semantic transformations. ABINet [79] refines character positions through iterative feedback, using a separately trained language model for contextual refinement. These strategies align semantic and visual data efficiently, but challenges remain in fully leveraging semantic relationships.
- **Training Advancements for Contextual Awareness:** Pre-training strategies adapted from natural language processing (NLP), such as BERT, have played a pivotal role in enhancing context-awareness in OCR tasks. Methods like VisionLAN [80] use masking to improve contextual understanding, introducing a Masked Language Perception Module (MLM) and a Visual Reasoning Module (VRM) for parallel reasoning. Similarly, Text-DIAE [81] applies degradation methods like masking, blurring, and noise addition during pre-training to improve OCR capabilities. PARSeq [82] modifies Permutation Language Modeling (PLM) to enhance text recognition by reordering encoded tags for better contextual sequences. While these pre-training approaches improve semantic learning, they often increase computational complexity and resource demands.

4.4 Text Spotting

Text spotting involves the detection and transcription of textual information from images, which encompasses both text detection and recognition tasks. Traditionally, these tasks have been handled independently, with a detector identifying text regions and a recognition module subsequently transcribing the identified text. Although this approach is straightforward, separate processing of detection and recognition may impact performance, as the final result’s effectiveness largely hinges on the accuracy of the text detection model. With advancements in deep learning, recent efforts have increasingly focused on developing end-to-end models that integrate text detection and recognition, thereby enhancing efficiency and accuracy by sharing feature representations. End-to-end text spotting models based on deep learning fall into two primary categories: two-stage and single-stage methods, each offering unique advantages. While both methods have been explored, recent research primarily emphasizes one-stage approaches.

- **Two-Stage Methods:**

Two-stage methods integrate text detection and recognition architectures, enabling joint training and feature alignment. This approach permits the sharing of information between detection and recognition tasks by extracting common features, often through shared convolutional layers, and linking tasks using a Region of Interest (RoI) mechanism. In the detection phase, the model identifies potential text regions and maps them onto the shared feature map in the recognition phase for transcription.

For instance, a foundational two-stage method combined a single-scan text detector with a sequence-to-sequence recognizer using rectangular RoIs [222]. Subsequent work improved multi-directional text detection using a similar architecture [64]. However, rectangular RoIs are most suited for organized text layouts and can be compromised by background elements, prompting researchers to explore alternative RoI methods. Some methods employed object detection technologies, such as FOTS [223] with the RoIRotate mechanism, and the Mask TextSpotter series [224, 225], AE TextSpotter [226], and ABINet++ [227] using RoIAlign. Notably, Mask TextSpotter v1 was the first to fully implement end-to-end OCR, allowing feedback between detection and recognition during joint training, while Mask TextSpotter v3 [225] introduced a Segmentation Proposal Network (SPN) to provide flexible text region representations.

Other two-stage methods, such as [228], integrated attention mechanisms with text alignment layers instead of RoI. Innovations in RoI mechanisms include TextDragon’s [229] RoLSide operator, which extracts and aligns arbitrary text regions, and BezierAlign in ABCNet [230],

which adapts to text contours rather than rectangular boundaries. PAN++ [231] uses a masked region of interest attention recognition head to balance accuracy and speed, while SwinTextSpotter [232] introduced a mechanism for detection-informed recognition. In 2022, GLASS [233] proposed Rotated-RoIAlign to enhance text feature extraction from shared backbones, addressing challenges posed by varying text sizes and orientations through a global attention module.

Despite these innovations, two-stage methods possess inherent limitations. Their reliance on precise detection results increases demands on the detection modules and necessitates high-quality annotated datasets. Furthermore, the RoI operations and post-processing steps are computationally intensive, especially for arbitrary text shapes.

- **One-Stage Methods:**

One-stage methods unify text detection and recognition within a single architecture, obviating the need for distinct modules. By sharing loss functions, the two tasks can be trained and optimized conjointly, avoiding potential performance losses from module separation. The first one-stage approach, proposed by [234], introduced Convolutional Character Networks that detect characters as fundamental units and produce character boundaries and labels, eliminating the need for RoI cropping. While effective for English text, this approach is computationally demanding. CRAFTS [235] continued this character-based approach, integrating detection results into an attention-based recognizer to propagate recognition loss across the network.

Subsequent developments, such as [236], incorporated Shape Transformer Modules to optimize end-to-end detection and recognition, while MANGO [237] employed a position-aware mask attention module to apply attention weights directly to character sequences. Recent encoder-decoder models have further evolved, with PGNet [238] and PageNet [239] decoding feature maps into sequences, while the SPTS series [240, 241] and TESTR [242] adopted Transformer-based architectures. Enhanced with cross-attention mechanisms, CLIP-based models [243] improved collaboration between image and text embeddings. In [244], the application of text spotting to video text is introduced with TransDETR, a Transformer-based framework that simplifies tracking and recognition of text over time, which may benefit document text spotting tasks.

Although one-stage models have demonstrated versatility and improved accuracy, they require more complex training processes than two-stage models and may not perform equally effectively in certain specialized text-processing tasks.

5 Mathematical Expression Detection and Recognition

5.1 Introduction to Mathematical Expression Detection and Recognition

Mathematical expression detection and recognition focus on identifying and interpreting mathematical expression within documents. This process is typically categorized by the type of mathematical expression: handwritten or printed. Handwritten mathematical expressions, facilitated by advances in stylus technology, can be further divided into online (real-time) and offline recognition; this discussion centers exclusively on offline mathematical expressions.

In documents, mathematical expressions manifest as either displayed mathematical expressions, which are distinct from regular text, or in-line expressions, which are embedded within text lines. Displayed mathematical expressions are easier to identify using document layout analysis, whereas in-line mathematical expressions present challenges due to their proximity to regular text, necessitating specialized detection techniques.

Early approaches to mathematical expression detection relied on rule-based methods [245–253] or adaptations of document layout analysis [254–256]. For in-line mathematical expressions, statistical and machine learning techniques, including Support Vector Machines and Bayesian models, were commonly used for feature extraction and classification [252, 253, 257, 256, 258].

With the advent of deep learning, mathematical expression detection increasingly resembles object detection in document images, utilizing bounding boxes or instance segmentation to isolate mathematical expression regions. Continuous mathematical expression blocks can be managed through advanced segmentation techniques.

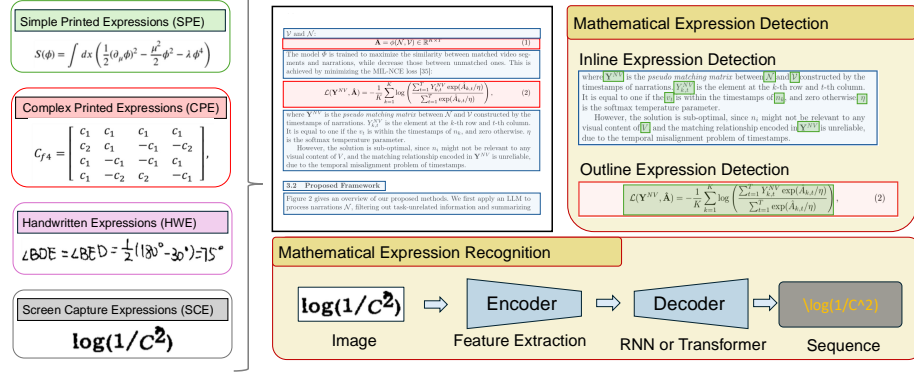


Figure 5: Overview of the Mathematical Expression Detection and Recognition

The challenge of recognizing printed mathematical expressions was first addressed in the 1960s [259], marking the initiation of efforts to convert mathematical expression images into structured code or tags. Unlike regular text, mathematical expressions pose a unique challenge due to their extensive symbol set, two-dimensional format, and context-dependent meanings. By the late 20th century, mathematical expression recognition had been segmented into symbol recognition and structure segmentation [260]. The sequential method became a pivotal approach, segmenting symbols in mathematical expressions prior to recognition via classification algorithms [261–264], complemented by grammar rules that convey the 2D structure of mathematical expressions [265–267].

An alternative perspective regards Mathematical Expression Recognition (MER) as a global optimization task, recognizing symbols and their interrelationships holistically without explicit segmentation [268–270].

Following the success of sequence-based deep learning in domains such as machine translation, the PAL model introduced the sequence-to-sequence "encoder-decoder" architecture to mathematical expression recognition, representing a significant advancement [92]. Although subsequent models have built upon this foundation, a comprehensive shortage of end-to-end solutions for mathematical expression detection and recognition remains, leading to distinct challenges that are addressed here.

5.2 Mathematical Expression Detection

5.2.1 Early Work and Convolutional Neural Networks

Initial endeavors in mathematical expression detection (MED) utilized convolutional neural networks (CNNs) for mathematical expression localization. Studies such as [271, 204, 272] harnessed CNNs alongside traditional manual feature extraction to generate bounding boxes for mathematical expression identification. Notably, [271] employed recurrent neural networks (RNNs) to extract character sequences; however, these models did not support fully end-to-end detection, limiting their generalization and performance. The Unet model, introduced for end-to-end detection in [83], concentrated on printed documents and circumvented complex segmentation tasks. While effective in detecting in-line mathematical expressions, it lacked robustness against noise.

5.2.2 Advances in Object Detection Algorithms

MED has evolved through adaptations of generic object detection algorithms into specialized forms, including both single-stage and two-stage approaches. Single-stage detectors, like DS-YOLOv5 [90], incorporated deformable convolutions and multi-scale architectures to enhance detection accuracy and speed. Similarly, the Single Shot MultiBox Detector (SSD) [85] accelerated computations using a sliding window strategy for scale-invariant detection. The 2021 ICDAR competition showcased advancements such as Generalized Focal Loss (GFL) to address class imbalance, bolstered by Feature Pyramid Networks to improve small mathematical expression detection.

Two-stage detectors, notably R-CNN variants [87], deliver high accuracy but at the expense of computational speed. Techniques such as Faster R-CNN and Mask R-CNN are applied directly, refined with region proposal networks (RPN) for enhanced performance [273, 274]. Despite the persistent challenges associated with multi-anchor configurations, anchor-free methods like FCOS and DenseBox have emerged, although they lack specific optimizations for MED [275, 276].

5.2.3 Instance Segmentation Techniques

Instance segmentation algorithms align well with MED, effectively managing non-linear and dense mathematical expression configurations through pixel-level segmentation. Mask R-CNN [277] advanced the field by incorporating pixel mask predictions within its framework, resulting in superior region recognition. PANet [278] and Hybrid Task Cascade (HTC) [279] further enhanced these approaches by improving semantic localization and integrating detection with segmentation tasks. In 2024, FormulaDet [91] innovated by framing MED as an entity and relation extraction problem, successfully utilizing contextual and layout-aware networks. This integrated approach demonstrated substantial improvements in understanding and detecting complex formula structures.

5.3 Mathematical Expression Recognition (MER)

Mathematical Expression Recognition (MER) models frequently utilize encoder-decoder architectures to convert visual representations into structured formats like LaTeX. These models predominantly rely on CNN-based encoders, with recent advancements integrating Transformer-based encoders. On the decoder side, RNN and Transformer architectures are commonly employed, with numerous enhancements improving model performance.

5.3.1 Encoder Strategies in MER

The fundamental task of MER encoders is to extract meaningful image features that encapsulate the complexity of mathematical expression. Traditional CNNs, known for their ability to capture local features, have been extensively utilized; however, they often struggle with the multi-scale and intricate nature of mathematical expression representations. Enhancements such as dense convolutional architectures and multi-directional scanning (e.g., MDLSTM) address these limitations by facilitating enriched spatial dependencies.

- **Convolutional Approaches:** A variety of convolutional architectures, including DenseNet and ResNet, have been proposed to improve feature extraction for MER [94, 95]. Subsequent developments involve augmenting CNNs with RNNs or positional encoding to better capture mathematical expression structures, thereby enhancing spatial and contextual interpretations [92, 93].
- **Transformer Encoders:** Acknowledging the limitations of CNNs in handling long-range dependencies, newer models utilize vision-based Transformers like the Swin Transformer [96], which offer superior capabilities for managing global context and complexity through self-attention mechanisms.

5.3.2 Decoder Approaches for MER

Decoding in MER entails sequential data processing akin to optical character recognition (OCR), employing architectures such as RNNs and Transformers. RNN-based decoders, enhanced with attention mechanisms, generate sequences that correspond to the inherent order of the input [92, 97]. These models excel in managing contextual dependencies, which are essential for accurately addressing nested and hierarchical expressions.

Advanced designs incorporate Gated Recurrent Units (GRUs) and attention mechanisms for resource-efficient processing, catering to the complexities of intricate mathematical expression structures. Meanwhile, tree-structured and Transformer-based decoders tackle challenges associated with vanishing gradients and computational overhead, thereby enhancing robustness in handling extensive formulaic notation [98, 99].

5.3.3 Other Improvement Strategies

In addition to improvements in encoder-decoder architectures, other strategies have emerged to enhance MER accuracy.

- **Character and Length Hints:** Incorporating character and length information aids in managing diverse handwriting styles and sequence lengths, often embedded as supplementary clues within traditional frameworks [100, 101].
- **Stroke Order Information:** Utilizing stroke sequence data is particularly beneficial for online handwritten mathematical expressions, providing deeper insights into structural semantics [102, 103].
- **Data Augmentation:** Innovative data manipulation techniques, such as pattern generation and pre-training augmentation, are crucial for enhancing dataset robustness and model performance, thereby mitigating architectural stagnation [93, 96].

6 Table Detection and Recognition

6.1 Introduction to Table Detection and Recognition

Tables serve as essential information carriers in various documents, including reports, academic papers, financial statements, and technical documentation. They present data in a structured and coherent manner, facilitating quick comprehension of relationships and hierarchies. Consequently, accurate table detection and recognition are critical for effective document analysis.

Table detection entails identifying and segmenting table areas within document images or electronic files. This process aims to locate tables while distinguishing them from other content, such as text or images. The precision of table detection directly impacts the success of subsequent structural recognition and data extraction, establishing it as a crucial initial step.

Current methods for table detection predominantly rely on Document Layout Analysis (DLA) and enhanced object detection algorithms. While DLA excels in identifying and segmenting table areas, it often encounters difficulties with tables that lack clear boundaries. Transformer-based object detection algorithms have also been adapted for table detection, refining general methodologies for improved outcomes.

As detection accuracy increases, research focus has shifted toward Table Structure Recognition, which involves analyzing the internal structure of tables post-detection. This includes tasks such as segmenting rows and columns, extracting cell content, and interpreting cell relationships into structured formats like LaTeX. Enhanced recognition capabilities facilitate automated information processing, minimizing manual intervention and enriching document analysis applications.

This section reviews target detection-based algorithms for table detection and discusses three deep learning-based table recognition methods derived from recent research.

6.2 Table Detection Based on Object Detection Algorithms

Table detection (TD) is often framed as an object detection task, where tables are treated as objects and models originally developed for natural images are applied. Despite the inherent differences between page elements and natural images, one-stage, two-stage, and transformer-based models can yield robust results with careful retraining and tuning, frequently serving as benchmarks for TD.

To adapt object detection for TD, numerous studies have sought to enhance standard methods. For instance, [104] integrates PDF features, such as character coordinates, into CNN-based models. [105] customizes Faster R-CNN for document images by modifying their representation and optimizing anchor points. [106] combines Deformable CNNs with Faster R-CNN to accommodate varying table scales, while [107] fine-tunes Faster R-CNN specifically for tables. [108] employs the YOLO series, enhancing anchor and post-processing techniques.

To address table sparsity, [109] expands SparseR-CNN with Gaussian Noise Augmented Image Size proposals and many-to-one label assignments, introducing the Information Coverage Score (ICS) to evaluate recognition accuracy.

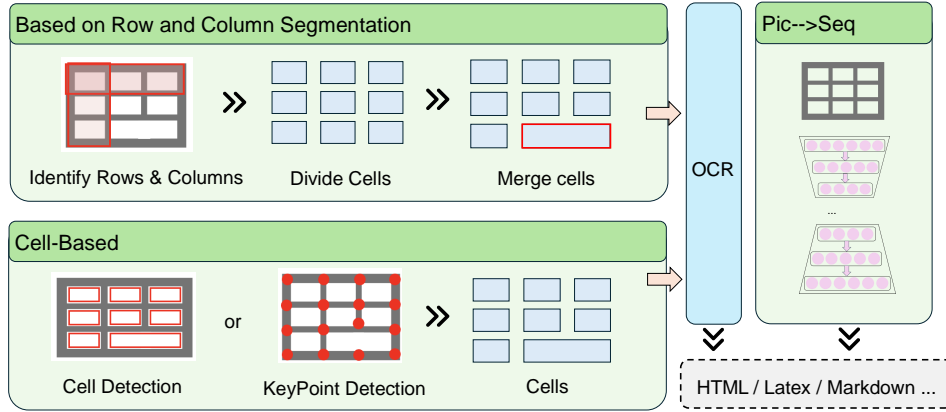


Figure 6: Overview of the Table Detection and Recognition

6.3 Table Structure Recognition

Historically, table structure recognition depended on manual rules and heuristics, such as the Hough Transform for line detection and blank space analysis for unframed tables. However, these methods struggled with complex table layouts. Recent advances have leveraged algorithms from document layout and formula detection to improve table structure recognition, categorizing approaches into row and column segmentation, cell detection, and sequence generation methods. TabNet [280] is an innovative deep learning model designed for table feature extraction, processing both numerical and categorical features of tabular data in an end-to-end manner. It introduces a highly efficient and interpretable learning architecture, optimized for diverse downstream tasks. TabNet’s sequential attention mechanism allows the model to selectively focus on relevant features progressively, using instance-level sparse feature selection and a multi-step decision architecture. This approach enhances TabNet’s ability to explain feature importance at both local and global levels. Building on this foundation, models like TabTransformer [281] have further advanced table feature extraction, contributing valuable insights for the development of robust table recognition models.

6.3.1 Methods Based on Row and Column Segmentation

The main challenge in table structure recognition often lies in detecting individual cells, particularly due to large blank spaces. Early deep learning approaches tackled this by segmenting tables into rows and columns. These algorithms generally follow a top-down strategy, first identifying the overall table region and subsequently segmenting it into rows and columns. This relatively straightforward method proves effective for tables with clear boundaries and simple layouts.

- **Row and Column Detection:** Initially, table structure recognition was perceived as an extension of table detection, primarily utilizing object detection algorithms to identify table bounding boxes. Segmentation algorithms then established relationships between rows and columns. Convolutional Neural Networks (CNNs) and Transformer architectures played pivotal roles in this context. For instance, Faster R-CNN was adapted for table detection, followed by a Fully Convolutional Network (FCN) for semantic segmentation, effectively capturing the table’s structure [110]. Similarly, models such as Unet and DeepLab facilitated end-to-end semantic segmentation for pixel-level recognition [111]. CNN-based techniques, like multi-task learning with VGG-19, focused on both table area and row/column segmentation [112]. In contrast, Transformers, such as DETR, excelled at recognizing global relationships within an image, thereby enhancing generalization. Innovations included row and column segmentation through transformer queries [282] and a dynamic query enhancement model, DQ-DETR [113]. Additionally, Bi-directional Gated Recurrent Units (Bi-GRUs) effectively captured row and column separators by scanning images bidirectionally [114].

- **Fusion Module:** While earlier methods emphasized detecting table lines, they often overlooked complex inter-cell relationships. Advanced algorithms have constructed models to estimate merging probabilities between cells, thereby improving recognition accuracy in tables that lack explicit row and column lines. For example, embedding modules were employed to integrate plain text within grid contexts, guiding merge predictions via GRU decoders [115]. Other techniques have utilized adjacency criteria and spatial compatibility to predict cell mergers [116]. The integration of global computational models, such as Transformers, further enhances the analysis of complex tables [117].

CNNs remain foundational for feature extraction in table images, although recent efforts aim to optimize architectures for table-specific characteristics. For instance, replacing ResNet18 with ShuffleNetv2 significantly reduced model parameters [283]. Despite progress, challenges persist in tables that lack explicit lines, such as those with sparse content or irregular arrangements.

6.3.2 Cell-based Methods

Cell-based methods, characterized as bottom-up approaches, construct tables by detecting individual cells and merging them based on visual or textual relationships. These methods typically involve two stages: detecting cell boundaries and subsequently associating cells to form the overall table structure, thereby offering advantages in handling complex tables and minimizing error propagation.

Early enhancements concentrated on improving cell keypoint detection and segmentation accuracy. For example, HRNet served as a backbone for high-resolution feature representation in tasks such as multi-stage instance segmentation [118]. Some approaches introduced new loss terms to enhance detection, including continuity and overlap loss [119]. Others developed dual-path models to learn local features and optimize table segmentation [120].

Vertex prediction, which focuses on the corners of cells, proved beneficial for addressing deformed cells resulting from angles or perspectives. Techniques like the Cycle-Pairing Module simultaneously predicted centers and vertices of cells [121]. Representing tables as graph structures enabled a more nuanced understanding, employing Graph Neural Networks (GNNs) to model complex relationships [122]. These methods effectively improved upon the limitations of traditional grid-based approaches in capturing intricate cell relationships.

Graph-based methods effectively leverage cell characteristics by treating tables as graphs, where cells represent vertices and relationships signify edges. This approach allows for comprehensive modeling of adjacency relationships, positioning GNNs as powerful tools for managing complex tables [123].

While effective, cell-based methods can be computationally demanding, as they involve independent detection and classification for each cell. Errors occurring at this stage can significantly affect the final table structure.

6.3.3 Image to Sequence

Building on advancements in OCR and formula recognition, image-to-sequence methods convert table images into structured formats, such as LaTeX, HTML, or Markdown. Encoder-decoder frameworks utilize attention mechanisms to encode table images into feature vectors, which decoders subsequently transform into descriptive text sequences.

Early efforts by [124] implemented encoder-decoder architectures to translate images from scientific papers into LaTeX code. Subsequent models refined these techniques with dual-decoder architectures, enabling concurrent handling of structural and textual information [125]. The MASTER architecture, adapted for scene text recognition, effectively distinguished between structural elements and positional information [126].

Recent advancements propose designing Transformer architectures specifically for scientific tables, enhancing robustness against the complex features found in particular contexts, such as medical reports [127]. Solutions like the VAST framework have demonstrated improved accuracy by employing dual-decoders for managing both HTML and coordinate sequences [128].

While these methods present considerable advantages in processing complex tables, the inherent challenge lies in training models to adequately capture diverse table structures without succumbing to error propagation.

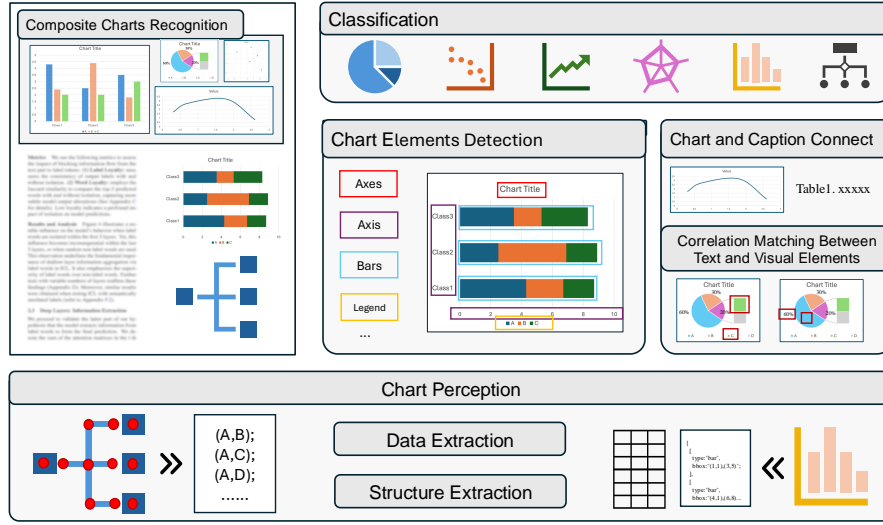


Figure 7: Overview of the Chart-related Tasks in Document

6.4 Chart Perception

6.4.1 Introduction to Tasks Related to Charts in Documents

Charts in documents are graphical representations that present data in a concise and intuitive manner, making it easier to visualize patterns, trends, and relationships. Common chart types include line graphs, bar charts, area charts, pie charts, and scatter plots, all of which play a critical role in conveying key insights.

Tasks related to processing charts in documents typically involve several subtasks, such as chart categorization, segmentation of composite charts, title matching, chart element identification, and data and structure extraction. The first step, chart categorization, is essential for distinguishing between different chart types, as each type requires a unique extraction method. For example, identifying key points such as turning points and endpoints is crucial in line charts, while in bar charts, both the data and the associated text labels are key components. Determining whether the chart is a line chart, bar chart, or another type is crucial for successful extraction.

Once categorized, titles are matched with charts, and composite charts are segmented. Elements such as axes, labels, scales, data points, and legends are detected and recognized, often using bounding box detection, to facilitate further extraction.

Chart perception, also known as chart information extraction, involves retrieving numerical and textual data from visual representations. The ultimate goal is to convert visual data into structured formats like tables or JSON, enabling more effective processing by models. For instance, scatter plot points, line chart inflection points, pie chart proportions, and bar chart elements are extracted and matched with corresponding labels to produce structured data. Additionally, for diagrams such as flowcharts, structural diagrams, and mind maps, extracting structural information aids downstream tasks such as reasoning.

The use of visual-language models in chart comprehension and reasoning tasks, including summarization and question-answering, offers great potential for interpreting and representing chart content in natural language [284].

6.5 Chart Classification

Chart classification categorizes various chart types by focusing on their visual characteristics and representational forms. This process aims to accurately identify charts—such as bar charts, pie charts,

line charts, scatter plots, and heat maps—either manually or through automation. A key challenge lies in the diversity of chart types and their often-subtle visual distinctions, which complicates automatic differentiation. Accurate classification is essential for subsequent tasks, such as data mining and chart analysis.

The success of AlexNet in the 2015 ImageNet competition sparked the dominance of deep learning models in image classification, including chart classification. Convolutional Neural Networks (CNNs) like VGG, ResNet, Inception, and EfficientNet have been utilized to extract high-level features from chart images. Transfer learning has further improved classification accuracy by leveraging features learned from natural images for chart tasks [285]. Networks such as ResNet and VGG perform particularly well on large datasets and complex chart types [129–132]. As datasets grew and chart complexity increased, models like DenseNet and EfficientNet demonstrated superior performance [133–135].

Despite these advances, CNN-based models often struggle with noisy or visually similar charts. To address these challenges, Vision Transformers have emerged as a promising development. In the 2022 chart classification competition, a pre-trained Swin Transformer outperformed other models [136]. The Swin Transformer, with its hierarchical structure and local window attention mechanism, effectively manages both global and local image features, surpassing competitors in handling complex charts [134]. The Swin-Chart model [137], which integrates a fine-tuned Swin Transformer, further enhanced performance through a weight-averaging strategy. Additionally, [138] proposed a coarse-to-fine curriculum learning strategy, significantly improving the classification of visually similar charts.

6.6 Chart Detection and Element Recognition

Detecting and segmenting charts in documents requires layout detection algorithms to accurately identify chart regions. Before extracting data, tasks like associating charts with titles, segmenting multi-panel charts, and recognizing chart elements are necessary.

6.6.1 Chart and Caption Association

Linking charts to their titles connects charts with their captions in documents, essential for understanding chart data and enabling efficient retrieval. Rule-based algorithms typically determine this relationship by analyzing the spatial layout of text around the chart [139, 140], employing methods like the Hungarian algorithm and greedy matching [286, 287]. Recent approaches use CNN-based classifiers and OCR to link titles with charts through large datasets [288, 289], and multimodal Transformers combine image and text features for improved title association [290, 291].

6.6.2 Recognition of Composite Charts

Composite charts compile multiple sub-charts within a single frame, each with distinct data. Segmenting these charts is critical for accurate data extraction. Early methods employed geometric features and pixel-contour-based segmentation [141, 292]. Viewing segmentation as an object detection task, approaches like YOLO and Faster R-CNN enable simultaneous detection of sub-charts and their elements [293, 140].

6.6.3 Detection of Chart Elements

Charts contain both text and visual elements critical for data extraction. Key tasks include detecting text and classifying it into categories like titles and labels. Recent methods for text detection in charts often use semi-automatic systems, with input from users to identify important elements like axis labels [288, 142–144]. Traditional systems like Microsoft OCR and Tesseract OCR, while limited in precision, are still widely used [142, 150]. Visual elements are detected similarly to text, with deep learning models increasingly replacing rule-based methods. The 2023 Context-Aware system uses Faster R-CNN to detect elements like legends and data points, relying on a Region Proposal Network [294].

6.6.4 Correlation Matching Between Text and Visual Elements

Linking text to corresponding visual elements is critical for interpreting chart data. Early methods were rule-based, focusing on positional relationships [295–297, 131]. Recent advancements, such as the Swin Transformer-based method introduced in 2022, have refined these techniques, providing better correlation matching through transformer architectures [136, 145].

6.7 Chart Perception

Chart perception retrieves data from charts like bar charts, line charts, pie charts, and scatter plots. This process extracts both data structure and text information. With the development of visual language models, it is also possible to implement end-to-end diagram-to-text conversion.

6.7.1 Chart Data Extraction

Chart data extraction focuses on basic charts and has evolved from manual methods to deep learning techniques. Semi-automated systems, such as those that allow user input for accuracy, remain common. For example, [146] allows users to specify label and axis locations, while [144] enables result correction. Modern methods are either multi-stage or end-to-end. Two-stage models like FigureSeer [142] classify charts, followed by object detection. End-to-end models, such as ChartDETR [147], integrate CNNs and Transformers for element detection, producing structured data. There are also research that utilize large amounts of image-table data to perform image-to-text conversion through self-supervised training [298, 299]. Specialized methods have been developed for specific chart types, like scatter plots, bar charts and line chart [148–150, 300].

6.7.2 Chart Structure Extraction

Extracting structural information from charts such as flowcharts and tree diagrams requires detecting components like cell boxes and connecting lines. Research on flowchart structure extraction has focused on both hand-drawn and machine-generated charts [301, 302]. Recent models, such as FR-DETR [151], combine DETR and LETR to simultaneously detect symbols and edges, improving accuracy. However, challenges remain, especially with complex connecting lines, as highlighted by [152], which focuses on organizational charts using a two-stage method for line detection.

7 Large Models for Document Parsing: Overview and Recent Advancements

Document Extraction Large Models (DELMs) employ Transformer-based architectures to extract multimodal information from documents (e.g., text, tables, and images) into structured data. Unlike traditional rule-based systems, DELMs integrate visual, linguistic, and structural information, enabling more effective document structure analysis, table extraction, and cross-modal associations. These capabilities make DELMs well-suited for end-to-end document parsing, facilitating deeper understanding for downstream tasks.

As Multimodal Large Language Models (MLLMs), particularly Visual-Language Models (LVLMs), have advanced, they have become increasingly adept at processing complex multimodal inputs such as documents and web pages, which combine text, images, and structured information. Despite these advancements, challenges remain in processing academic and professional documents efficiently, especially in OCR and detailed document structure extraction. The following subsections chronicle the evolution of DELMs, highlighting their solutions to these challenges and illustrating how each model builds upon previous efforts.

7.1 Initial Developments in Document Multimodal Processing

Early document extraction models, such as LLaVA-Next [156], Qwen-VL [157], and InternVL [160], aimed to understand multimodal content (i.e., images and text) in documents. These models laid the foundation for large-scale document analysis by training on broad datasets containing images and text. However, their general-purpose image understanding was insufficient for more complex academic and professional documents, where domain-specific tasks like OCR and detailed document structure analysis were essential. These models were effective at understanding visual content but lacked the granularity needed for text-heavy documents, such as technical reports or academic papers.

To address this gap, models such as DocOwl1.5 [165] and Qwen-VL were fine-tuned on document-specific datasets. Enhancements to the CLIP-ViT architecture improved their performance in document-related tasks. Additionally, techniques such as sliding windows, used by models like TextMonkey[161] and Ureader[163], helped break large, high-resolution documents into smaller segments, improving OCR accuracy. Nevertheless, these early models still struggled with aligning extensive textual and visual information, as evidenced by the GOT model [6], where the focus on visual reasoning conflicted with fine-grained text extraction.

7.2 Advances in OCR and End-to-End Document Parsing

In 2023, Nougat[170] marked a significant leap in document extraction by being the first end-to-end Transformer model designed for academic document processing. Built on Donut and using a Swin Transformer encoder with an mBART[303] decoder, Nougat allowed direct conversion of academic documents into Markdown format. This innovation integrated mathematical expression recognition and page relationship organization, making it particularly suitable for scientific documents. Nougat represented a shift from modular OCR systems that handled text extraction, formula recognition, and page formatting separately. However, it faced limitations in processing non-Latin scripts and suffered from slower conversion speeds due to high computational demands.

While Nougat addressed many previous models' shortcomings, its focus on academic documents left room for improvement in other areas, such as fine-grained OCR tasks and chart interpretation. Vary[304] emerged to tackle these challenges by improving chart and document OCR. Vary expanded the visual vocabulary library by integrating a SAM-style visual vocabulary into its framework, allowing better OCR and chart understanding without fragmenting document pages. However, Vary still struggled with handling language diversity and multi-page documents, demonstrating the ongoing need for more specialized models.

7.3 Handling Multi-Page Documents and Fine-Grained Tasks

In 2024, Fox[168] introduced a novel approach for multi-page document understanding and fine-grained focus tasks. By leveraging multiple pre-trained visual vocabularies, such as CLIP-ViT and SAM-style ViT, Fox enabled simultaneous processing of both natural images and document data without modifying the pretrained weights. Additionally, Fox employed hybrid data generation strategies that synthesized datasets combining textual and visual elements, improving performance in tasks such as cross-page translation and summary generation. This model addressed the limitations of earlier DELMs that struggled with complex, multi-page document structures.

Although Fox excelled in multi-page document processing, its approach to hierarchical document structures was further refined by models like Detect-Order-Construct[305]. Detect-Order-Construct introduced a tree-construction-based method for hierarchical document analysis, dividing the process into three stages: detection, ordering, and construction. By detecting page objects, assigning logical roles, and establishing reading order, the model reconstructed hierarchical structures for entire documents. This unified relation prediction approach outperformed traditional rule-based methods in understanding and reconstructing complex document structures.

7.4 Unified Frameworks for Document Parsing and Structured Data Extraction

The introduction of models like OmniParser [306] marked a shift toward unified frameworks combining multiple document processing tasks, such as text parsing, key information extraction, and table recognition. OmniParser's two-stage decoder architecture enhanced the extraction of structural information, offering a more interpretable and efficient method for managing complex relationships within documents. By decoupling OCR from structural sequence processing, OmniParser outperformed earlier task-specific models like TESTER and SwinTextSpotter in both text detection and table recognition, while also reducing inference time.

In parallel, GOT [6], released in 2024, introduced a universal OCR paradigm by treating all characters (text, formulas, tables, musical scores) as objects. This approach enabled the model to handle a wide range of document types, from scene text OCR to fine-grained document OCR. GOT's use of 5 million text-image pair dataset and its three-stage training strategy—pre-training, joint training, and fine-tuning—allowed it to surpass previous document-specific models in handling complex charts,

non-traditional content such as musical scores, and geometric shapes. GOT represents a step toward a general OCR system capable of addressing the diverse content found in modern documents.

In conclusion, the evolution of DELMs has been marked by progressive advancements addressing specific limitations in earlier models. Initial developments improved multimodal document processing, while later models like Nougat and Vary advanced OCR capabilities and fine-grained extraction tasks. Models like Fox and Detect-Order-Construct further refined multi-page and hierarchical document understanding. Finally, unified frameworks like OmniParser and universal OCR models like GOT are paving the way for more comprehensive, efficient, and general-purpose document extraction solutions. These advancements represent significant strides in how complex documents are analyzed and processed, benefiting both academic and professional fields.

8 Datasets

8.1 Datasets for Single Tasks

8.1.1 Datasets for DLA

Datasets for Document Layout Analysis (DLA) are primarily classified into synthetic, real-world (Documents and scanned images), and hybrid datasets. Early efforts focused on historical documents, such as IMPACT [307], Saint Gall [14], and GW20 [308]. More comprehensive datasets, like IIT-CDIP [309], containing 7 million documents with intricate layouts, have also emerged. Post-2010, research interest has transitioned towards complex printed layouts alongside the continued examination of handwritten historical texts. Table 1 lists key datasets used in DLA research over the last ten years. Furthermore, major conferences such as the International Conference on Document Analysis and Recognition (ICDAR) host competitions that introduce datasets with high-quality, standardized annotations. These are essential for model evaluation and benchmarking. The ICDAR 2013 Page Segmentation Competition, for instance, focused on document layout analysis using newspapers, journals, and magazines with multiple annotation types. The ICDAR 2021 Competitions emphasized historical documents, addressing layout challenges due to aging, and scientific literature parsing for extracting structured information.

Table 1: Summary of Common Used Datasets for DLA

Dataset	Class	Instance	Document Type	Language
PRImA [310]	10	305	Multiple Types	English
BCE-Atabic-v1 [311]	3	1833	Arabic books	Arabic
Diva-hisdb [312]	Text Block	150	Handwritten Historical Document	Multiple Languages
DSSE200 [313]	6	200	Magazines, Academic papers	English
OHG [314]	6	596	Handwritten Historical Document	English
CORD [315]	5	1000	Receipts	Indonesian
FUNSD [316]	4	199	Form document	English
PubLayNet [317]	5	360000	Academic papers	English
Chn [318]	5	8005	Chinese Wikipedia pages	Chinese
DocBank [319]	13	500000	Academic papers	English, Chinese
BCE-Atabic-v1 [320]	21	9000	Arabic books	Arabic
DAD [321]	5	5980	Articles	English
DocLayNet [322]	11	80863	Multiple Types	Primarily English
D4LA [27]	27	11092	Multiple Types	English
M6Doc [323]	74	9080	Multiple Types	English, Chinese

8.1.2 Dataset for OCR

This section provides an overview of common OCR datasets for printed text. Among the most notable and widely used are those introduced in various ICDAR competitions, such as ICDAR2013 and ICDAR2015, which include real-world scene and document images and are frequently used to evaluate scene text detection algorithms. Additionally, datasets like Street View Text Perspective and MSRA-TD500 focus on detecting irregular text in challenging contexts. Synthetic datasets, such as SynthText and SynthAdd, are artificially generated and contain large volumes of data, making them popular for text detection and recognition tasks. Moreover, datasets like ICDAR2015 and ICDAR2019 offer both regional annotations and textual content, supporting end-to-end OCR tasks. A summary of commonly used OCR datasets is provided in Table 2.

Table 2: Summary of Common Used Datasets for OCR

Dataset	Instance	Task	Feature	Language
IIIT5K [324]	5000	TR	Real-world scene text	English
Street View Text [325]	647	TD	Street View	English
Street View Text Perspective [326]	645	TD	Street View with perspective distortion	English
ICDAR 2003 [327]	507	TD & TR	Real-world short scene text	English
ICDAR 2013 [328]	462	TD & TR	Real-world short scene text	English
MSRA-TD500 [329]	500	TD	Rotated text	English, Chinese
CUTE80 [330]	13000	TD & TR	Curved text	English
COCO-Text [331]	63,686	TD & TR	Real-world short scene text	English
ICDAR 2015 [332]	1500	TD & TR & TS	Incidental Scene Text	English
SCUT-CTW1500 [333]	1500	TD	Curved text	English, Chinese
Total-Text [334]	1555	TD & TR	Multi-oriented scene text	English, Chinese
SynthText [335]	800,000	TD & TR	Synthetic images	English
SynthAdd [336]	1,200,000	TD & TR	Synthetic images	English
Occlusion Scene Text [80]	4832	TD	Occlusion text	English
WordArt [337]	6316	TR	Artistic text	English
ICDAR2019-ReCTS [338]	25,000	TD & TR & TS	TD & TR & Document Structure Analysis	Chinese

TD: Text Detection; TR: Text Recognition; TS: Text Spotting.

8.1.3 Datasets for MED and MER

In document analysis, mathematical expression detection and recognition are crucial research areas. With specialized datasets, researchers now achieve improved recognition of diverse mathematical mathematical expressions. Table 3 lists common benchmark datasets for mathematical expression detection and recognition, covering both printed and handwritten mathematical expressions across various document formats like images and Documents. These datasets support tasks such as mathematical expression detection, extraction, localization, and mathematical expression recognition. Significant datasets include UW-III, InftyCDB-1, and Marmot, often used for assessing printed mathematical expression detection, addressing both inline and standalone mathematical expressions. The ICDAR series has advanced the field through competitions with datasets like ICDAR-2017 POD and ICDAR-2021 IBEM, presenting extensive and complex scenarios. These resources improve recognition model robustness and emphasize the challenges of detecting mathematical expressions in intricate document structures. Additionally, datasets like FormulaNet and ArxivFormula emphasize large-scale detection, particularly mathematical expression extraction from images. Despite advancements, the availability of datasets for mathematical expression detection and recognition remains limited, necessitating improvement in multi-format support and robustness.

Table 3: Summary of Common Used Datasets for MED and MER

Dataset	Image	Instance	Type	Task
UW-III [339]	100	/	Inline and displayed Formula	MED
InfyCDB-1 [340]	467	21000	Inline and displayed Formula	MED
Marmo [341]t	594	9500	Inline and displayed Formula	MED
ICDAR-2017 POD [342]	3900	5400	Only displayed Formula	MED
TFD-ICDAR 2019 [343]	851	38000	Inline and displayed Formula	MED
ICDAR-2021 IBEM [344]	8900	166000	Inline and displayed Formula	MED
FormulaNet [345]	46,672	1000,00	Inline and displayed Formula	MED
ArxivFormula [91]	700000	813.3	Inline and displayed Formula	MED
Pix2tex [346]		189117	Printed	MER
CROHME [347]		12178	Handwritten	MER
HME100K [348]		99109	Handwritten	MER
UniMERNet [96]		1,061,791	Printed and Handwritten	MER

Table 4: Summary of Common Used Datasets for TD and TSR

Dataset	Instance	Type	Language	Task	Feature
ICDAR2013 [349]	150	Government Documents	English	TD & TSR	Covers complex structures and cross-page tables
ICDAR2017 POD [342]	1548	Scientific papers	English	TD	Includes shape and formula detection
ICDAR2019 [350]	2439	Multiple Types	English	TD & TSR	Includes historical and modern tables
TABLE2LATEX-450K [124]	140000	Scientific papers	English	TSR	
RVL-CDIP (subset) [351]	518	Receipts	English	TD	Derived from RVL-CDIP
IIIT-AR-13K [352]	17,000 (not only tables)	Annual Reports	Multi-languages	TD	Does not only contain tables
CamCap [353]	85	Table images	English	TD & TSR	Used for evaluating table detection in camera-captured images
UNLV Table [354]	2889	Journals, Newspapers, Business Letters	English	TD	
UW-3 Table [355]	1,600 (around 120 tables)	Books, Magazines	English	TD	Manually labeled bounding boxes
Marmot [356]	2000	Conference Papers	English and Chinese	TD	Includes diversified table types; still expanding
TableBank [357]	417234	Multiple Types	English	TD & TSR	Automatically created by weakly supervised methods
DeepFigures [287]	5,500,000 (tables and figures)	Scientific papers	English	TD	Supports figure extraction
PubTabNet [125]	568000	Scientific papers	English	TSR	Structure and content recognition of tables
PubTables-1M [358]	1000000	Scientific papers	English	TSR [122]	Evaluates the oversegmentation issue
SciTSR [359]	15000	Scientific papers	English	TSR	
FinTable [359]	112887	Scientific and Financial Tables	English	TD & TSR	Automatic Annotation methods
SynthTabNet [360]	600000	Multiple Types	English	TD & TSR	Synthetic tables
Wired Table in the Wild [121]	14582 (pages)	Photos, Files, and Web Pages	English	TSR	Deformed and occluded images
WikiTableSet [361]	50000000	Wikipedia	English, Japanese, French	TSR	
STDW [362]	7000	Multiple Types	English	TD	
TableGraph-350K [363]	358,767	Academic Table	English	TSR	including TableGraph-24K
TabRecSet [364]	38100	Multiple Types	English and Chinese	TSR	
DECO [365]	1165	Multiple Types	English	TD	Enron document electronic table files
iFLYTAB [366]	17291	Multiple Types	Chinese and English	TD & TSR	Online and offline tables from various scenarios
FinTab [367]	1,600	Financial Table	Chinese	TSR	

TD: Table Detection; TSR: Table Structure Recognition

8.1.4 Dataset for TD and TSR

Given the wide range and complex structure of tabular data, numerous representative datasets have emerged for table-related tasks. Foundational tabular datasets with broad applicability are primarily sourced from official ICDAR competitions, such as ICDAR2013 and ICDAR2017, which offer diverse sources and an appropriate level of complexity. To enhance the variety of tables in datasets, researchers introduced TableBank, which contains high-quality annotated tables from various fields, including scientific literature and business documents. This dataset increases table diversity, providing broader application scenarios and more realistic data for table detection and recognition tasks. Additionally, datasets like TabStructDB enrich existing datasets by providing more detailed structured information, such as internal cell representations and table structural details, facilitating more accurate structural analysis. Some datasets specifically address irregular table samples. For instance, the Marmot dataset focuses on the detection of both wired and wireless tables, while CamCap collects irregular tables photographed on curved surfaces, and Wired Table in the Wild (WTW) includes tables with common real-world challenges, such as occlusion and blurring. These datasets enhance the robustness of table recognition systems in complex environments. Certain datasets are tailored to specific table-related tasks. For example, FinTabNet focuses on the detection and recognition of financial tables, and SciTSR specializes in recognizing table structures in academic articles. Such datasets offer targeted support for specialized table analysis tasks and advance the progress of segmented research areas. Moreover, datasets like WikiTableSet and Marmot cover tables in multiple languages, including Chinese, helping to address the lack of linguistic diversity and enabling cross-language table detection and structural analysis. Although existing table datasets provide rich data sources and support diverse scenarios for tasks such as table detection and structural recognition, there is still room for improvement in terms of scene diversity, task specificity, and language coverage.

8.1.5 Dataset for Chart-related Task

Charts in documents involve several key tasks, including chart classification, data extraction, structure extraction, and chart interpretation. Various datasets exist to support these tasks, and those related to chart classification and information extraction are listed in the Table 5. The chart classification domain is relatively mature, with numerous widely used authoritative datasets. For instance, DeepChart (2015) comprises 5,000 charts across five categories for classification tasks. VIEW (2012) includes 300 charts in three categories, focusing on improving the accessibility of chart images. ReVision (2011) contains 2,601 instances across 10 chart categories, enabling automatic classification, analysis, and redesign of charts. These datasets provide strong support for advancing chart classification research. In contrast, the field of chart data and structure extraction often relies on custom-built datasets, such as UB-PMC (2019-2022) and Synth(2020). The UB-PMC dataset collects real charts from scientific publications, with subsets published across different years. It covers 4 to 15 chart categories, with instance counts ranging from 2,000 to 30,000. Synth 2020 consists of 9,600 synthetic charts generated using the Matplotlib library. While these datasets are valuable for data extraction tasks, publicly available data sources remain limited. In the realm of chart understanding and reasoning, recent advancements in large visual-language models have led to the creation of high-quality datasets. For example, LineEX430k (2023) focuses on line chart data extraction and contains 430,000 line chart instances. OneChart (2023) is a large-scale dataset with 10 million charts, supporting complex tasks like chart information extraction, question answering, and reasoning. These datasets significantly advance research in chart comprehension and reasoning.

8.2 Other Datasets for Document-related Tasks

In addition to specific task-oriented datasets, there are others supporting multiple document-related tasks. A notable example is the FUNSD dataset, widely used in form processing and document OCR. Although small, with only 199 images, it provides valuable resources for parsing form and document structures via annotated text blocks and relationships. This dataset is ideal for early-stage training and testing in structured document understanding.

In contrast, the SROIE dataset [380], featuring 1,000 receipt images, focuses on extracting key information such as company names, dates, addresses, and total costs, making it particularly suitable for automating document information extraction in finance and retail industries.

Table 5: Summary of Common Used Datasets for Chart-related Tasks

Dataset	Year	Instance	Class	Task	Feature
DeepChart [368]	2015	5000	5	Chart Classification	-
VIEW [369]	2012	300	3	Chart Classification	-
ReVision [288]	2011	2601	10	Chart Classification	Based on ChartSense dataset
CHART 2019 [370]	2019	4242	multi-class	Chart Classification	Real charts from scientific publications
- PMC					
CHART 2019 - Synthetic [371]	2019	202,550	multi-class	Chart Classification	Synthetic charts
DocFigure [370]	2019	33000	28	Chart Classification	Includes various figure images
UB-PMC 2019 [370]	2019	4242	7	Chart Classification	Competition dataset
UB-PMC 2020 [372]	2020	2123	4	Chart Data Extraction	Real charts from PubMedCentral
UM-PMC 2021 [373]	2021	22924	15	Chart Classification	Competition dataset
UB-PMC 2022 [136]	2022	33186	15	Chart Classification	Competition dataset
Synth 2020 [373]	2020	9600	4	Chart Data Extraction	Synthetic charts
LINEEX430k [300]	2023	430,000	Line charts	Chart Data Extraction	Focused on line charts
ICPR 2022 [136]	2022	26,596	15	Chart Classification	Charts with embedded text
ExcelChart400K [374]	2021	400,000	Pie and bar charts	Chart Data Extraction	Extracted from Excel charts with JSON annotations
CHARTER [375]	2021	32334	4	Chart Data Extraction	Sourced from document pages, web pages, PubMed, FigureQA, etc.
StructChart dataset [376]	2023	16466	Organization and structure charts	Chart Structure Extraction	-
OneChart [377]	2023	10000000	5	Chart Information Extraction, QA, and Inference	Synthesized using Matplotlib
Chart-to-Text [378]	2023	8305	6	Chart Information Extraction	Contains chart samples and corresponding data
ChartLlama [379]	2023	1500	10	7 comprehensive chart tasks including chart information extraction	GPT-4 generates charts and instruction data
ChartX [299]	2024	48000	18	7 comprehensive chart tasks including chart information extraction	Automatically generated by GPT-4 and manually checked

The LOCR dataset [74], dedicated to academic document tasks, is a large-scale resource derived from arXiv academic papers. Covering 125,738 pages and containing over 77 million text-location pairs, it offers annotated bounding boxes for complex typographical elements, making it suitable for academic OCR tasks that necessitate recognizing intricate layouts.

Another valuable resource, the DocGenome dataset [5], is an open-source large-scale benchmark comprising 500,000 scientific documents from arXiv, spanning 153 disciplines and 13 document components (e.g., charts, mathematical expressions, tables). Created using the DocParser annotation tool, it supports multimodal tasks like document classification, layout detection, and visual positioning, as well as document component conversion to LaTeX. This dataset is designed to evaluate and train large multimodal models for document understanding tasks.

OCRBench [381] serves as a comprehensive evaluation platform, integrating 29 datasets that cover various OCR-related tasks such as text recognition, visual question answering, and handwritten mathematical expression recognition. It highlights the complexity of OCR tasks and the potential of multimodal models for cross-task performance.

In specialized domains, the CHEMU [382] and ChEMBL25 [383] datasets focus on recognizing molecular mathematical expressions and chemical structures in chemical literature, thus expanding OCR applications to scientific symbol extraction and analysis. MUSCIMA++ [384] and DeepScores [385] target music score OCR by annotating handwritten music scores and symbols, thereby advancing music symbol recognition. These datasets illustrate the potential and challenges of OCR in highly technical fields.

Recent developments in datasets for large document models have opened new avenues for document parsing and large-scale model training. For instance, Nougat utilizes datasets from arXiv, PubMed Central (PMC), and the Industrial Document Library (IDL), constructed by pairing Document pages with source code, particularly for preserving semantic information in mathematical expressions and tables.

The Vary dataset includes 2 million Chinese and English document image-text pairs, 1.5 million chart image-text pairs, and 120,000 natural image negative sample pairs. This dataset merges new visual vocabulary with CLIP vocabulary, making it suitable for tasks like OCR, Markdown/LaTeX conversion, and chart understanding in both Chinese and English contexts.

The GOT model dataset contains about 5 million image-text pairs sourced from Laion-2B, Wukong, and Common Crawl, covering Chinese and English data. It includes 2 million scene-text data points

and 3 million document-level data points, with synthetic datasets supporting tasks such as music score recognition, molecular mathematical expressions, geometric figures, and chart analysis. This diversity positions GOT to address a wide range of OCR tasks, from general document OCR to specialized and fine-grained OCR.

The diversity and complexity of document parsing datasets fuel advancements in document-related algorithms and large models. These datasets provide a broad testing ground for models and offer new solutions for document processing across various fields.

9 Evaluation Metrics

9.1 Metrics for DLA

In document layout detection, the results typically include the coordinate region information and classification of document elements. Therefore, the evaluation metrics for Document Layout Analysis (DLA) emphasize the accuracy of element position recognition, recognition accuracy, and the importance of structural hierarchy to comprehensively reflect the model’s performance in segmenting, recognizing, and reconstructing document structure. For the accuracy of element position recognition, Intersection over Union (IoU) is mainly used to measure the overlap between the predicted and actual boxes. Regarding model recognition accuracy, commonly used metrics include Precision, Recall, and F1-score. Apart from the traditional evaluation metrics mentioned above, adjustments can be made flexibly according to specific analysis goals. In the following sections, for text detection, mathematical expressions, table detection, etc., metrics such as Precision, Recall, F1-score, and IoU are mainly used for evaluation, so detailed introductions will not be provided.

Table 6: Metrics Common Used for DLA

Metric	Definition	Description
IoU	$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$	Measures the overlap between predicted and ground truth boxes.
ReCall	$\text{ReCall} = \frac{TP}{TP + FN}$	Measures how many true positive samples are correctly predicted by the model.
mAP	$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i$	Average precision across all classes, assessing overall model performance.
mAP@IoU[a:b]	$\text{mAP@IoU[a:b]} = \frac{1}{M} \sum_{j=1}^M \text{mAP}_{\text{IoU}_j}$	Computes over a range of IoU thresholds [a, b], calculating at specified intervals and averaged.
F1-score	$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Balances precision and recall and useful in imbalanced class scenarios.

9.2 Metrics for OCR

Text detection and text recognition are two crucial steps in the OCR task, each with different evaluation metrics. Text detection focuses more on localization accuracy and coverage, primarily using precision, recall, F1 score, and IoU to evaluate performance. In contrast, text recognition emphasizes the correctness of the recognition results and is typically assessed using character error rate, word error rate, edit distance, and BLEU score. In projects like LOCR [74], METEOR is also introduced to compensate for some of BLEU’s shortcomings, providing a more comprehensive evaluation of the similarity between machine-generated text and reference text.

9.3 Metrics for MER

Although mathematic expression can be evaluated using OCR task metrics after being converted into formatted code, BLEU, edit distance, and ExpRate are the most commonly used evaluation metrics in the current field of mathematical expression recognition, each with its own limitations. Since mathematical expression can have multiple valid representations, metrics solely relying on text matching cannot fairly and accurately assess recognition results. Some studies have attempted to apply image evaluation metrics to mathematical expression recognition, but the results have not been ideal [386]. Evaluating the results of mathematical expression recognition remains an area that

Table 7: Metrics Common Used for OCR

Metric	Definition	Description
CER	$\text{CER} = \frac{S + D + I}{N}$	Measures the character-level discrepancy between recognized and ground truth text, suitable for OCR tasks requiring high precision.
Edit Distance	$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \text{Cost}(s_1[i], s_2[j]) \end{cases}$	Measures the minimum edit distance needed to convert recognized text into ground truth text.
BLEU	$\text{BLEU} = BP \times \exp \left(\sum_{n=1}^N w_n \log P_n \right)$	Measures the minimum edit distance needed to convert recognized text into ground truth text.
METEOR	$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty})$	Accounts for both precision and recall, and supports stem and synonym matching.
ROUGE-N	$\text{ROUGE-N} = \frac{\sum_{\text{ngram} \in \text{Reference}} \min(\text{Count}_{\text{match}}(\text{ngram}), \text{Count}_{\text{candidate}}(\text{ngram}))}{\sum_{\text{ngram} \in \text{Reference}} \text{Count}_{\text{reference}}(\text{ngram})}$	An improved version of BLEU that focuses on recall rather than precision.

requires further exploration and development. [387] proposed Character Detection Matching (CDM), a metric that eliminates issues arising from different LaTeX representations, offering a more intuitive, accurate, and fair evaluation approach.

Table 8: Metrics Common Used for MER

Metric	Definition	Description
ExpRate	$\text{ExpRate} = \frac{\text{Number of exact matches}}{\text{Total number of samples}}$	Measures the proportion of samples that are completely correct, suitable for scenarios requiring high accuracy.
MSE	$\text{MSE} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (I(i, j) - K(i, j))^2$	Measures the average squared difference between corresponding pixels in two images.
SSIM	$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$	Measures the structural similarity of images, taking into account brightness, contrast, and structural information.
CDM	$\text{CDM} = \frac{2 \times TP}{2 \times TP + FP + FN}$	Converts LaTeX mathematical expression into image and matches it with the corresponding image structure.

9.4 Metrics for Table Recognition

In table detection tasks, in addition to common character-level recall, precision, and F1-score, purity and completeness can also be used for detection. Table structure recognition mainly focuses on analyzing the layout structure inside the table and the relationships between cells. Besides traditional metrics like precision and recall, recently developed detailed evaluation methods provide more dimensions for evaluating table recognition tasks, such as row and column accuracy, multi-column recall (MCR), and multi-row recall (MRR) [388]. With the continuous development of the table recognition field, some universal evaluation metrics have also been proposed, such as cell adjacency relations (CAR) and tree-edit-distance-based similarity (TEDS)[125]. [128] introduced a simplified version of the S-TEDS metric, which only considers the logical structure of tables, ignoring cell content, and focuses on the matching of row, column, spanning row, and spanning column information. The performance evaluation metrics in TGRNet [152] provide several innovative ideas, proposing metrics such as Aall, which describes four logical positions simultaneously, and F_β , which measures comprehensive performance. It also uses weighted average F-score to evaluate the performance of adjacency relation prediction at different IoU thresholds. Tasks involving the conversion of tables into LaTeX or other structured languages, character-level evaluation is typically the primary evaluation method. Alpha-Numeric Tokens Evaluation (AN) assesses the degree of matching between the structured code generated by the model and the alphanumeric symbols in the ground truth. LaTeX Tokens and Non-LaTeX Symbols Evaluation (LT) measures the accuracy of the model in generating LaTeX-specific symbols. Additionally, the Average Levenshtein Distance (ALD) computes the edit distance between the generated structured code and the true value, quantifying the similarity between the two strings. Due to the particularity of table detection and recognition tasks, there is a wide variety of evaluation metrics. Many studies propose different metrics with specific focuses based on their needs. Using a combination of multiple metrics provides a more comprehensive evaluation of

model performance. As the complexity of tasks increases, future evaluation work may rely more on fine-grained evaluation metrics.

Table 9: Metrics Common Used for Table Detection and Recognition

Metric	Definition	Description
Purity	$\text{Purity} = \frac{1}{N} \sum_{i=1}^k \max_j C_i \cap L_j $	Measures the level of noise contained in the detected results.
Completeness	$\text{Completeness} = \frac{1}{N} \sum_{j=1}^k \max_i L_j \cap C_i $	Measure the proportion of table areas detected within the tables.
CAR	$\text{CAR} = \frac{\sum_{i=1}^n 1(\text{predicted adjacency}(C_i) = \text{true adjacency}(C_i))}{n}$	Evaluates boundary detection and relative positioning of table cells, reflecting the structural relationships of the table.
TEDS	$\text{TEDS}(T_1, T_2) = 1 - \frac{\text{TED}(T_1, T_2)}{\max(\text{size}(T_1), \text{size}(T_2))}$	Measures similarity based on tree edit distance, focusing on table structure, including tags and content.
Aall	$\text{Aall} = \frac{ \{i \mid A_{\text{rowSt}}(i) \cap A_{\text{rowEd}}(i) \cap A_{\text{colSt}}(i) \cap A_{\text{colEd}}(i)\} }{N}$	A cell's prediction is considered correct if and only if all four of its logical positions are accurately predicted.
F_beta	$F_{\beta=0.5} = \frac{(1 + 0.5^2) \cdot H \cdot A_{\text{all}}}{0.5^2 \cdot H + A_{\text{all}}}$	Combines spatial positioning and logical accuracy, balancing layout and spatial location evaluation better than F1-score.
WAF	$\text{WAF} = \frac{\sum_{i=1}^4 \text{IoU}_i \cdot F_{\beta=1} @ \text{IoU}_i}{\sum_{i=1}^4 \text{IoU}_i}$	Evaluates adjacency relation prediction based on intersection over union (IoU).

9.5 Metrics for Chart-related Tasks

In chart classification, evaluation metrics are similar to those in standard classification tasks, so we will not detail them here. For chart element detection, metrics like Average IoU, Recall, and Precision are typically used to evaluate the detection of elements (e.g., text areas, bars) [373]. Additionally, for data conversion, metrics like s_0 (visual element detection score), s_1 (average name score for legend matching accuracy), s_2 (average data series score for data conversion accuracy), and s_3 (comprehensive score across all indicators) are employed. These metrics thoroughly assess the effectiveness and robustness of data extraction frameworks for various types of chart data.

The task of extracting data and structure from charts remains underdeveloped, with no standard evaluation metrics established. For instance, in the ChartOCR project, custom metrics are used for different chart types, such as bar, pie, and line charts. Bar chart evaluation uses a distance function between predicted and ground truth bounding boxes, with scores derived from solving an allocation problem. For pie charts, data value importance and order are considered in a sequence matching framework with scores calculated via dynamic programming. ChartDETR uses Precision, Recall, and F1-score.

For line charts, Strict and Relaxed Object Keypoint Similarity metrics are used, offering a balanced perspective incorporating accuracy and flexibility. This method is also adopted by LINEEX.

For charts with structural relationships (e.g., tree diagrams), structured data extraction evaluators modify existing metrics. For instance, in [152], tuples like ownership or subordinate relationships are deemed correct only if all components are accurately extracted, and metrics such as Precision, Recall, and F1 Score are computed.

StructChart [376] introduces the Structuring Chart-oriented Representation Metric (SCRM) for evaluating chart perception tasks. SCRM includes Precision under a fixed similarity threshold and mean Precision (mPrecision) across variable thresholds. The formulas are:

$$\text{Precision}_{\text{IoU}_{\text{thr}, \text{tol}}} = \frac{\sum_{i=1}^L d(i)_{\text{IoU}_{\text{thr}, \text{tol}}}}{L}$$

$$\text{mPrecision}_{\text{tol}} = \frac{\sum_{t=10}^{19} \sum_{i=1}^L d(i, 0.05t)_{\text{tol}}}{10L}$$

Here, L denotes the total number of images, and $d(i)_{\text{IoU}_{\text{thr}, \text{tol}}}$ is the discriminant function, outputting 1 if the IoU of the i -th image meets the threshold within tolerance; otherwise, 0. Similarly, $d(i, 0.05t)_{\text{tol}}$ is another discriminant function for varying thresholds t from 0.5 to 0.95.

In conclusion, chart data and structure extraction tasks present significant developmental opportunities due to diverse and complex evaluation criteria. As research progresses, establishing a comprehensive and universally applicable evaluation system for chart extraction becomes increasingly necessary.

Table 10: Open Source Tools for Document Context Extraction

Tools	Developer	Time	Introduction
GROBID	Patrice Lopez	2011	A machine learning library that focuses on extracting and restructuring original documents, converting them into structured formats such as XML/TEI encoding.
PyMuPDF	Jorj X. McKie	2011	A Python library for extracting, analyzing, converting, and processing data from PDFs and other documents, supporting tables, figures, and other types of content.
doc2text	Joe Sutherland	2016.9	Specializes in extracting low-quality documents; only ensures compatibility in Linux.
pdfplumber	Jeremy Singer-Vine	2019.1	Tools for extraction and parsing of characters, images, lines, tables, and other elements from digital PDF documents.
Parsr	axa-group	2019.8	A tool for cleaning, parsing, and extracting content from various document types, with outputs including JSON, Markdown, CSV/pandasDF, and txt formats.
PP-StructureV2	Baidu	2021.8	Intelligent document analysis system, supports layout analysis of Chinese and English documents, table recognition, and semantic recognition.
DocxChain	Alibaba	2023.9	A system for non-structured or semi-structured document conversion into various information and formats, including complex document applications based on computational capabilities.
pdf2htmlEX	Lu Wang	2023.12	A project to convert PDF documents into HTML format.
MinerU	OpenDataLab	2024.4	A system for extracting content from PDF and converting it into markdown or JSON formats.
PDF-Extract-Kit	OpenDataLab	2024.7	A system based on MinerU to extract various content from PDF, including layout analysis, OCR, table recognition, and formula recognition tasks.
OmniParser	Adithya S Kolavi	2024.6	A platform for extracting and parsing any unstructured data, transforming it into structured, actionable data optimized for GenAI applications.
LLM_aided_ocr	Jeff Emanuel	2024.8	Uses Tesseract for document OCR, followed by LLM-based error correction, with final output in markdown or similar formats.

10 Open Source Tools for Document Extraction

Table 10 lists several open-source document extraction tools with over 1,000 stars on GitHub, designed to handle various document formats and conversion tasks.

Optical Character Recognition (OCR) is a critical aspect of document processing and content extraction. It uses computer vision techniques to detect and extract characters and text from documents, converting images into editable and searchable data. Modern OCR tools have significantly improved in terms of accuracy, speed, and multi-language support. Widely-used general-purpose OCR systems, such as Tesseract and PaddleOCR, have made substantial contributions to this field. Tesseract, an open-source engine, offers robust text recognition and flexible configuration, making it particularly effective for large-scale text extraction tasks. PaddleOCR, with its strong multi-language capabilities, performs exceptionally well in terms of accuracy and speed, especially in complex scenarios.

While general-purpose tools like Tesseract and PaddleOCR are highly effective for document OCR, specialized tools like Unstructured and Zeros demonstrate excellent performance in handling complex document structures, such as nested tables or documents containing both text and images. These tools are particularly adept at extracting structured information.

Beyond OCR, large models have increasingly been employed for document parsing. Recent models such as Nougat, Fox, Vary, and GOT excel at processing complex documents, particularly in PDF format. Nougat, for instance, is tailored for parsing scientific documents and is proficient in extracting formulas and symbols. Fox incorporates multi-modal information, enhancing its effectiveness in semantic understanding and information retrieval. Vary specializes in parsing diverse document formats, including those with embedded images, text boxes, and tables. GOT, a leading model in the OCR 2.0 era, uses a unified end-to-end architecture with advanced visual perception, enabling it to process a wide range of content, such as text, tables, mathematical formulas, molecular structures, and geometric figures. Additionally, OCR at the region level, high-resolution processing, and batch operations for multi-page documents.

Moreover, large multi-modal models commonly utilized in image and language tasks, such as GPT-4, QwenVL, InternVL, and the LLaMA series, can also perform document parsing to a certain extent.

11 Discussion

Modular document parsing systems and Visual-Language Models (VLMs) used for document parsing continue to encounter several challenges and limitations.

Challenges and Future Directions for Pipeline-Based Systems. Pipeline-based document parsing systems face key challenges such as integrating multiple modules, standardizing output formats, and addressing irregular reading orders in complex documents. For instance, systems like MinerU involve intensive pre-processing of input documents, intricate post-processing and specific training for each module to achieve the desired outcome. Additionally, research on document reading order remains limited, with many approaches still relying on rules that struggle with complex layouts, such as multi-column formats. Pipeline systems typically process documents page by page, further limiting their convenience.

The effectiveness of these systems also depends heavily on the performance of individual modules; therefore, advancements in each module are essential for overall system improvement. Despite some progress, specific challenges remain in individual modules:

- **Layout Analysis (DLA):** The accuracy of analyzing complex document layouts with nested elements requires improvement. Future DLA technologies should focus on integrating semantic information to enhance the understanding of fine-grained layouts, such as multi-level heading structures.
- **Document OCR:** Current systems struggle with accurately recognizing large blocks of densely packed text and handling multiple font formats, such as bold and italics. Additionally, balancing general OCR tasks with specialized tasks, like table recognition, remains an issue.
- **Table Detection and Recognition:** The shape of tables significantly impacts detection performance. For example, detecting tables without clear boundaries or those spanning multiple pages remains challenging. In terms of recognition, processing nested tables, tables without cell borders, and cells containing multi-line text still requires improvement.
- **Mathematical expression Recognition:** Detecting and recognizing both inline and multi-line mathematical expressions in documents remains difficult. For printed mathematical expressions, structural extraction needs improvement, while robustness in handling screen-captured mathematical expressions across varying font sizes, noise, and distortions also requires attention. Handwritten mathematical expressions pose additional challenges. Furthermore, current evaluation metrics for mathematical expression recognition are insufficient, calling for more granular and consistent benchmarks.
- **Diagram Extraction:** Diagram extraction from documents is a growing field, but it lacks unified definitions and standard transformation paradigms. Existing methods are often semi-automated or designed for specific diagram types, facing application limitations. End-to-end models are promising, but need to be improved in recognizing diagram elements, OCR, and understanding diagram structure. Although current multimodal large language models (MLLMs) have great potential in handling complex diagram types, they are difficult to integrate into modular document parsing systems.

Challenges and Future Directions for Large Visual Models. In contrast, large visual models for document parsing offer significant advantages by providing end-to-end solutions that eliminate the need for complex module connections and post-processing. They also address some of the limitations of pipeline-based systems in understanding document structure and producing outputs with greater semantic coherence. However, large models are not without their own challenges.

- **Performance Limitations:** Most notably, large models for document parsing do not consistently outperform modular systems, particularly in distinguishing page elements like headers and footers or handling high-density text and complex table structures. This is partly due to the lack of fine-tuned models for tasks involving complex documents and high-resolution images.
- **Frozen Parameters and OCR Capabilities:** Especially, many LVMs freeze LLM parameters during training, hindering their optical character recognition (OCR) capabilities when

dealing with extensive text. Although current models excel at encoding document images, challenges like repeated output and format errors in long document generation remain. These can be mitigated by developing better decoding strategies or employing regularization techniques.

- **Resource Efficiency:** Large models also have significant training and deployment costs, and their inefficiency in processing high-density text results in considerable resource waste. When handling large volumes of text, existing methods for aligning image and text features are insufficient, particularly in dense document formats like A4-sized Documents. While large models inherently require extensive parameters, optimization through architectural improvements and data enhancement can help reduce their size.

Beyond technical challenges, current document parsing research often focuses on structured document types, such as scientific papers and textbooks, leaving more complex documents like instructions, posters, and newspapers underexplored. This narrow focus limits the field’s overall development. There is a need for larger and more diverse datasets to support both training and evaluation efforts.

12 Conclusion

This paper provides a detailed overview of document parsing, covering both modular systems and large models. It reviews datasets, evaluation metrics, and open-source tools, and highlights current limitations in the field. Document parsing technology is of growing interest due to its wide range of applications, such as retrieval-augmented generation (RAG), information storage, and as a source of training data. While modular systems are widely used, end-to-end large models show great potential for future development. In the future, document parsing is expected to evolve into a more accurate, multi-language, and downstream-task-friendly technology that supports diverse OCR tasks.

References

- [1] Cong Yao. Docxchain: A powerful open-source toolchain for document parsing and beyond. *arXiv preprint arXiv:2310.12430*, 2023.
- [2] Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. Visualwordgrid: information extraction from scanned documents using a multimodal approach. In *International Conference on Document Analysis and Recognition*, pages 389–402. Springer, 2021.
- [3] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020.
- [4] Dipali Baviskar, Swati Ahirrao, Vidyasagar Potdar, and Ketan Kotecha. Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, 9:72894–72936, 2021.
- [5] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024.
- [6] Adel Got, Djaafar Zouache, Abdelouahab Moussaoui, Laith Abualigah, and Ahmed Alsayat. Improved manta ray foraging optimizer-based svm for feature selection problems: a medical case study. *Journal of Bionic Engineering*, 21(1):409–425, 2024.
- [7] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.
- [8] Demiao Lin. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *arXiv preprint arXiv:2401.12599*, 2024.
- [9] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.
- [10] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7092–7101, 2023.
- [11] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2023.
- [12] Roldano Cattoni, Tarcisio Coianiz, Stefano Messelodi, and Carla Maria Modena. Geometric layout analysis techniques for document image understanding: a review. *ITC-irst Technical Report*, 9703(09), 1998.
- [13] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document structure analysis algorithms: a literature survey. *Document recognition and retrieval X*, 5010:197–207, 2003.
- [14] Galal M Binmakhashen and Sabri A Mahmoud. Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36, 2019.
- [15] Sakshi and Vinay Kukreja. Machine learning and non-machine learning methods in mathematical recognition systems: Two decades’ systematic literature review. *Multimedia Tools and Applications*, 83(9):27831–27900, 2024.
- [16] Ridhi Aggarwal, Shilpa Pandey, Anil Kumar Tiwari, and Gaurav Harit. Survey of mathematical expression recognition for printed and handwritten documents. *IETE Technical Review*, 39(6): 1245–1253, 2022.

- [17] Vinay Kukreja et al. Recent trends in mathematical expressions recognition: An l₁-based analysis. *Expert Systems with Applications*, 213:119028, 2023.
- [18] Mahmoud Kasem, Abdelrahman Abdallah, Alexander Berendeyev, Ebrahim Elkady, Mohamed Mahmoud, Mahmoud Abdalla, Mohamed Hamada, Sebastiano Vascon, Daniyar Nurseitov, and Islam Taj-Eddin. Deep learning for table detection and structure recognition: A survey. *ACM Computing Surveys*, 2022.
- [19] Mohammad Minouei, Khurram Azeem Hashmi, Mohammad Reza Soheili, Muhammad Zeshan Afzal, and Didier Stricker. Continual learning for table detection in document images. *Applied Sciences*, 12(18):8969, 2022.
- [20] Chixiang Ma, Weihong Lin, Lei Sun, and Qiang Huo. Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognition*, 133:109006, 2023.
- [21] Kenny Davila, Srirangaraj Setlur, David Doermann, Bhargava Urala Kota, and Venu Govindaraju. Chart mining: A survey of methods for automated chart analysis. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3799–3819, 2020.
- [22] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022.
- [23] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*, 2019.
- [24] Siwen Luo, Yihao Ding, Siqu Long, Josiah Poon, and Soyeon Caren Han. Doc-gcn: Heterogeneous graph convolutional networks for document layout analysis. *arXiv preprint arXiv:2208.10970*, 2022.
- [25] Jilin Wang, Michael Krundick, Baojia Tong, Hamima Halim, Maxim Sokolov, Vadym Barda, Delphine Vendryes, and Chris Tanner. A graphical approach to document layout analysis. In *International Conference on Document Analysis and Recognition*, pages 53–69. Springer, 2023.
- [26] Timo I Denk and Christian Reisswig. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*, 2019.
- [27] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19462–19472, 2023.
- [28] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 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*, pages 1192–1200, 2020.
- [29] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020.
- [30] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.
- [31] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition-ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 115–130. Springer, 2021.
- [32] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021.

- [33] Mengxi Wei, Yifan He, and Qiong Zhang. Robust layout-aware ie for visually rich documents with pre-trained language models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2367–2376, 2020.
- [34] Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. *arXiv preprint arXiv:2009.14457*, 2020.
- [35] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024.
- [36] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [37] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.
- [38] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 56–72. Springer, 2016.
- [39] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708, 2020.
- [40] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021.
- [41] Zhuoyao Zhong, Lianwen Jin, and Shuangping Huang. Deeptext: A new approach for text proposal generation and text detection in natural images. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1208–1212. IEEE, 2017.
- [42] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.
- [43] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R 2 cnn: Rotational region cnn for arbitrarily-oriented scene text detection. In *2018 24th International conference on pattern recognition (ICPR)*, pages 3610–3615. IEEE, 2018.
- [44] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE transactions on multimedia*, 20(11):3111–3122, 2018.
- [45] Qiangpeng Yang, Mengli Cheng, Wenmeng Zhou, Yan Chen, Minghui Qiu, Wei Lin, and Wei Chu. Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection. *arXiv preprint arXiv:1805.01167*, 2018.
- [46] Yuliang Liu, Tong He, Hao Chen, Xinyu Wang, Canjie Luo, Shuaitao Zhang, Chunhua Shen, and Lianwen Jin. Exploring the capacity of sequential-free box discretization network for omnidirectional scene text detection. *arXiv preprint arXiv:1912.09629*, 3:15, 2019.
- [47] Shanyu Xiao, Liangrui Peng, Ruijie Yan, Keyu An, Gang Yao, and Jaesik Min. Sequential deformation for accurate scene text detection. In *European Conference on Computer Vision*, pages 108–124. Springer, 2020.

- [48] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [49] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8440–8449, 2019.
- [50] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019.
- [51] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9038–9045, 2019.
- [52] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4234–4243, 2019.
- [53] Jiachen Li, Yuan Lin, Rongrong Liu, Chiu Man Ho, and Humphrey Shi. Rsca: Real-time segmentation-based context-aware scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2349–2358, 2021.
- [54] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.
- [55] Tao Sheng, Jie Chen, and Zhouhui Lian. Centripetaltext: An efficient text instance representation for scene text detection. *Advances in Neural Information Processing Systems*, 34:335–346, 2021.
- [56] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Adaptive boundary proposal network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1305–1314, 2021.
- [57] Shi-Xue Zhang, Chun Yang, Xiaobin Zhu, and Xu-Cheng Yin. Arbitrary shape text detection via boundary transformer. *IEEE Transactions on Multimedia*, 26:1747–1760, 2023.
- [58] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could be better than all: Feature sampling and grouping for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4572, 2022.
- [59] Sibong Song, Jianqiang Wan, Zhibo Yang, Jun Tang, Wenqing Cheng, Xiang Bai, and Cong Yao. Vision-language pre-training for boosting scene text detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15681–15691, 2022.
- [60] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.
- [61] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8714–8721, 2019.
- [62] Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12120–12127, 2020.

- [63] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.
- [64] Michal Busta, Lukas Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE international conference on computer vision*, pages 2204–2212, 2017.
- [65] Lamia Mosbah, Ikram Moalla, Tarek M Hamdani, Bilel Neji, Taha Beyrouthy, and Adel M Alimi. Adocrnet: A deep learning ocr for arabic documents recognition. *IEEE Access*, 2024.
- [66] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative rectification. *Cornell University Library*, pages 1–8, 2018.
- [67] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5571–5579, 2018.
- [68] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019.
- [69] Fenfen Sheng, Zhineng Chen, and Bo Xu. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 781–786. IEEE, 2019.
- [70] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8610–8617, 2019.
- [71] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, pages 319–334. Springer, 2021.
- [72] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 546–547, 2020.
- [73] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.
- [74] Yu Sun, Dongzhan Zhou, Chen Lin, Conghui He, Wanli Ouyang, and Han-Sen Zhong. Locr: Location-guided transformer for optical character recognition. *arXiv preprint arXiv:2403.02127*, 2024.
- [75] Hui Jiang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Yi Niu, Wenqi Ren, Fei Wu, and Wenming Tan. Reciprocal feature learning via explicit and implicit tasks in scene text recognition. In *International Conference on Document Analysis and Recognition*, pages 287–303. Springer, 2021.
- [76] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Chenxia Li, Yuning Du, and Yugang Jiang. Context perception parallel decoder for scene text recognition. *arXiv preprint arXiv:2307.12270*, 2023.
- [77] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12113–12122, 2020.
- [78] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13528–13537, 2020.

- [79] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 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*, pages 7098–7107, 2021.
- [80] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021.
- [81] Mohamed Ali Souibgui, Sanket Biswas, Andres Mafla, Ali Furkan Biten, Alicia Fornés, Yousri Kessentini, Josep Lladós, Lluís Gomez, and Dimosthenis Karatzas. Text-diae: a self-supervised degradation invariant autoencoder for text recognition and document enhancement. In *proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2330–2338, 2023.
- [82] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022.
- [83] Wataru Ohyama, Masakazu Suzuki, and Seiichi Uchida. Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. *IEEE Access*, 7: 144030–144042, 2019.
- [84] Bui Hai Phong, Thang Manh Hoang, and Thi-Lan Le. A hybrid method for mathematical expression detection in scientific document images. *IEEE Access*, 8:83663–83684, 2020.
- [85] Parag Mali, Puneeth Kukkadapu, Mahshad Mahdavi, and Richard Zanibbi. Scanssd: Scanning single shot detector for mathematical formulas in pdf document images. *arXiv preprint arXiv:2003.08005*, 2020.
- [86] Yuxiang Zhong, Xianbiao Qi, Shanjun Li, Dengyi Gu, Yihao Chen, Peiyang Ning, and Rong Xiao. 1st place solution for icdar 2021 competition on mathematical formula detection. *arXiv preprint arXiv:2107.05534*, 2021.
- [87] Junaid Younas, Syed Tahseen Raza Rizvi, Muhammad Imran Malik, Faisal Shafait, Paul Lukowicz, and Sheraz Ahmed. Ffd: Figure and formula detection from document images. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2019.
- [88] Junaid Younas, Shoaib Ahmed Siddiqui, Mohsin Munir, Muhammad Imran Malik, Faisal Shafait, Paul Lukowicz, and Sheraz Ahmed. Fi-fo detector: figure and formula detection using deformable networks. *Applied Sciences*, 10(18):6460, 2020.
- [89] Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Cascade network with deformable composite backbone for formula detection in scanned document images. *Applied Sciences*, 11(16):7610, 2021.
- [90] Minh-Thang Nguyen, Thi-Lan Le, Lan Huong Nguyen Thi, and Thu Ha Nguyen. Ds-yolov5: Deformable and scalable yolov5 for mathematical formula detection in scientific documents. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE, 2021.
- [91] Kai Hu, Zhuoyao Zhong, Lei Sun, and Qiang Huo. Mathematical formula detection in document images: A new dataset and a new approach. *Pattern Recognition*, 148:110212, 2024.
- [92] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, pages 980–989. PMLR, 2017.
- [93] Anh Duc Le, Bipin Indurkha, and Masaki Nakagawa. Pattern generation strategies for improving recognition of handwritten mathematical expressions. *Pattern Recognition Letters*, 128:255–262, 2019.

- [94] Jianshu Zhang, Jun Du, and Lirong Dai. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2245–2250. IEEE, 2018.
- [95] Zhe Li, Lianwen Jin, Songxuan Lai, and Yecheng Zhu. Improving attention-based handwritten mathematical expression recognition with scale augmentation and drop attention. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 175–180. IEEE, 2020.
- [96] Bin Wang, Zhuangcheng Gu, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*, 2024.
- [97] Wei Zhang, Zhiqiang Bai, and Yuesheng Zhu. An improved approach based on cnn-rnns for mathematical expression recognition. In *Proceedings of the 2019 4th international conference on multimedia systems and signal processing*, pages 57–61, 2019.
- [98] Wenqi Zhao, Liangcai Gao, Zuoyu Yan, Shuai Peng, Lin Du, and Ziyin Zhang. Handwritten mathematical expression recognition with bidirectionally trained transformer. In *Document analysis and recognition–ICDAR 2021: 16th international conference, Lausanne, Switzerland, September 5–10, 2021, proceedings, part II 16*, pages 570–584. Springer, 2021.
- [99] Wenqi Zhao and Liangcai Gao. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *European conference on computer vision*, pages 392–408. Springer, 2022.
- [100] Bohan Li, Ye Yuan, Dingkan Liang, Xiao Liu, Zhilong Ji, Jinfeng Bai, Wenyu Liu, and Xiang Bai. When counting meets hmer: counting-aware network for handwritten mathematical expression recognition. In *European conference on computer vision*, pages 197–214. Springer, 2022.
- [101] Jianhua Zhu, Liangcai Gao, and Wenqi Zhao. Ical: Implicit character-aided learning for enhanced handwritten mathematical expression recognition. In *International Conference on Document Analysis and Recognition*, pages 21–37. Springer, 2024.
- [102] Chungkwong Chan. Stroke extraction for offline handwritten mathematical expression recognition. *IEEE Access*, 8:61565–61575, 2020.
- [103] Jiaming Wang, Jun Du, Jianshu Zhang, and Zi-Rui Wang. Multi-modal attention network for handwritten mathematical expression recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1181–1186. IEEE, 2019.
- [104] Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang. A table detection method for pdf documents based on convolutional neural networks. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292. IEEE, 2016.
- [105] Azka Gilani, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. Table detection using deep learning. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 771–776. IEEE, 2017.
- [106] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 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)*, volume 1, pages 1162–1167. IEEE, 2017.
- [107] Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. Decnt: Deep deformable cnn for table detection. *IEEE access*, 6:74151–74161, 2018.
- [108] Yilun Huang, Qinqin Yan, Yibo Li, Yifan Chen, Xiong Wang, Liangcai Gao, and Zhi Tang. A yolo-based table detection method. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 813–818. IEEE, 2019.
- [109] Bin Xiao, Murat Simsek, Burak Kantarci, and Ala Abu Alkheir. Table detection for visually rich document images. *Knowledge-Based Systems*, 282:111080, 2023.

- [110] Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. Deeptabstr: Deep learning based table structure recognition. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1403–1409. IEEE, 2019.
- [111] Yajun Zou and Jinwen Ma. A deep semantic segmentation model for image-based table structure recognition. In *2020 15th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 274–280. IEEE, 2020.
- [112] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 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)*, pages 128–133. IEEE, 2019.
- [113] Jiawei Wang, Weihong Lin, Chixiang Ma, Mingze Li, Zheng Sun, Lei Sun, and Qiang Huo. Robust table structure recognition with dynamic queries enhanced detection transformer. *Pattern Recognition*, 144:109817, 2023.
- [114] Saqib Ali Khan, Syed Muhammad Daniyal Khalid, Muhammad Ali Shahzad, and Faisal Shafait. Table structure extraction with bi-directional gated recurrent unit networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1366–1371. IEEE, 2019.
- [115] Zhenrong Zhang, Jianshu Zhang, Jun Du, and Fengren Wang. Split, embed and merge: An accurate table structure recognizer. *Pattern Recognition*, 126:108565, 2022.
- [116] Weihong Lin, Zheng Sun, Chixiang Ma, Mingze Li, Jiawei Wang, Lei Sun, and Qiang Huo. Tsrformer: Table structure recognition with transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6473–6482, 2022.
- [117] Nam Quan Nguyen, Anh Duy Le, Anh Khoa Lu, Xuan Toan Mai, and Tuan Anh Tran. Formerge: Recover spanning cells in complex table structure using transformer network. In *International Conference on Document Analysis and Recognition*, pages 522–534. Springer, 2023.
- [118] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 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*, pages 572–573, 2020.
- [119] Sachin Raja, Ajoy Mondal, and CV Jawahar. Visual understanding of complex table structures from document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2299–2308, 2022.
- [120] Duc-Dung Nguyen. Tablesegnet: a fully convolutional network for table detection and segmentation in document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(1):1–14, 2022.
- [121] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 944–952, 2021.
- [122] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019.
- [123] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. Rethinking table recognition using graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 142–147. IEEE, 2019.
- [124] Yuntian Deng, David Rosenberg, and Gideon Mann. Challenges in end-to-end neural scientific table recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 894–901. IEEE, 2019.

- [125] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020.
- [126] Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. Pingan-vcgroup’s solution for icdar 2021 competition on scientific literature parsing task b: table recognition to html. *arXiv preprint arXiv:2105.01848*, 2021.
- [127] Honglin Wan, Zongfeng Zhong, Tianping Li, Huaxiang Zhang, and Jiande Sun. Contextual transformer sequence-based recognition network for medical examination reports. *Applied Intelligence*, 53(14):17363–17380, 2023.
- [128] Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11134–11143, 2023.
- [129] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [130] Paulo Chagas, Rafael Akiyama, Aruanda Meiguins, Carlos Santos, Filipe Saraiva, Bianchi Meiguins, and Jefferson Moraes. Evaluation of convolutional neural network architectures for chart image classification. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.
- [131] Wenjing Dai, Meng Wang, Zhibin Niu, and Jiawan Zhang. Chart decoder: Generating textual and numeric information from chart images automatically. *Journal of Visual Languages & Computing*, 48:101–109, 2018.
- [132] Tiago Araújo, Paulo Chagas, Joao Alves, Carlos Santos, Beatriz Sousa Santos, and Bianchi Serique Meiguins. A real-world approach on the problem of chart recognition using classification, detection and perspective correction. *Sensors*, 20(16):4370, 2020.
- [133] Jennil Thiyam, Sanasam Ranbir Singh, and Prabin K Bora. Chart classification: an empirical comparative study of different learning models. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021.
- [134] Anurag Dhote, Mohammed Javed, and David S Doermann. A survey and approach to chart classification. In *International Conference on Document Analysis and Recognition*, pages 67–82. Springer, 2023.
- [135] Jennil Thiyam, Sanasam Ranbir Singh, and Prabin Kumar Bora. Chart classification: a survey and benchmarking of different state-of-the-art methods. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(1):19–44, 2024.
- [136] Kenny Davila, Fei Xu, Saleem Ahmed, David A Mendoza, Srirangaraj Setlur, and Venu Govindaraju. Icp2022: Challenge on harvesting raw tables from infographics (chart-infographics). In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4995–5001. IEEE, 2022.
- [137] Anurag Dhote, Mohammed Javed, and David S Doermann. Swin-chart: An efficient approach for chart classification. *Pattern Recognition Letters*, 185:203–209, 2024.
- [138] Nour Shaheen, Tamer Elsharnouby, and Marwan Torki. C2f-chart: A curriculum learning approach to chart classification. *arXiv preprint arXiv:2409.04683*, 2024.
- [139] Piotr Adam Praczyk and Javier Nogueras-Iso. Automatic extraction of figures from scientific publications in high-energy physics. *Information Technology and Libraries*, 32(4):25–52, 2013.
- [140] Luis D Lopez, Jingyi Yu, Cecilia Arighi, Catalina O Tudor, Manabu Torii, Hongzhan Huang, K Vijay-Shanker, and Cathy Wu. A framework for biomedical figure segmentation towards image-based document retrieval. *BMC systems biology*, 7:1–16, 2013.

- [141] Emilia Apostolova, Daekeun You, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, and George R Thoma. Image retrieval from scientific publications: Text and image content processing to separate multipanel figures. *Journal of the American Society for Information Science and Technology*, 64(5):893–908, 2013.
- [142] Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. Figureseer: Parsing result-figures in research papers. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 664–680. Springer, 2016.
- [143] Sagnik Ray Choudhury, Shuting Wang, and C Lee Giles. Scalable algorithms for scholarly figure mining and semantics. In *Proceedings of the International Workshop on Semantic Big Data*, pages 1–6, 2016.
- [144] Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. Chartsense: Interactive data extraction from chart images. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 6706–6717, 2017.
- [145] Osama Mustafa, Muhammad Khizer Ali, Momina Moetesum, and Imran Siddiqi. Charteye: A deep learning framework for chart information extraction. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 554–561. IEEE, 2023.
- [146] Daniel Drevon, Sophie R Fursa, and Allura L Malcolm. Intercoder reliability and validity of webplotdigitizer in extracting graphed data. *Behavior modification*, 41(2):323–339, 2017.
- [147] Wenyuan Xue, Dapeng Chen, Baosheng Yu, Yifei Chen, Sai Zhou, and Wei Peng. Chartdetr: A multi-shape detection network for visual chart recognition. *arXiv preprint arXiv:2308.07743*, 2023.
- [148] Mathieu Cliche, David Rosenberg, Dhruv Madeka, and Connie Yee. Scatteract: Automated extraction of data from scatter plots. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 135–150. Springer, 2017.
- [149] Rabah Al-Zaidy and C Giles. A machine learning approach for semantic structuring of scientific charts in scholarly documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4644–4649, 2017.
- [150] Jorge Poco and Jeffrey Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. In *Computer graphics forum*, volume 36, pages 353–363. Wiley Online Library, 2017.
- [151] Lianshan Sun, Hanchao Du, and Tao Hou. Fr-detr: End-to-end flowchart recognition with precision and robustness. *IEEE Access*, 10:64292–64301, 2022.
- [152] Meixuan Qiao, Jun Wang, Junfu Xiang, Qiyu Hou, and Ruixuan Li. Structure diagram recognition in financial announcements. In *International Conference on Document Analysis and Recognition*, pages 20–44. Springer, 2023.
- [153] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023.
- [154] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023.
- [155] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

- [156] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge (january 2024). URL <https://llava-vl.github.io/blog/2024-01-30-llava-next>, 2(5):8.
- [157] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.
- [158] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [159] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [160] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [161] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- [162] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023.
- [163] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
- [164] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.
- [165] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *ACM Multimedia 2024*, 2024.
- [166] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024.
- [167] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2025.
- [168] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.
- [169] Xudong Xie, Liang Yin, Hao Yan, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. Wukong: A large multimodal model for efficient long pdf reading with end-to-end sparse sampling. *arXiv preprint arXiv:2410.05970*, 2024.
- [170] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

- [171] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [172] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.
- [173] George Nagy and Sharad C Seth. Hierarchical representation of optically scanned documents. 1984.
- [174] Don Sylwester and Sharad Seth. A trainable, single-pass algorithm for column segmentation. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 615–618. IEEE, 1995.
- [175] Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. Document page decomposition by the bounding-box project. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 1119–1122. IEEE, 1995.
- [176] Lawrence O’Gorman. The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11):1162–1173, 1993.
- [177] Anikó Simon, J-C Pret, and A Peter Johnson. A fast algorithm for bottom-up document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):273–277, 1997.
- [178] Tuan Anh Tran, In-Seop Na, and Soo-Hyung Kim. Hybrid page segmentation using multilevel homogeneity structure. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, pages 1–6, 2015.
- [179] Irina Rabaev, Ofer Biller, Jihad El-Sana, Klara Kedem, and Itshak Dinstein. Text line detection in corrupted and damaged historical manuscripts. In *2013 12th International Conference on Document Analysis and Recognition*, pages 812–816. IEEE, 2013.
- [180] Friedrich M Wahl, Kwan Y Wong, and Richard G Casey. Block segmentation and text extraction in mixed text/image documents. *Computer graphics and image processing*, 20(4): 375–390, 1982.
- [181] Zhixin Shi and Venu Govindaraju. Line separation for complex document images using fuzzy runlength. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 306–312. IEEE, 2004.
- [182] Nikos Nikolaou, Michael Makridis, Basilis Gatos, Nikolaos Stamatopoulos, and Nikos Papamarkos. Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths. *Image and Vision Computing*, 28(4):590–604, 2010.
- [183] Wassim Swaileh, Kamel Ait Mohand, and Thierry Paquet. Multi-script iterative steerable directional filtering for handwritten text line extraction. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1241–1245. IEEE, 2015.
- [184] Mark J Burge and Gladys Monagan. Using the voronoi tessellation for grouping words and multipart symbols in documents. In *Vision Geometry IV*, volume 2573, pages 116–124. SPIE, 1995.
- [185] Koich Kise, Akinori Sato, and Keinosuke Matsumoto. Document image segmentation as selection of voronoi edges. In *Proceedings Workshop on Document Image Analysis (DIA’97)*, pages 32–39. IEEE, 1997.
- [186] Mudit Agrawal and David Doermann. Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1011–1015. IEEE, 2009.

- [187] Xinyuan Wang, Victor Shea-Jay Huang, Renmiao Chen, Hao Wang, Chengwei Pan, Lei Sha, and Minlie Huang. Blackdan: A black-box multi-objective approach for effective and contextual jailbreaking of large language models. *arXiv preprint arXiv:2410.09804*, 2024.
- [188] Hao Liang, Linzhuang Sun, Jingxuan Wei, Victor Shea-Jay Huang, Linkun Sun, Bihui Yu, Conghui He, and Wentao Zhang. Synth-empathy: Towards high-quality synthetic empathy data. *arXiv preprint arXiv:2407.21669*, 2024.
- [189] Zheng Liu, Hao Liang, Victor Shea-Jay Huang, Wentao Xiong, Qinhan Yu, Conghui He, Bin Cui, and Wentao Zhang. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2024.
- [190] Hao Liang, Jiapeng Li, Tianyi Bai, Victor Shea-Jay Huang, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024.
- [191] Jingkun An, Yinghao Zhu, Zongjian Li, Haoran Feng, Victor Shea-Jay Huang, Bohua Chen, Yemin Shi, and Chengwei Pan. Agfsync: Leveraging ai-generated feedback for preference optimization in text-to-image generation. *arXiv preprint arXiv:2403.13352*, 2024.
- [192] Xijie Huang, Xinyuan Wang, Hantao Zhang, Jiawen Xi, Jingkun An, Hao Wang, and Chengwei Pan. Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. *arXiv preprint arXiv:2405.20775*, 2024.
- [193] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024.
- [194] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.
- [195] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. *arXiv preprint arXiv:2405.02363*, 2024.
- [196] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.
- [197] Yangyang Tian, Chenqiang Gao, and Xiaoming Huang. Table frame line detection in low quality document images based on hough transform. In *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*, pages 818–822. IEEE, 2014.
- [198] G Louloudis, B Gatos, I Pratikakis, and K Halatsis. A block-based hough transform mapping for text line detection in handwritten documents. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.
- [199] Chen Zhang, Yang Liu, and Niu Tie. Forest land resource information acquisition with sentinel-2 image utilizing support vector machine, k-nearest neighbor, random forest, decision trees and multi-layer perceptron. *Forests*, 14(2):254, 2023.
- [200] Zaidah Ibrahim, Dino Isa, Rajprasad Rajkumar, and Graham Kendall. Document zone content classification for technical document images using artificial neural networks and support vector machines. In *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, pages 345–350. IEEE, 2009.
- [201] CS Shin, KI Kim, MH Park, and Hang Joon Kim. Support vector machine-based text detection in digital video. In *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No. 00TH8501)*, volume 2, pages 634–641. IEEE, 2000.

- [202] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [203] Dario Augusto Borges Oliveira and Matheus Palhares Viana. Fast cnn-based document layout analysis. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1173–1180. IEEE, 2017.
- [204] Xiaohan Yi, Liangcai Gao, Yuan Liao, Xiaode Zhang, Runtao Liu, and Zhuoren Jiang. Cnn based page object detection in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 230–235. IEEE, 2017.
- [205] Christoph Wick and Frank Puppe. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 287–292. IEEE, 2018.
- [206] Tobias Grüning, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. A two-stage method for text line detection in historical documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(3):285–302, 2019.
- [207] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [208] Horst Bunke and Kaspar Riesen. Recent advances in graph-based pattern recognition with applications in document analysis. *Pattern Recognition*, 44(5):1057–1067, 2011.
- [209] Xiao-Hui Li, Fei Yin, and Cheng-Lin Liu. Page segmentation using convolutional neural network and graphical model. In *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, pages 231–245. Springer, 2020.
- [210] Charles Jacobs, Wilmot Li, Evan Schrier, David Barger, and David Salesin. Adaptive grid-based document layout. *ACM transactions on graphics (TOG)*, 22(3):838–847, 2003.
- [211] Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*, 2018.
- [212] Xiaohui Zhao, Endi Niu, Zhuo Wu, and Xiaoguang Wang. Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*, 2019.
- [213] Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. Semidocseg: harnessing semi-supervised learning for document layout analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–18, 2024.
- [214] Abdelrahman Abdallah, Daniel Eberharter, Zoe Pfister, and Adam Jatowt. Transformers and language models in form understanding: A comprehensive review of scanned document analysis. *arXiv preprint arXiv:2403.04080*, 2024.
- [215] Md Mutasim Billah Abu Noman Akanda, Maruf Ahmed, AKM Shahariar Azad Rabby, and Fuad Rahman. Optimum deep learning method for document layout analysis in low resource languages. In *Proceedings of the 2024 ACM Southeast Conference*, pages 199–204, 2024.
- [216] Qilin Deng, Mayire Ibrayim, Askar Hamdulla, and Chunhu Zhang. The yolo model that still excels in document layout analysis. *Signal, Image and Video Processing*, 18(2):1539–1548, 2024.
- [217] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition*, 96:106954, 2019.
- [218] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

- [219] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 3304–3308. IEEE, 2012.
- [220] Yi Zheng, Qitong Wang, and Margrit Betke. Deep neural network for semantic-based text recognition in images. *arXiv preprint arXiv:1908.01403*, 2019.
- [221] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2021.
- [222] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5238–5246, 2017.
- [223] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.
- [224] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–83, 2018.
- [225] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 706–722. Springer, 2020.
- [226] Wenhai Wang, Xuebo Liu, Xiaozhong Ji, Enze Xie, Ding Liang, ZhiBo Yang, Tong Lu, Chunhua Shen, and Ping Luo. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 457–473. Springer, 2020.
- [227] Shancheng Fang, Zhendong Mao, Hongtao Xie, Yuxin Wang, Chenggang Yan, and Yongdong Zhang. Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7123–7141, 2022.
- [228] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5020–5029, 2018.
- [229] Wei Feng, Wenhao He, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Textdragon: An end-to-end framework for arbitrary shaped text spotting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9076–9085, 2019.
- [230] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020.
- [231] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Zhibo Yang, Tong Lu, and Chunhua Shen. Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5349–5367, 2021.
- [232] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4593–4603, 2022.
- [233] Roi Ronen, Shahar Tsiper, Oron Anschel, Inbal Lavi, Amir Markovitz, and R Manmatha. Glass: Global to local attention for scene-text spotting. In *European Conference on Computer Vision*, pages 249–266. Springer, 2022.

- [234] Linjie Xing, Zhi Tian, Weilin Huang, and Matthew R Scott. Convolutional character networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9126–9136, 2019.
- [235] Youngmin Baek, Seung Shin, Jeonghun Baek, Sungrae Park, Junyeop Lee, Daehyun Nam, and Hwalsuk Lee. Character region attention for text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 504–521. Springer, 2020.
- [236] Liang Qiao, Sanli Tang, Zhanzhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. Text perceptron: Towards end-to-end arbitrary-shaped text spotting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11899–11907, 2020.
- [237] Liang Qiao, Ying Chen, Zhanzhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. Mango: A mask attention guided one-stage scene text spotter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2467–2476, 2021.
- [238] Pengfei Wang, Chengquan Zhang, Fei Qi, Shanshan Liu, Xiaoqiang Zhang, Pengyuan Lyu, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. Pgnnet: Real-time arbitrarily-shaped text spotting with point gathering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2782–2790, 2021.
- [239] Dezhi Peng, Lianwen Jin, Yuliang Liu, Canjie Luo, and Songxuan Lai. Pagenet: Towards end-to-end weakly supervised page-level handwritten chinese text recognition. *International Journal of Computer Vision*, 130(11):2623–2645, 2022.
- [240] Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiabin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. Spts: single-point text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4272–4281, 2022.
- [241] Yuliang Liu, Jiabin Zhang, Dezhi Peng, Mingxin Huang, Xinyu Wang, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, et al. Spts v2: single-point scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [242] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9519–9528, 2022.
- [243] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a clip model into a scene text detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6978–6988, 2023.
- [244] Weijia Wu, Yuanqiang Cai, Chunhua Shen, Debing Zhang, Ying Fu, Hong Zhou, and Ping Luo. End-to-end video text spotting with transformer. *International Journal of Computer Vision*, 132(9):4019–4035, 2024.
- [245] Hsi-Jian Lee and Jiumn-Shine Wang. Design of a mathematical expression understanding system. *Pattern Recognition Letters*, 18(3):289–298, 1997.
- [246] Afef Kacem, Abdel Belaïd, and Mohamed Ben Ahmed. Embedded formulas extraction. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 676–680. IEEE, 2000.
- [247] Utpal Garain, BB Chaudhuri, and Adrish Ray Chaudhuri. Identification of embedded mathematical expressions in scanned documents. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 384–387. IEEE, 2004.
- [248] Derek M Drake and Henry S Baird. Distinguishing mathematics notation from english text using computational geometry. In *Eighth international conference on document analysis and recognition (ICDAR’05)*, pages 1270–1274. IEEE, 2005.

- [249] Tzu-Yuan Chang, Yusuke Takiguchi, and Minoru Okada. Physical structure segmentation with projection profile for mathematic formulae and graphics in academic paper images. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 1193–1197. IEEE, 2007.
- [250] Utpal Garain. Identification of mathematical expressions in document images. In *2009 10th international conference on document analysis and recognition*, pages 1340–1344. IEEE, 2009.
- [251] Josef B Baker, Alan P Sexton, and Volker Sorge. Towards reverse engineering of pdf documents. 2011.
- [252] Xiaoyan Lin, Liangcai Gao, Zhi Tang, Xuan Hu, and Xiaofan Lin. Identification of embedded mathematical formulas in pdf documents using svm. In *Document recognition and retrieval xix*, volume 8297, pages 93–100. SPIE, 2012.
- [253] Bui Hai Phong, Thang Manh Hoang, and Thi-Lan Le. A new method for displayed mathematical expression detection based on fft and svm. In *2017 4th NAFOSTED Conference on Information and Computer Science*, pages 90–95. IEEE, 2017.
- [254] SP Chowdhury, Sekhar Mandal, Amit Kumar Das, and Bhabatosh Chanda. Automated segmentation of math-zones from document images. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 755–759. IEEE, 2003.
- [255] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [256] Kenichi Iwatsuki, Takeshi Sagara, Tadayoshi Hara, and Akiko Aizawa. Detecting in-line mathematical expressions in scientific documents. In *Proceedings of the 2017 ACM symposium on document engineering*, pages 141–144, 2017.
- [257] Xing Wang and Jyh-Charn Liu. A font setting based bayesian model to extract mathematical expression in pdf files. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 759–764. IEEE, 2017.
- [258] Xing Wang, Zelun Wang, and Jyh-Charn Liu. Bigram label regularization to reduce over-segmentation on inline math expression detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 387–392. IEEE, 2019.
- [259] Robert H Anderson. Syntax-directed recognition of hand-printed two-dimensional mathematics. In *Symposium on interactive systems for experimental applied mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium*, pages 436–459, 1967.
- [260] Kam-Fai Chan and Dit-Yan Yeung. Mathematical expression recognition: a survey. *International Journal on Document Analysis and Recognition*, 3:3–15, 2000.
- [261] Richard Zanibbi, Dorothea Blostein, and James R. Cordy. Recognizing mathematical expressions using tree transformation. *IEEE Transactions on pattern analysis and machine intelligence*, 24(11):1455–1467, 2002.
- [262] Christopher Malon, Seichi Uchida, and Masakazu Suzuki. Mathematical symbol recognition with support vector machines. *Pattern Recognition Letters*, 29(9):1326–1332, 2008.
- [263] Ba-Quy Vuong, Yulan He, and Siu Cheung Hui. Towards a web-based progressive handwriting recognition environment for mathematical problem solving. *Expert Systems with Applications*, 37(1):886–893, 2010.
- [264] Lei Hu and Richard Zanibbi. Hmm-based recognition of online handwritten mathematical symbols using segmental k-means initialization and a modified pen-up/down feature. In *2011 International conference on Document analysis and Recognition*, pages 457–462. IEEE, 2011.
- [265] Stéphane Lavirotte and Loic Pottier. Mathematical formula recognition using graph grammar. In *Document Recognition V*, volume 3305, pages 44–52. SPIE, 1998.

- [266] Kam-Fai Chan and Dit-Yan Yeung. Error detection, error correction and performance evaluation in on-line mathematical expression recognition. *Pattern Recognition*, 34(8):1671–1684, 2001.
- [267] Francisco Álvaro, Joan-Andreu Sánchez, and José-Miguel Benedí. Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. *Pattern Recognition Letters*, 35:58–67, 2014.
- [268] Ahmad-Montaser Awal, Harold Mouchere, and Christian Viard-Gaudin. A global learning approach for an online handwritten mathematical expression recognition system. *Pattern Recognition Letters*, 35:68–77, 2014.
- [269] Francisco Álvaro, Joan-Andreu Sánchez, and José-Miguel Benedí. An integrated grammar-based approach for mathematical expression recognition. *Pattern Recognition*, 51:135–147, 2016.
- [270] Anh Duc Le and Masaki Nakagawa. A system for recognizing online handwritten mathematical expressions by using improved structural analysis. *International Journal on Document Analysis and Recognition (IJ DAR)*, 19:305–319, 2016.
- [271] Liangcai Gao, Xiaohan Yi, Yuan Liao, Zhuoren Jiang, Zuoyu Yan, and Zhi Tang. A deep learning-based formula detection method for pdf documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 553–558. IEEE, 2017.
- [272] Xiao-Hui Li, Fei Yin, and Cheng-Lin Liu. Page object detection from pdf document images by deep structured prediction and supervised clustering. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3627–3632. IEEE, 2018.
- [273] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.
- [274] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2061–2069, 2019.
- [275] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.
- [276] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020.
- [277] Jian-Hua Shu, Fu-Dong Nian, Ming-Hui Yu, and Xu Li. An improved mask r-cnn model for multiorgan segmentation. *Mathematical Problems in Engineering*, 2020(1):8351725, 2020.
- [278] Miao Hu, Yali Li, Lu Fang, and Shengjin Wang. A2-fpn: Attention aggregation based feature pyramid network for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15343–15352, 2021.
- [279] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.
- [280] Serkan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.
- [281] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

- [282] Zengyuan Guo, Yuechen Yu, Pengyuan Lv, Chengquan Zhang, Haojie Li, Zhihui Wang, Kun Yao, Jingtuo Liu, and Jingdong Wang. Trust: An accurate and end-to-end table structure recognizer using splitting-based transformers. *arXiv preprint arXiv:2208.14687*, 2022.
- [283] Tao Zhang, Yi Sui, Shun Yao Wu, Fengjing Shao, and Rencheng Sun. Table structure recognition method based on lightweight network and channel attention. *Electronics*, 12(3):673, 2023.
- [284] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022.
- [285] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [286] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152, 2016.
- [287] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 223–232, 2018.
- [288] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 393–402, 2011.
- [289] Ales Mishchenko and Natalia Vassilieva. Chart image understanding and numerical data extraction. In *2011 Sixth International Conference on Digital Information Management*, pages 115–120. IEEE, 2011.
- [290] Haixia Liu and Tim Brailsford. Reproducing show, attend and tell: Neural image caption generation with visual attention. In *Journal of Physics: Conference Series*, volume 2589, page 012012. IOP Publishing, 2023.
- [291] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*, 2015.
- [292] Sameer Antani, Dina Demner-Fushman, Jiang Li, Balaji V Srinivasan, and George R Thoma. Exploring use of images in clinical articles for decision support in evidence-based medicine. In *Document Recognition and Retrieval XV*, volume 6815, pages 230–239. SPIE, 2008.
- [293] Beibei Cheng, Sameer Antani, R Joe Stanley, and George R Thoma. Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval. In *Document Recognition and Retrieval XVIII*, volume 7874, pages 294–304. SPIE, 2011.
- [294] Daliang Xu, Hao Zhang, Liming Yang, Ruiqi Liu, Gang Huang, Mengwei Xu, and Xuanzhe Liu. Empowering 1000 tokens/second on-device llm prefilling with mllm-npu. *arXiv preprint arXiv:2407.05858*, 2024.
- [295] Weihua Huang, Chew Lim Tan, and Wee Kheng Leow. Associating text and graphics for scientific chart understanding. In *Eighth International Conference on Document Analysis and Recognition (ICDAR’05)*, pages 580–584. IEEE, 2005.
- [296] Weihua Huang and Chew Lim Tan. A system for understanding imaged infographics and its applications. In *Proceedings of the 2007 ACM symposium on Document engineering*, pages 9–18, 2007.
- [297] Sagnik Ray Choudhury, Shuting Wang, Prasenjit Mitra, and C Lee Giles. Automated data extraction from scholarly line graphs. In *Proc. Int. Workshop Graph. Recognit*, 2015.

- [298] Chinmayee Rane, Seshasayee Mahadevan Subramanya, Devi Sandeep Endluri, Jian Wu, and C Lee Giles. Chartreader: Automatic parsing of bar-plots. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 318–325. IEEE, 2021.
- [299] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024.
- [300] Muhammad Yusuf Hassan, Mayank Singh, et al. Lineex: data extraction from scientific line charts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6213–6221, 2023.
- [301] Céres Carton, Aurélie Lemaitre, and Bertrand Couïasnon. Fusion of statistical and structural information for flowchart recognition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1210–1214. IEEE, 2013.
- [302] Marçal Rusinol, Lluís-Pere de las Heras, Joan Mas, Oriol Ramos Terrades, Dimosthenis Karatzas, Anjan Dutta, Gemma Sánchez, and Josep Lladós. Cvc-uab’s participation in the flowchart recognition task of clef-ip 2012. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
- [303] Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544, 2022.
- [304] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.
- [305] Jiawei Wang, Kai Hu, Zhuoyao Zhong, Lei Sun, and Qiang Huo. Detect-order-construct: A tree construction based approach for hierarchical document structure analysis. *arXiv preprint arXiv:2401.11874*, 2024.
- [306] Jianqiang Wan, Sibong Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15641–15653, 2024.
- [307] Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. The impact dataset of historical document images. In *Proceedings of the 2Nd international workshop on historical document imaging and processing*, pages 123–130, 2013.
- [308] Mukkai Krishnamoorthy, George Nagy, Sharad Seth, and Mahesh Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):737–747, 1993.
- [309] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666, 2006.
- [310] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300. IEEE, 2009.
- [311] Rana SM Saad, Randa I Elanwar, NS Abdel Kader, Samia Mashali, and Margrit Betke. Bce-arabic-v1 dataset: Towards interpreting arabic document images for people with visual impairments. In *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–8, 2016.
- [312] Fotini Simistira, Manuel Bouillon, Mathias Seuret, Marcel Würsch, Michele Alberti, Rolf Ingold, and Marcus Liwicki. Icdar2017 competition on layout analysis for challenging medieval manuscripts. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1361–1370. IEEE, 2017.

- [313] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5315–5324, 2017.
- [314] Lorenzo Quirós. Multi-task handwritten document layout analysis. *arXiv preprint arXiv:1806.08852*, 2018.
- [315] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.
- [316] Narendra Sahu and Manoj Sonkusare. A study on optical character recognition techniques. *International Journal of Computational Science, Information Technology and Control Engineering*, 4(1):01–15, 2017.
- [317] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.
- [318] Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I Morariu, Varun Manjunatha, Tong Sun, and Yun Fu. Cross-domain document object detection: Benchmark suite and method. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12915–12924, 2020.
- [319] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.
- [320] Randa Elanwar, Wenda Qin, Margrit Betke, and Derry Wijaya. Extracting text from scanned arabic books: a large-scale benchmark dataset and a fine-tuned faster-r-cnn model. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(4):349–362, 2021.
- [321] Logan Markewich, Hao Zhang, Yubin Xing, Navid Lambert-Shirzad, Zhexin Jiang, Roy Ka-Wei Lee, Zhi Li, and Seok-Bum Ko. Segmentation for document layout analysis: not dead yet. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–11, 2022.
- [322] B Pfitzmann, C Auer, M Dolfi, AS Nassar, and PWJ Staar. Doclaynet: A large humanannotated dataset for document-layout analysis (2022). URL: <https://arxiv.org/abs/2206.1062>.
- [323] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15138–15147, 2023.
- [324] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012.
- [325] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116: 1–20, 2016.
- [326] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016.
- [327] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7:105–122, 2005.

- [328] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluís Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013.
- [329] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012.
- [330] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014.
- [331] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.
- [332] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.
- [333] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90: 337–345, 2019.
- [334] Chee Kheng Ch’ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017.
- [335] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016.
- [336] Ron Litman, Oron Anschel, Shahar Tsiper, Roei Litman, Shai Mazor, and R Manmatha. Scatter: selective context attentional scene text recognizer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11962–11972, 2020.
- [337] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European conference on computer vision*, pages 303–321. Springer, 2022.
- [338] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019.
- [339] Jisheng Liang, Ihsin T Phillips, and Robert M Haralick. Performance evaluation of document layout analysis algorithms on the uw data set. In *Document Recognition IV*, volume 3027, pages 149–160. SPIE, 1997.
- [340] Masakazu Suzuki, Seiichi Uchida, and Akihiro Nomura. A ground-truthed mathematical character and symbol image database. In *Eighth International Conference on Document Analysis and Recognition (ICDAR’05)*, pages 675–679. IEEE, 2005.
- [341] Xiaoyan Lin, Liangcai Gao, Zhi Tang, Xiaofan Lin, and Xuan Hu. Performance evaluation of mathematical formula identification. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 287–291. IEEE, 2012.
- [342] Liangcai Gao, Xiaohan Yi, Zhuoren Jiang, Leipeng Hao, and Zhi Tang. Icdar2017 competition on page object detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1417–1422. IEEE, 2017.

- [343] Mahshad Mahdavi, Richard Zanibbi, Harold Mouchere, Christian Viard-Gaudin, and Utpal Garain. Icdar 2019 crohme+ tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1533–1538. IEEE, 2019.
- [344] Dan Anitei, Joan Andreu Sánchez, José Manuel Fuentes, Roberto Paredes, and José Miguel Benedí. Icdar 2021 competition on mathematical formula detection. In *International Conference on Document Analysis and Recognition*, pages 783–795. Springer, 2021.
- [345] Felix M Schmitt-Koopmann, Elaine M Huang, Hans-Peter Hutter, Thilo Stadelmann, and Alireza Darvishy. Formulanet: A benchmark dataset for mathematical formula detection. *IEEE Access*, 10:91588–91596, 2022.
- [346] Latex-ocr. <https://github.com/lukas-blecher/LaTeX-OCR>.
- [347] Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 791–796. IEEE, 2014.
- [348] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4553–4562, 2022.
- [349] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *2013 12th international conference on document analysis and recognition*, pages 1449–1453. IEEE, 2013.
- [350] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1510–1515. IEEE, 2019.
- [351] Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127. IEEE, 2019.
- [352] Ajoy Mondal, Peter Lipps, and CV Jawahar. Iit-ar-13k: A new dataset for graphical object detection in documents. In *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, pages 216–230. Springer, 2020.
- [353] Wonkyo Seo, Hyung Il Koo, and Nam Ik Cho. Junction-based table detection in camera-captured document images. *International Journal on Document Analysis and Recognition (IJDA)*, 18:47–57, 2015.
- [354] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 113–120, 2010.
- [355] Ihsin Tsaiyun Phillips. User’s reference manual for the uw english/technical document image database iii. *UW-III English/technical document image database manual*, 1996.
- [356] Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 445–449. IEEE, 2012.
- [357] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, 2020.
- [358] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.

- [359] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 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*, pages 697–706, 2021.
- [360] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623, 2022.
- [361] Nam Tuan Ly, Atsuhiko Takasu, Phuc Nguyen, and Hideaki Takeda. Rethinking image-based table recognition using weakly supervised methods. *arXiv preprint arXiv:2303.07641*, 2023.
- [362] Mrinal Haloi, Shashank Shekhar, Nikhil Fande, Siddhant Swaroop Dash, et al. Table detection in the wild: A novel diverse table detection dataset and method. *arXiv preprint arXiv:2209.09207*, 2022.
- [363] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. Tgrnet: A table graph reconstruction network for table structure recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1295–1304, 2021.
- [364] Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data*, 10(1):110, 2023.
- [365] Elvis Koci, Maik Thiele, Josephine Rehak, Oscar Romero, and Wolfgang Lehner. Deco: A dataset of annotated spreadsheets for layout and table recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1280–1285. IEEE, 2019.
- [366] Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, Baocai Yin, Bing Yin, and Cong Liu. Semv2: Table separation line detection based on instance segmentation. *Pattern Recognition*, 149:110279, 2024.
- [367] Yiren Li, Zheng Huang, Junchi Yan, Yi Zhou, Fan Ye, and Xianhui Liu. Gfte: graph-based financial table extraction. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 644–658. Springer, 2021.
- [368] Binbin Tang, Xiao Liu, Jie Lei, Mingli Song, Dapeng Tao, Shuifa Sun, and Fangmin Dong. Deepchart: Combining deep convolutional networks and deep belief networks in chart classification. *Signal Processing*, 124:156–161, 2016.
- [369] Jinglun Gao, Yin Zhou, and Kenneth E Barner. View: Visual information extraction widget for improving chart images accessibility. In *2012 19th IEEE international conference on image processing*, pages 2865–2868. IEEE, 2012.
- [370] Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju, Christopher Tensmeyer, Sumit Shekhar, and Ritwick Chaudhry. Icdar 2019 competition on harvesting raw tables from infographics (chart-infographics). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1594–1599. IEEE, 2019.
- [371] KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.
- [372] Kenny Davila, Chris Tensmeyer, Sumit Shekhar, Hrituraj Singh, Srirangaraj Setlur, and Venu Govindaraju. Icpr 2020-competition on harvesting raw tables from infographics. In *International Conference on Pattern Recognition*, pages 361–380. Springer, 2021.
- [373] Weihong Ma, Hesuo Zhang, Shuang Yan, Guangshun Yao, Yichao Huang, Hui Li, Yaqiang Wu, and Lianwen Jin. Towards an efficient framework for data extraction from chart images. In *International Conference on Document Analysis and Recognition*, pages 583–597. Springer, 2021.

- [374] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925, 2021.
- [375] Joseph Shtok, Sivan Harary, Ophir Azulai, Adi Raz Goldfarb, Assaf Arbelle, and Leonid Karlin-sky. Charter: heatmap-based multi-type chart data extraction. *arXiv preprint arXiv:2111.14103*, 2021.
- [376] Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*, 2023.
- [377] Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Onechart: Purify the chart structural extraction via one auxiliary token. *arXiv preprint arXiv:2404.09987*, 2024.
- [378] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020.
- [379] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.
- [380] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.
- [381] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701*, 2024.
- [382] Karin Verspoor, Dat Quoc Nguyen, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Jiayuan He, and Zenan Zhai. Chemu dataset for information extraction from chemical patents. *Mendeley Data*, 2(10):17632, 2020.
- [383] Shivalika Tanwar, Patrick Auberger, Germain Gillet, Mario DiPaola, Katya Tsaïoun, and Bruno O Villoutreix. A new chembl dataset for the similarity-based target fishing engine fasttargetpred: Annotation of an exhaustive list of linear tetrapeptides. *Data in Brief*, 42: 108159, 2022.
- [384] Jan Hajič and Pavel Pecina. The muscima++ dataset for handwritten optical music recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 39–46. IEEE, 2017.
- [385] Lukas Tuggener, Ismail Elezi, Jurgen Schmidhuber, Marcello Pelillo, and Thilo Stadelmann. Deepscores-a dataset for segmentation, detection and classification of tiny objects. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3704–3709. IEEE, 2018.
- [386] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(1):63–75, 2021.
- [387] Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. Cdm: A reliable metric for fair and accurate formula recognition evaluation. *arXiv preprint arXiv:2409.03643*, 2024.
- [388] Pratik Kayal, Mrinal Anand, Harsh Desai, and Mayank Singh. Tables to latex: structure and content extraction from scientific tables. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(2):121–130, 2023.