Published: 22 June 2024
Online AM: 12 April 2024
Accepted: 02 April 2024
Revised: 26 January 2024
Received: 21 March 2023



SURVEY

# From Detection to Application: Recent Advances in Understanding Scientific Tables and Figures

**JIANI HUANG**, Wuhan University, Wuhan, Hubei, China

**HAIHUA CHEN**, University of North Texas, Denton, TX, United States

**FENGCHANG YU**, Wuhan University, Wuhan, Hubei, China

**WEI LU**, Wuhan University, Wuhan, Hubei, China

# From Detection to Application: Recent Advances in Understanding Scientific Tables and Figures

JIANI HUANG, Wuhan University, Wuhan, China
HAIHUA CHEN, University of North Texas, Denton, United States
FENGCHANG YU, Wuhan University, Wuhan, China
WEI LU, Wuhan University, Wuhan, China

Tables and figures are usually used to present information in a structured and visual way in scientific documents. Understanding the tables and figures in scientific documents is significant for a series of downstream tasks, such as academic search, scientific knowledge graphs, and so on. Existing studies mainly focus on detecting figures and tables from scientific documents, interpreting their semantics, and integrating them into downstream tasks. However, a systematic and comprehensive literature review on the mining and application of tables and figures in academic papers is still missing. In this article, we introduce the research framework and the whole pipeline for understanding tables and figures, including detection, structural analysis, interpretation, and application. We deliver a thorough analysis of benchmark datasets, recent techniques, and their pros and cons. Additionally, a quantitative analysis of the effectiveness of different models on popular benchmarks is presented. We further outline several important applications that exploit the semantics of scientific tables and figures. Finally, we highlight the challenges and some potential directions for future research. We believe this is the first comprehensive survey in understanding scientific tables and figures that covers the landscape from detection to application.

CCS Concepts: • **Applied computing → Graphics recognition and interpretation**; • **General and reference → Surveys and overviews**

Additional Key Words and Phrases: Scientific documents, figure understanding, table understanding

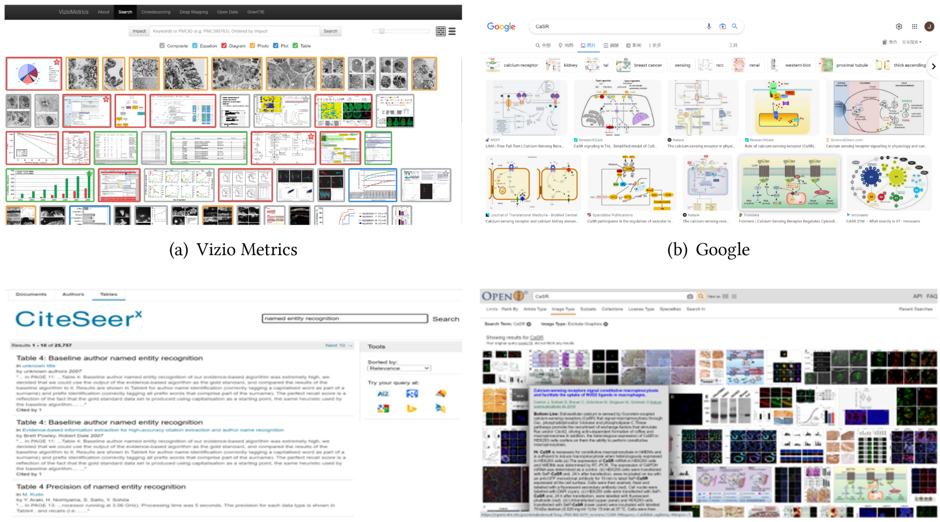
**ACM Reference Format:**
Jiani Huang, Haihua Chen, Fengchang Yu, and Wei Lu. 2024. From Detection to Application: Recent Advances in Understanding Scientific Tables and Figures. *ACM Comput. Surv.* 56, 10, Article 261 (June 2024), 39 pages.
https://doi.org/10.1145/3657285

## 1 INTRODUCTION

The rise in the volume of digitized documents over the last two decades has posed a challenge to traditional manual analysis methods, and AI technologies are bringing document analysis into a new era. Therefore, significant efforts have been made in employing **Natural Language Processing (NLP)** and **Computer Vision (CV)** techniques to tackle the tasks involved in understanding

Authors' Contact Information: Jiani Huang, Wuhan University, Wuhan, Hubei, China; e-mail: huangjiani@whu.edu.cn; Haihua Chen, University of North Texas, Denton, Texas, United States; e-mail: haihua.chen@unt.edu; Fengchang Yu (Corresponding author), Wuhan University, Wuhan, Hubei, China; e-mail: yufc2002@whu.edu.cn; Wei Lu, Wuhan University, Wuhan, Hubei, China; e-mail: weilu@whu.edu.cn.

ACM 0360-0300/2024/06-ART261
https://doi.org/10.1145/3657285



(a) Vizio Metrics (b) Google

(c) CiteSeer (d) Open-i

Fig. 1. Academic table and figure retrieval systems.

document elements. Texts, tables, and figures are the three crucial elements of documents. A table presents data in a structured form, while a figure can eliminate potential language differences due to its intuitive nature. Therefore, semantic analysis in tables and figures receive broad attention in the existing literature [14, 44, 56, 70, 71, 113, 144, 167, 195].

Tables and figures frequently appear in scientific documents. Yu et al. [192] discovered an average of five figures per biomedical paper in **Proceedings of the National Academy of Sciences (PNAS)**. They are typically used to present the experimental setup and results, contextual information, and term definitions. Due to the innovation and credibility of academic papers, the tables and figures in them have higher knowledge density and reliability than in ordinary documents.

Understanding the tables and figures in scientific documents is significant for a series of downstream tasks, such as academic search, scientific knowledgebase construction, and so on. For example, an increasing number of retrieval systems, such as Vizio Metrics,[1] Google,[2] CiteSeer,[3] and Open-i,[4] integrate table and figure retrieval into their functions to enhance search performance, as shown in Figure 1. Moreover, Zhu et al. [204] found that taking the content of figures into account can significantly improve user satisfaction with the informativeness of academic article summaries. Tab2Know [85], a knowledgebase of tables in scientific papers, could assist users in finding answers without accessing the papers. Additionally, it can serve various purposes, such as categorizing papers, identifying inconsistencies, and detecting plagiarized content.

Over the last 30 years, there has been a growing focus within the research community on scientific tables and figures, as illustrated in Figure 2, derived from Web of Science searches using the query “scientific documents figure/table”. We collected the survey paper on understanding tables and figures over the past decade, as presented in Table 1. Previous surveys primarily emphasized either tables or figures. Although Bhatt et al. [13] addressed both figures and tables, their focus is primarily on the detection task. The works of [13, 35, 56, 106, 109], involved figures or tables

[1] http://viziometrics.org/
[2] https://www.google.com
[3] https://citeseer.ist.psu.edu/
[4] https://openi.nlm.nih.gov/

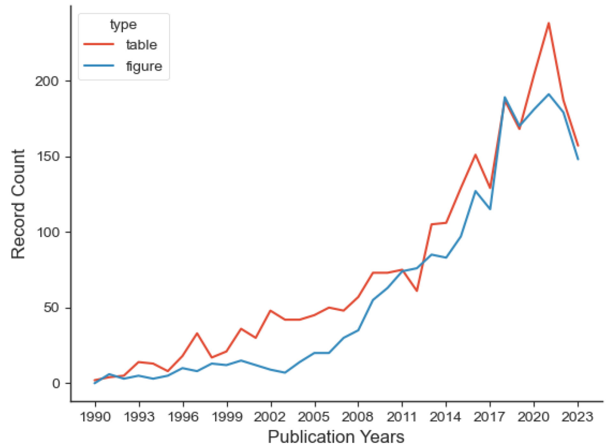



Fig. 2. The number of papers on scientific tables and figures retrieved in the Web of Science, from the period 1990 to 2023.

Table 1. Previous Surveys Related to Table and Figure Understanding within Ten Years

| Survey | Year | Scope | Table /Figure | Task | Experiment Results | Evaluation Metrics | Dataset Summary | Application Summary |
|---|---|---|---|---|---|---|---|---|
| [109] | 2013 | general topics | 🟠 | segmentation, classification, interpretation | ✓ | ✗ | ✗ | ✗ |
| [184] | 2020 | scientific topics | 🟢 | retrieval | ✗ | ✗ | ✗ | ✓ |
| [56] | 2021 | general topics | 🟠 | detection, structure analysis | ✓ | ✓ | ✓ | ✗ |
| [157] | 2021 | statistical topics | 🟢 | interpretation, reasoning | ✓ | ✓ | ✓ | ✓ |
| [13] | 2021 | general topics | 🟠🟢 | detection | ✓ | ✓ | ✓ | ✗ |
| [35] | 2021 | general topics | 🟢 | detection, classification, data extraction | ✗ | ✓ | ✓ | ✓ |
| [106] | 2022 | general topics | 🟠 | interpretation | ✓ | ✓ | ✓ | ✗ |
| [181] | 2022 | visualization | 🟢 | reasoning, assessment, etc. | ✗ | ✗ | ✗ | ✓ |
| [47] | 2023 | medical, scientific topics | 🟢 | interpretation | ✗ | ✓ | ✓ | ✗ |
| [10] | 2023 | statistical topics | 🟢 | classification, data extraction, description generation | ✗ | ✗ | ✗ | ✗ |

🟠 denotes table, while 🟢 is figure. **Scope** describes the scope or area of the figures or tables in this survey.

on general topics, deviating somewhat from scientific figures and tables. Yang et al. [184], Shahira and Lijiya [157], and Farahani et al. [47] focused on scientific or statistical charts but are limited in the range of tasks they cover. Therefore, a systematic and comprehensive literature review on the mining and applying tables and figures in academic papers is still lacking. Prior research has primarily focused on individual subtasks while disregarding the interconnection of various subtasks and applications. Furthermore, the absence of an explicit framework inhibits future research in this area. With the growing interest and work on this topic, it is time for a paper of our kind to:

— define the research framework for understanding figure and table tasks and sort out benchmark datasets built from scientific documents, as well as identify the main evaluation metrics;
— depict the history of research methodologies over time and summarize the performance of competitive models on benchmark datasets to compare the advantages and disadvantages of different methods;
— outline the application of scientific tables and figures in various downstream tasks; and
— identify key challenges to motivate and orient interests in this area effectively.

The survey is outlined as follows: in Section 2, we establish the research framework for understanding tables and figures, dividing it into detection, structure analysis, and interpretation subtasks. Subsequently, Sections 3–5 provide a summary of research on these three subtasks, respectively. Finally, some applications and potential future directions are discussed in Sections 6 and 7.

## 2 RESEARCH FRAMEWORK FOR UNDERSTANDING TABLE AND FIGURE TASKS

In the following section, we formally present the definitions of "table" and "figure" and establish a framework for understanding tables and figures.

In this paper, tables and figures are defined as follows:

- Table: A table is a structured display of data organized in rows and columns, facilitating the systematic presentation, comparison, and analysis of information. Each row signifies a record, while each column represents an attribute.
- Figure: A figure encompasses diverse visual elements, serving as a visual representation of data. Prior research may focus solely on a specific type of figure. Here, we categorize figures into three distinct types:
  - Chart: Charts visually represent quantitative data using axes, labels, and data points to illustrate trends, comparisons, or relationships, such as bar charts, line charts, or pie charts.
  - Diagram: Diagrams employ shapes and lines to illustrate relationships, concepts, or processes, such as flowcharts and Venn diagrams.
  - Image: Images represent real-world scenes or objects through pixel-based representations, including photographs, satellite imagery, and microscopic imagery.

Inspired by Hurst [70], a pipeline of understanding tables and figures in document images can be divided into three main subtasks, as shown in Figure 3.

- `Detection`: detecting tables and figures and returning their coordinates in documents.
- `Structure analysis`: for tables, this task includes identifying the rows, columns, blocks, cells, and data in the table. In addition, the metadata, including notes and titles, are crucial components that interest many researchers. For figures, this task mainly aims at extracting and classifying figure elements such as X-axis, Y-axis, data values, legend, and so on.
- `Interpretation`: extracting the meaningful and unambiguously information; in other words, understanding the semantics of the tables and figures.

The initial step involves utilizing document images as inputs to identify tables and figures during the detection phase, yielding their respective categories and location coordinates. Subsequently, the structural analysis phase is employed to acquire components of the figures or tables, such as cells and rows of tables. In the interpretation phase, the primary objective is to extract meaningful information and comprehend the semantics of figures and tables. Upon completion of these processes, fine-grained mining results for academic tables and figures are obtained. These results can be utilized for various downstream applications, including knowledgebase construction, summary generation, and beyond. In the following sections, we will systematically survey the three steps and the applications of scientific tables and figures, respectively.

## 3 TABLE AND FIGURE DETECTION

Table and figure detection provides a basis for analyzing the structure and extracting semantics from table and figure contents. Next, we summarize the benchmark datasets, popular techniques, and their performances, respectively.

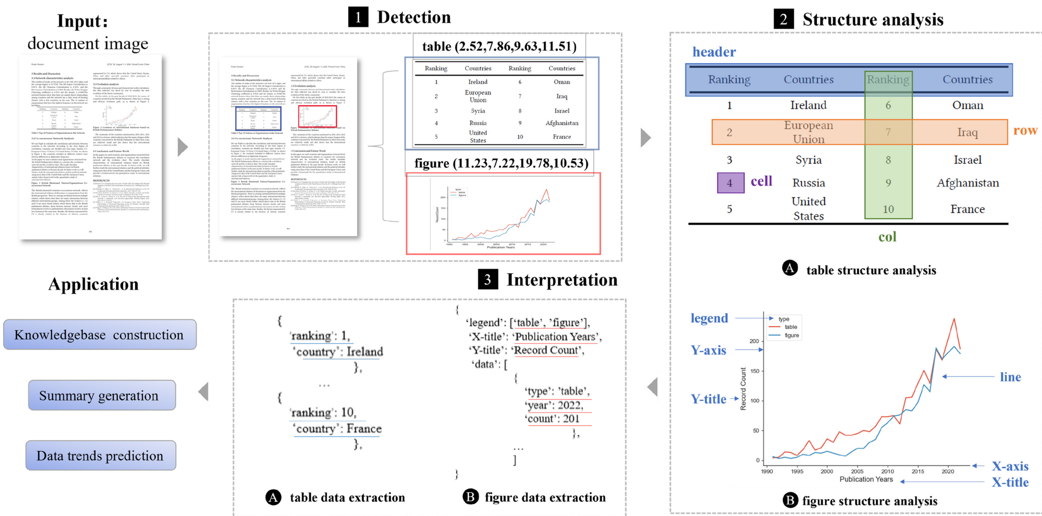



Fig. 3. Pipeline of understanding tables and figures.

Table 2. Available Datasets for Scientific Table and Figure Detection

| Type | Dataset | Source | Format | Size: Figure | Size: Table | Year | Link |
|---|---|---|---|---|---|---|---|
| both | GROTOAP2 [172] | PubMed | PDF,XML | 42,777 | 505,958 | 2014 | Link |
| | CS-150 [31] | CS conferences | PDF,JSON | 458 | 191 | 2015 | Link |
| | CS-Large [30] | Semantic Scholar | PDF, JSON | 957 | 300 | 2016 | Link |
| | DeepFigures [162] | PubMed | LaTeX, XML | 4,095,622 | 1,431,820 | 2018 | Link |
| | Article Regions [166] | PubMed | PDF,XML | 299 | 148 | 2019 | Link |
| | PubLayNet [202] | PubMed | PNG, JSON | 126,938 | 113,128 | 2019 | Link |
| | DocBank [98] | arXiv | LaTeX, PNG | 113,270 | 24,517 | 2020 | Link |
| | IIIT-AR-13K⋆ [130] | Business documents | PNG,XML | 2,948 | 15,981 | 2020 | Link |
| | ScanBank [78] | MIT repository | PNG, JSON | 3,375 in total | | 2021 | Link |
| | ACL-FIG [80] | ACL repository | PNG, JSON | 112,052 | 151,900 | 2023 | Link |
| table | UW3⋆ [141] | Books | PNG,XML | – | 147 | 1996 | Link |
| | Marmot* | Citeseer | PDF | – | 958 | 2012 | Link |
| | TableBank* [96] | arXiv | LaTeX, PNG | – | 253,817 | 2019 | Link |
| | SciTSR [26] | arXiv | LaTeX, JSON | – | 1,500 | 2019 | Link |
| | ICDAR2019* [50] | Websites | PNG, XML | – | 2,371 | 2019 | Link |
| | TNCR⋆ [1] | Websites | JPG, XML | – | 9,428 | 2021 | Link |
| | PubTables-1M [165] | PubMed | PNG, JSON | – | 947,642 | 2021 | Link |
| | FintabNet [199] | Business documents | PDF,JSON | – | 112,887 | 2021 | Link |
| | TabRecSet [183] | Wild scenarios | JPG, JSON | – | 38,177 | 2023 | Link |
| figure | VisImages [39] | IEEE InfoVis and VAST | PNG, JSON, CSV | 12,267 | – | 2022 | Link |

* represents that the dataset includes not only scientific papers but also various domain documents, such as financial records. ⋆ indicates that the dataset is not built on academic papers.

### 3.1 Datasets

High-quality, large-scale datasets are the basis for training a deep learning model. This section introduces publicly available and well-known datasets for figure and table detection. While some of these datasets may not derive from scientific documents, the models trained on them exhibit potential transferability to scientific tables and figures. Consequently, these datasets are included in our survey. Table 2 displays an overview of available datasets for table and figure detection. Considering space limitations, we only introduce popular datasets built upon academic literature in detail.

**DeepFigures.** DeepFigures [162] is derived from the arXiv[5] and PubMed[6] datasets and is composed of 1,400k papers on various subjects. The authors introduce a distantly supervised method to induce high-quality labels for figures and tables. This dataset contains 5.5 million induced labels with a precision of 96.8% on average.

**PubLayNet.** PubLayNet [202] is designed for the document layout analysis task and built from the PubMed dataset. The annotations for tables and figures are generated by matching the PDF and XML formats of papers. This large dataset contains over 1 million PDF articles and 360,000 document images, with 126,938 figures and 113,128 tables in total.

**DocBank.** DocBank [98] is a document layout analysis benchmark, consisting of 500K document pages with 12 types of semantic units, such as table, figure, and so on. According to the authors, DocBank is a natural extension of the TableBank dataset and is fully annotated at the token level.

**ACL-FIG.** Karishma et al. [80] downloaded 55,760 articles from the ACL Anthology repository and developed a pipeline to extract and classify the figures of these papers. They published two datasets; namely, ACL-FIG and ACL-FIG-PILOT. The former includes 112,052 figures and 151,900 tables, while the latter consists of 1,671 figures annotated across 19 distinct figure types, including bar charts, pie charts, and others.

**TableBank.** TableBank [97] is an image-based table detection and recognition dataset containing 417K high-quality annotated tables. The documents within TableBank are sourced from the arXiv dataset and are all in English. TableBank could be utilized for both table detection and table structure recognition tasks.

**SciTSR.** The SciTSR dataset, constructed by Chi et al. [26], contains 15,000 tables from scientific articles and their corresponding high-quality structure labels derived from LaTeX source files. This dataset has many complex tables, with an average of 48 cells, 9 rows, and 5 columns per table. To evaluate the model performance in recognizing complex tables, the authors constructed a test subset named SciTSR-COMP, including 716 complex tables extracted from the test set.

**PubTables-1M.** PubTables-1M [165] is a large, detailed, high-quality dataset for training and evaluating models for table detection, table structure recognition, and functional analysis. It provides nearly one million tables from scientific articles in the PubMed database. PubTables-1M contains rich annotation information, including annotations for projected row headers and bounding boxes for all rows, columns, and cells, even blank cells.

## 3.2 Methods

Based on the model architecture, we categorize previous work into heuristic-based, CNN-based, Transformer-based, and GNN-based. Figure 4 depicts the history and evolution of table and figure detection research. Before 2015, nearly all existing methods were heuristic-based; since 2015, most studies in this area have focused on deep learning techniques. Next, we will outline the different models for the scientific table and figure detection task.

### *3.2.1 Heuristic-based models.*

Previous studies heavily rely on heuristic algorithms and probabilistic models, which are difficult to transfer to academic papers in different disciplines and layouts. Lopez et al. [112] proposed

[5] https://arxiv.org/
[6] https://pubmed.ncbi.nlm.nih.gov/

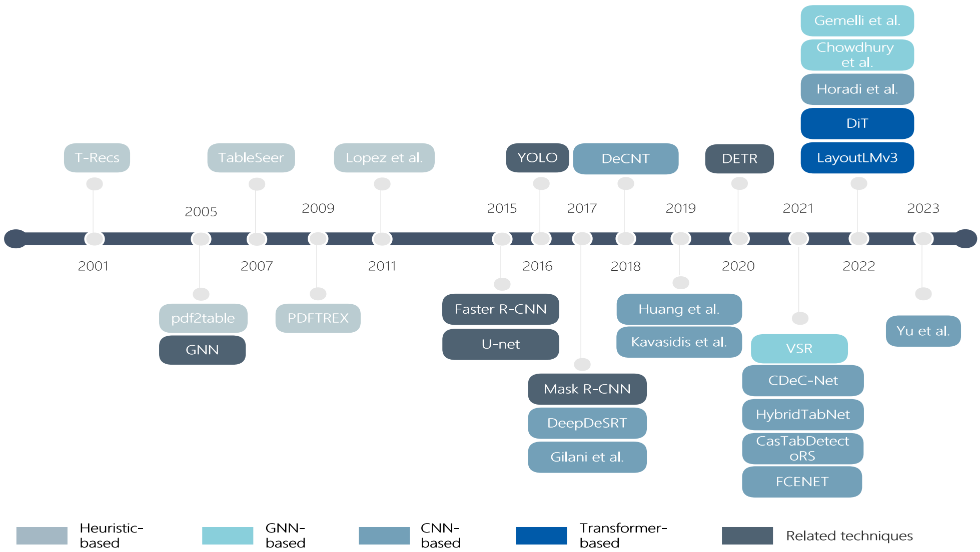



Fig. 4. The history and evolution of table and figure detection techniques.

an automatic system for extracting figures from the biomedical literature. This system exploits PDF stream content and designs several rules to recognize figures. Researchers in the fields of chemistry [27], high energy physics [59], and computer science [31] also investigated heuristic methods. Below we summarize the advantages and disadvantages of heuristic-based models.

— Advantages
  – Heuristic-based models usually perform well on lined tables and regular layouts, with relatively high precision.
  – They are less demanding on computing resources and annotated data.
— Disadvantages
  – Most of them rely on PDF stream content to detect tables and texts; therefore, they cannot handle scanned images.
  – Heuristic-based systems are generally complex and comprised of hand-crafted rules, which makes them less generalizable and lacking in robustness.
  – Heuristic-based methods usually suffer from low recall.

*3.2.2 CNN-Based Models.* The superior performance of **Convolutional Neural Networks (CNN)** in computer vision has prompted researchers to investigate CNN for the table and figure detection task. Object detection and instance segmentation are two branches of this type of research. Object detection has been an active research area in recent years, and its original goal is detecting target objects in natural scene images, which presents notable differences with detecting objects in document images. Document objects, such as tables and figures, typically exhibit a square shape, facilitating their easy distinction from the background. In contrast, natural objects are diverse and may share similar colors with the background. Moreover, the factors influencing detection performance vary between these two tasks. In the context of table and figure detection, the diverse document layouts and object formats play pivotal roles, while natural object detection may be affected by factors like blurriness and illumination. Another noteworthy distinction lies in the precision required for bounding boxes. In table and figure detection tasks, precise bounding boxes are crucial due to the potential lack of titles or legends hindering comprehension. However, missing small parts of natural objects is less likely to impede understanding.

With the development of backbones and networks for object detection and instance segmentation, such as VGG [163], ResNet [59], Faster R-CNN [150], Mask R-CNN [58], and U-Net [152], table and figure detection has achieved state-of-the-art results. Researchers have explored various methods to apply object detection models to the document analysis domain. Gilani et al. [52] fine-tuned the Faster R-CNN model for table detection. To align document images more closely with natural images, the authors employed a pre-processing step, which involves computing three distance metrics between text regions and white spaces, and setting them as the values of RGB channels. In contrast, DeepDeSRT [156] is an end-to-end model without any preprocessing technique but it fails to detect complicated tables. Huang et al. [69] proposed an anchor optimization technique to make anchors used in the YOLOv3 model more suitable for tables rather than natural objects. Chowdhury et al. [29] pretrained an image classifier on document layout datasets as the backbone in the Faster R-CNN model, rather than directly using a backbone trained on natural image datasets, such as ResNet.

Researchers have further explored building more robust models that effectively handle complex tables and diverse layouts. In [2, 135, 160], deformable convolution was widely used to improve the model's ability to handle tables with different layouts. DeCNT [160] replaced the traditional convolution with deformable convolution in the Faster R-CNN model and found that it could adapt to tables of different layouts well. In addition, Agarwal et al. [2] considered that existing models are trained on a fixed IoU threshold, which leads to a noisy detection at higher IoU thresholds. They addressed this issue by proposing the cDeC-Net network, which contains a series of detectors trained with increasing IoU thresholds. Compared with models trained on a single IoU threshold, cDeC-Net [2] achieves high accuracy and tighter bounding box detection at a higher IoU threshold. Although these methods achieve better results, they are more computationally expensive. To solve this problem, Hashmi et al. [57] presented CasTabDetectoRS, which employs a relatively lightweight backbone with **Switchable Atrous Convolution (SAC)** to achieve comparable performance.

Liu et al. [110] initially adopted an instance segmentation model for figure detection, yielding competitive results. Their proposed model, based on BlendMask, integrates horizontal and vertical attention modules to enhance adaptability to document images. Kavasidis et al. [82] proposed a saliency-based CNN model designed for figure and table detection. The authors formulated the detection problem as a semantic image segmentation problem, predicting each pixel's likelihood of being a graphical object. Yu et al. [190] utilized a cascade semantic segmentation model and designed a novel loss function aimed at improving the weighting of boundary parts. This adjustment allows the model to predict complete figures without losing information near the boundary.

The advantages and disadvantages of CNN-based models are:

- Advantages
  - CNN-based models are the most widely used framework for table and figure detection tasks, reporting superior results in all well-known benchmark datasets.
  - In contrast to heuristic-based models, the majority of CNN-based models use document images as input, which is more in line with practical needs.
  - Many techniques are designed to improve the robustness and generalization capability of the model.
- Disadvantages
  - Compared with natural objects, tables exhibit diverse sizes and layouts, with some extreme cases including short and wide, long and narrow tables. Designing appropriate scales and ratios for region proposals in the object detection process becomes challenging due to this variability.

  - These models rely on large-scale labeled datasets and are more computationally intensive. However, there have been relatively few studies that consider inference efficiency.

*3.2.3 Transformer-Based Models.* Originally crafted for NLP tasks, the Transformer is a model architecture that discards recurrent units and relies entirely on an attention mechanism to draw global dependencies between input and output. In contrast to CNN-based models, the Transformer architecture excels in capturing global features while conserving computing resources. Drawing inspiration from the Transformer, researchers proposed **Detection Transformer (DETR)** [18] for object detection. Smock et al. [165] first applied the DETR model to table detection, table structure recognition, and function analysis, and reported promising results. Biswas et al. [15] built a **document image segmentation Transformer (DocSegTr)** to analyze complicated document layouts from an instance segmentation perspective. DocSegTr is more computationally efficient in inference than the state-of-the-art models based on Mask-RCNN. Researchers have tried to pre-train the Transformer model on a large amount of unlabeled image data. Li et al. [95] proposed DiT, a self-supervised pre-trained document image transformer model for general document AI tasks. In the DiT framework, images are randomly masked and split into 16×16 patches, and the learning objective is to recover corrupted image patches. Huang et al. [68] introduced LayoutLMv3 to pre-train multimodal Transformers for document AI with unified text and image masking. They presented a **Word-Patch Alignment (WPA)** objective to learn cross-modal alignment effectively.

The advantages and disadvantages of Transformer-based models are:

- Advantages
  - Pre-trained general document AI models exploit large-scale unlabeled data and thus may have greater generalization capabilities. In addition, information from other page objects (e.g., equations) may help the model distinguish between tables and other unrelated objects.
  - Transformer architecture performs better at capturing global features, which is crucial for tables with multi-rows or multi-columns.
- Disadvantages
  - In contrast to CNN-based models, the Transformer has deficiencies in capturing local information.

*3.2.4 GNN-Based Models.* The inherent structural nature of a table makes it well-suited for representation as a graph. Consequently, researchers explore constructing **graph neural networks (GNN)** that explicitly model tabular structure.

Riba et al. [151] developed a GNN model that formulates document entities as nodes and detects tables by classifying these nodes. In their method, cells are defined as nodes and edges are constructed when a horizontal or vertical line connects cells' bounding boxes. Additionally, Gemelli et al. [51] enriched node and edge representations by adopting static NLP-based embeddings (SciBert [11] and Spacy). Compared with Riba et al. [151], who exclusively utilized the box lines of the table, Gemelli et al. [51] further calculated the distance between cells to derive edge features. It is worth noting that, due to the reliance on lines, these methods face challenges when applied to unlined tabular layouts.

Zhang et al. [196] proposed VSR, which considers the document graph as a fully connected graph and employs self-attention to automatically learn the edges instead of explicitly defining the nodes' relations. This idea addresses the issue of detecting tables without lines. Additionally, several GNN-based models have solved both table detection and table structure analysis tasks [51, 146, 151], and we will discuss them in Section 4.1.

The advantages and disadvantages of GNN-based models are:

- Advantages
  - The presentation of the table suggests that it is well-suited for modeling with a graph, as it inherently incorporates and leverages structured information from the table.
- Disadvantages
  - Node and edge definitions are critical to the performance of graph models, and designing these features is difficult.

### 3.3 Evaluation Metrics

It is necessary to discuss the evaluation metrics before looking into the performance of current research.

(1) `IoU`
**Intersection Over Union (IoU)** [138] is commonly used in the object detection task. It quantifies how much the predicted region overlaps with the actual ground truth region. Given the IoU threshold, a sample is positive if its IoU value is greater than the threshold; otherwise, it is negative. This is how it is defined:

$$\text{IoU} = \frac{\text{Area of Overlap Region}}{\text{Area of Union Region}} \tag{1}$$

Although IoU is widely adopted in natural object detection, it has certain limitations in document object detection. Yu et al. [190] recognized a gap between high IoU and detection entirety in the scientific figure and table detection task. For instance, a low IoU result, which includes more blank backgrounds but retains the entirety of the figure, is preferable to a high IoU detection result that loses critical boundary information, such as an axis label.

(2) `Recall`
Recall [138] is the percentage of correct positive predictions among all given ground truths. TP represents a correct detection of a ground-truth bounding box; while FN denotes an undetected ground-truth bounding box.

$$\text{Recall } = \frac{TP}{TP+FN} \tag{2}$$

(3) `Precision`
Precision [138] is the percentage of correct positive predictions. The formula is as follows. FP represents an incorrect detection of a nonexistent object or a misplaced detection of an existing object.

$$\text{Precision } = \frac{TP}{TP+FP} \tag{3}$$

(4) `F-Measure`
F-Measure [56] is calculated by taking the harmonic mean of Precision and Recall. The formula for F-Measure is:

$$\text{F-Measure } = \frac{2 \times \text{ Recall } \times \text{ Precision}}{\text{Recall } + \text{ Precision}} \tag{4}$$

(5) `AP`
AP [138] is defined as the average detection precision under different recalls. This involves computing the average of precision values derived from the Precision-Recall curve. There are typically two methods to calculate AP: the 11-point interpolation and all-point interpolation. The 11-point interpolation was first adopted by the Pascal VOC 2008 challenge. Precision values, denoted as $p_{interp}(R)$ for recall values distributed at 10 equal intervals ranging from 0 to 1, are averaged to yield the final $AP_{11}$. It is important to note that $p_{interp}(R)$ does not

use the precision at Recall = R on the curve but rather represents the maximum precision when the recall value exceeds R.

$$AP_{11} = \frac{1}{11} \sum_{0,0.1,..,1} p_{interp}(R) \tag{5}$$

The 11-point interpolation has limitations due to precision loss with only 11 sampling points. To address this, the all-point interpolation method was proposed in the Pascal VOC 2010 challenge. This method involves generating a smoothed Precision-Recall curve and calculating the area under the curve through integral operation to calculate AP.

$$AP = \int_0^1 p_{interp}(r)dr \tag{6}$$

(6) mAP

The **mean AP (mAP)** [138] is a metric used to measure the accuracy of object detectors over all classes. The mAP is simply the average AP over all classes and the formula for that is:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{7}$$

where *APi* is the AP in the *i*th class and *N* is the number of classes.

(7) AR

The COCO dataset[7] defined AR as the maximum recall given a fixed number of detections per image, averaged over categories and IoUs. In this benchmark, there is no distinction between AR and mAR (and likewise AP and mAP).

AP and mAP are originally introduced in the VOC 2007 challenge,[8] and after that, the object detecting task typically employs 0.5-IoU-based mAP as an evaluation metric. MS-COCO proposed a new AP calculating method in 2014. Instead of using a fixed IoU threshold, MS-COCO AP averages multiple IoU thresholds ranging from 0.5 to 0.95. This shift in metric directs the model's attention to the accuracy of the bounding box region, which can be significant in some scenarios. More AP variants were summarized in Padilla et al. [138]. There are other metrics proposed by researchers, such as **Localization Recall Precision (LRP)** [137], but the dominant metrics are still IoU-based.

### 3.4 Performance

In this section, we summarize the performances of competitive models on popular benchmark datasets. Table 3 shows the results of different models on table detection datasets. Some researchers do not specify the IoU threshold they set, but they compare their results with others in which the IoU metrics are given. Hence, we can infer that they used the same IoU threshold.

On the Marmot dataset, CDeC-Net [2] achieves the highest Precision of 0.975, while Ajij et al. [4] reported the highest Recall and F1 score of 0.984 and 0.972, respectively, at an IoU setting of 0.5. CasTabDetectoRS [57] wins first place at the IoU setting of 0.9 and second place at the IoU setting of 0.5, with F1 scores of 0.904 and 0.958, respectively.

On the TableBank (LaTeX) dataset, both CasTabDetectoRS [57] and Phan et al. [140] achieved the highest Recall score of 0.984 at an IoU threshold of 0.5, while CDeC-Net [2] did best on Precision and F1 score. When the IoU threshold is set as 0.9, CasTabDetectoRS [57] has a slight advantage over HybridTabNet [135] with a 0.001 higher F1 score.

[7] https://cocodataset.org/

[8] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html

Table 3. Competitive Models' Performances on Table Detection Datasets

| Dataset | Method | IoU | Score | | |
|---|---|---|---|---|---|
| | | | Recall | Precision | F1 |
| Marmot | DeCNT* [160] | 0.5 | 0.946 | 0.849 | 0.895 |
| | **CDeC-Net*** [2] | **0.5** | 0.93 | **0.975** | 0.952 |
| | HybridTabNet* [135] | 0.5 | 0.961 | 0.962 | 0.956 |
| | CasTabDetectoRS*[57] | 0.5 | 0.965 | 0.952 | 0.958 |
| | **Ajij et al.**⋆ [4] | 0.5 | **0.984** | 0.96 | **0.972** |
| | CDeC-Net* [2] | 0.9 | 0.765 | 0.774 | 0.769 |
| | **HybridTabNet*** [135] | 0.9 | **0.903** | 0.900 | 0.901 |
| | **CasTabDetectoRS***[57] | 0.9 | 0.901 | **0.906** | **0.904** |
| TableBank(LaTeX) | U-SSD* [91] | 0.5 | - | - | 0.93 |
| | Ajij et al.⋆ [4] | 0.5 | 0.948 | 0.981 | 0.965 |
| | CascadeTabNet* [143] | 0.5 | 0.972 | 0.959 | 0.966 |
| | Li et al.* [97] | 0.5 | 0.962 | 0.872 | 0.915 |
| | HybridTabNet* [135] | 0.5 | – | – | 0.980 |
| | **CasTabDetectoRS***[57] | 0.5 | **0.984** | 0.983 | 0.984 |
| | **Phan et al.*** [140] | 0.5 | **0.984** | 0.985 | 0.984 |
| | **CDeC-Net*** [2] | 0.5 | 0.979 | **0.995** | **0.987** |
| | HybridTabNet* [135] | 0.9 | - | - | 0.934 |
| | **CasTabDetectoRS***[57] | 0.9 | **0.935** | **0.935** | **0.935** |
| ICDAR17-POD | **HybridTabNet***[135] | 0.6 | **0.997** | 0.882 | 0.936 |
| | CDeC-Net* [2] | 0.6 | 0.931 | 0.977 | 0.954 |
| | CasTabDetectoRS*[57] | 0.6 | 0.941 | 0.972 | 0.956 |
| | DeCNT* [160] | 0.6 | 0.971 | 0.965 | 0.968 |
| | GOD* [153] | 0.6 | – | – | 0.989 |
| | **Huang Y et al.*** [69] | 0.6 | 0.972 | **0.978** | **0.975** |
| | **HybridTabNet***[135] | 0.8 | **0.994** | 0.879 | 0.933 |
| | CDeC-Net* [2] | 0.8 | 0.924 | 0.970 | 0.947 |
| | CasTabDetectoRS∗[57] | 0.8 | 0.932 | 0.962 | 0.947 |
| | DeCNT* [160] | 0.8 | 0.952 | 0.946 | 0.949 |
| | GOD* [153] | 0.8 | – | – | 0.971 |
| | **Huang Y et al.*** [69] | 0.8 | 0.968 | **0.975** | **0.971** |

* denotes CNN-based models while ⋆ represents hybrid models.

On the ICDAR17-POD dataset, the models rank almost equally for different IoU thresholds. HybridTabNet [135] reports the highest Recall at both IoU of 0.6 and 0.8, with 0.997 and 0.994, respectively. Huang et al. [69] ranked first in Precision and F1 score.

From the perspective of the model, the result indicates that the CNN architecture is most commonly used and highly competitive. However, the performance of the same model on different datasets varies widely, and no model can achieve SOTA results on all datasets. Additionally, some models struggle to maintain a balance between Precision and Recall, such as CDeC-Net [2], which on all three datasets reports nearly the highest precision but relatively low recall.

From the perspective of the IoU thresholds, we observe that some models' performances drop drastically at the high IoU threshold. Taking the CDeC-Net [2] model as an example, when the IoU is 0.9, its recall score on the Marmot dataset is 18.1% lower than when the IoU is 0.5. Hashmi et al.

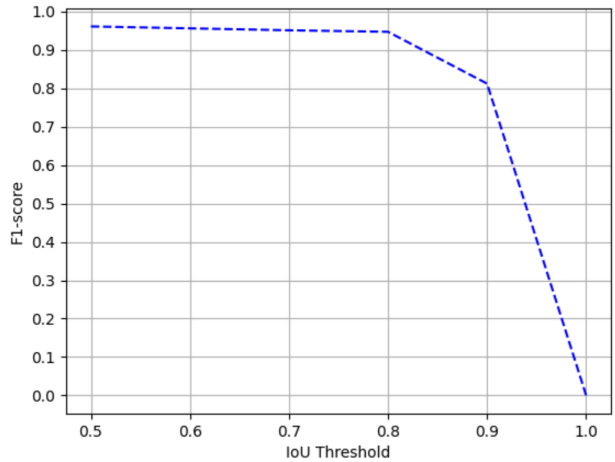



Fig. 5. The F1 score of CasTabDetectoRS [57] over the varying IoU thresholds ranging from 0.5 to 1.0 on the ICDAR17-POD table detection dataset.

Table 4. Competitive Models' Performances on Document Layout Analysis Datasets

| Dataset | Method | mAP@IOU[0.50:0.95] | | |
|---|---|---|---|---|
| | | Table | Figure | Overall |
| PubLayNet | **DocSegTr** [68] | 0.966 | **0.975** | 0.894 |
| | **DiT** [95] | 0.978 | **0.972** | 0.949 |
| | **LayoutLMv3**∗ [68] | 0.979 | 0.970 | **0.951** |

[57] visualized the F1 score of CasTabDetectoRS over the varying IoU thresholds ranging from 0.5 to 1.0 on the ICDAR17-POD dataset, as Figure 5 shows. Figure 5 indicates that when the IoU exceeds 0.9, the F1 score declines sharply. Moreover, we discover that the rankings of the models do not change under different IoU thresholds in each dataset.

Table 4 displays three document layout models' performances on the PubLayNet dataset, which are all Transformer-based. The PubLayNet dataset classifies page objects into five categories, and we focus on the table and figure results in this study. As can be seen from the table, the results of DiT [95], and LayoutLMv3 [68] are very close. The lower overall score of DocSegTr [15] is due to its lower accuracy in predicting titles and texts.

### 3.5 Observations

The table and figure detection task serves as the foundation for subsequent analysis and downstream tasks, and the detection quality significantly influences follow-up research. Despite the excellent performances of current research, there are still some areas for improvement.

— Available datasets are mainly in English, and research on detecting scientific tables/figures in other languages is scarce. Nevertheless, the papers' disciplines are primarily computer science and biomedical, and an accurately annotated dataset that spans multiple disciplines and languages is still lacking.
— Most detection models only use document images as input, limiting their ability to fully leverage the valuable information embedded within PDF stream content. Consequently, it

Table 5. Available Datasets for Scientific Table Structure Analysis

| Dataset | Source | Format | Tables | Year | CT | CC | CL | Link |
|---|---|---|---|---|---|---|---|---|
| UW3 [141] | Books | PNG, XML | 147 | 1996 | ✓ | × | ✓ | Link |
| TableBank [96] | arXiv | LaTeX, PNG | 145k | 2019 | ✓ | × | × | Link |
| SciTSR [26] | arXiv | PNG, JSON | 1.5k | 2019 | ✓ | ✓ | × | Link |
| TABLE2LATEX-450K△ [41] | arXiv | PNG, JSON | 450k | 2019 | ✓ | ✓ | × | Link |
| TabStructDB△ [159] | CiteSeer | XML, PNG | 1k | 2019 | × | × | × | Link |
| ICDAR2019 [50] | Websites | XML, PNG | 2.3k | 2019 | ✓ | × | ✓ | Link |
| DECO [84] | Enron corpus | Excel | 854 | 2019 | ✓ | ✓ | ✓ | Link |
| PubTabNet [201] | PubMed | PNG, JSON | 568k | 2020 | ✓ | ✓ | ✓* | Link |
| TabLeX [42] | arXiv | LaTeX, PNG | 4M | 2021 | ✓ | ✓ | × | Link |
| PubTables-1M [165] | PubMed | PNG, JSON | 1M | 2021 | ✓ | ✓ | ✓ | Link |
| FinTabNet [199] | Company reports | PDF, JSON | 110k | 2021 | ✓ | ✓ | ✓ | Link |
| WTW [111] | Multiple wild scenarios | JPG, XML | 16k | 2021 | ✓ | × | ✓ | Link |
| WikiTableSet [118] | Wikipedia | PNG, JSON | 5.23M | 2023 | ✓ | ✓ | ✓ | Link |
| TabRecSet [183] | Multiple wild scenarios | JPG, JSON | 38.1K | 2023 | ✓ | ✓ | ✓ | Link |

**CT** denotes cell topology, **CC** is cell content whereas **CL** is cell location, and * represents datasets that cell bounding boxes are only provided for non-blank cells. Unaccessible datasets are denoted with △.

is crucial to develop a model that can exploit information from PDF stream content and function optimally when document images are the only available input.

— Existing studies still struggle with dense tables and atypical table layouts, such as tables with only a few rows. Also, content that resembles a table, such as a graph with grids, aligned formulas, directories, and so on, may be misjudged as a table.
— Current research regarding model robustness, generalization, complexity, and inference efficiency in the context of table and figure detection tasks still has significant room for advancement.
— The entirety and completeness of scientific table and figure detection are fundamental to downstream tasks, yet there are few relevant evaluating research studies or techniques.

## 4 STRUCTURE ANALYSIS FOR TABLES AND FIGURES

This section presents research on structure analysis of scientific tables and figures. While the primary goal of the table structure analysis is to identify the roles and relations of cells, the figure structure analysis focuses on extracting figure components and the relations between them. Due to the distinct components of tables and figures, the associated tasks exhibit notable differences. Therefore, this section separately presents the datasets, evaluation metrics, and research progress for these two problems.

### 4.1 Table Structure Analysis

According to Hashmi et al. [56], there are two tasks related to table structure analysis: table structure recognition and table recognition. The former identifies the table structure solely, while the latter extracts the table content. Given the considerable similarity between these two tasks, we categorize related studies based on the research method rather than the specific task. Next, we will introduce the datasets, evaluation metrics, and performances, respectively.

*4.1.1 Datasets.* Table 5 summarizes datasets commonly employed in table structure analysis. In addition to datasets derived from the academic literature, we incorporate datasets obtained from diverse real-world scenarios to provide readers with a comprehensive summary. Subsequently, detailed information regarding datasets built upon academic literature is presented, excluding TableBank [96], SciTSR [26], and PubTables-1M [165], which are introduced in Section 3.1.

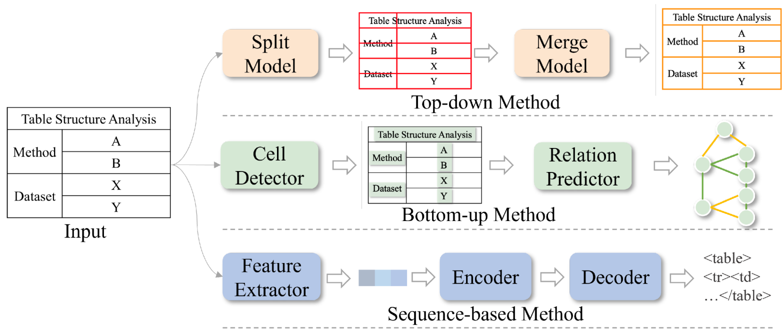



Fig. 6. Three types of existing methods for table structure recognition.

**PubTabNet.** PubTabNet [201] is a publicly available table recognition dataset with 568k table images and structured HTML representations. It is generated automatically by comparing XML and PDF representations of scientific articles from the PubMed dataset. The authors created a balanced test set by randomly choosing 5,000 tables with spanning cells and the same amount of tables without spanning cells. PubTabNet is employed as the competition dataset of the ICDAR 2021 Scientific Literature Analysis Competition Task B - Table Recognition [83].

**TabLeX.** To our knowledge, TabLeX is the largest dataset derived from scientific papers for the table recognition task. It comprises table images and corresponding LaTeX sources from arXiv papers and is divided into two subsets for table structure extraction and table content extraction, respectively. Notably, the authors augment the LaTeX codes with 12 distinct font styles and subsequently render them into table images with ratio variations. Distinguishing itself from other datasets that predominantly focus on biomedical and computer science papers, TabLeX incorporates a substantial number of papers on physics and mathematics.

*4.1.2 Methods.* Initially, research on table structure analysis relies on heuristic rules. For instance, Namysł et al. [133] introduced the heuristic-based method and design rules for fully bordered tables and for partially bordered or borderless tables, respectively. In recent years, deep learning models have become the most popular methods for table recognition, categorized into three main types: *top-down models*, *bottom-up models*, and *sequence-based models*, as illustrated in Figure 6. *Top-down* methods typically predict table splitting lines first and then merge over-split cells. On the other hand, *bottom-up* methods detect cells first and subsequently predict cell relations. These two methods can be considered two-stage frameworks, while *sequence-based* methods are end-to-end, directly outputting HTML or LaTeX codes to represent table structure. We present the distribution of these techniques from 2017 to 2023 in Figure 7.

*Top-down models.* The core idea of top-down models is to divide the table image into row and column grids using detection or segmentation models and then locate cells by intersecting rows and columns. DeepDeSRT [156] was the initial approach that employs a semantic segmentation model for table structure recognition. However, DeepDeSRT encounters challenges when confronted with tables containing multi-rows or multi-columns, as it primarily relies on local information. Similarly, DeepTabStR [159] also faces limitations as it is unable to recognize cells that span across rows or columns.

SPLERGE [170] addresses this issue by proposing two models: the Split model and the Merge model. The Split model predicts row and column separators, and the Merge model extracts cells by row and column intersection and predicts which cells should be merged to reconstruct multi-rows

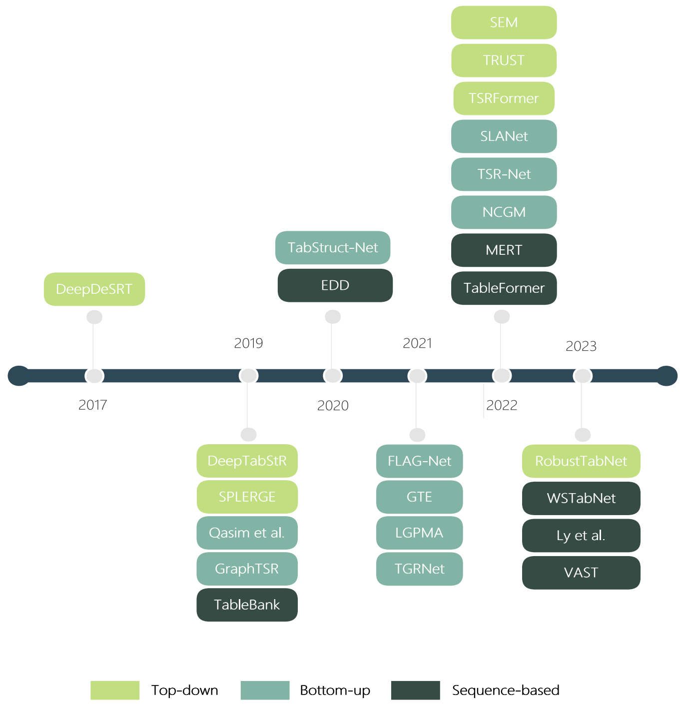



Fig. 7. The distribution of table structure analysis methods from 2017 to 2023.

or multi-columns. The limitation is that the two models are trained separately, which may make optimization more complex than end-to-end training.

In contrast to SPLERGE, **SEM (Split, Embed, and Merge)** proposed by Zhang et al. [198] considers the text information of cells, thus achieving higher accuracy. The Embedder extracts the grid-level visual and textual features from BERT and RoIAlign and fuses them for the Merger. The Merger is a GRU decoder that predicts the grid-merged results step-by-step based on the fused features provided by the Embedder. A notable limitation of SPLERGE is that the two models undergo separate training, potentially introducing complexity to the optimization process compared to end-to-end training.

TRUST adopts an end-to-end Transformer-based framework, including a CNN backbone as the visual feature encoder, a Query-Based Splitting Module for row and column splitting lines generation, and a Vertex-based Merging Module for cell relation prediction. It demonstrates outstanding results on various complex tables, including rotating and unlined tables, and those with spanning or empty cells. Similarly, RobusTabNet [120] reported promising results on recognizing tables with large empty cells and distorted regions. This is due to the novel spatial CNN-based separation line prediction module, which effectively propagates contextual information across the whole table image.

Unlike the mentioned methods that first predict table splitting lines and then merge over-split cells, GridFormer [119] directly explores predicting vertices and edges, demonstrating satisfactory results on complicated tables.

Top-down models share both common advantages and disadvantages, including:

— Advantages
  - Top-down methods have good generalization because they aim at predicting the basic table grid pattern, which is similar across different kinds of tables.

  - By intersecting row and column separators, top-down methods generate more accurate cell bounding boxes.
- Disadvantages
  - Top-down methods assume axis-aligned tables, so they may fail when processing distorted or rotating tables. However, this rarely occurs in academic paper data unless it is a photo of the paper or a scanned image.
  - Top-down models usually predict table grids first and then merge some cells that belong to the same spanning cell. However, this two-stage strategy may lead to error propagation, where incorrect predictions in the table grid may result in entirely inaccurate outcomes.

*Bottom-up models.* Bottom-up models consider texts or cells as table elements and leverage GNNs or LSTM networks to learn cell relations. Relevant studies can further be categorized according to table elements into text-based and cell-based. Text-based methods detect text bounding boxes and treat texts as nodes of a table graph. The acquisition of table content mainly depends on PDF stream content and OCR results, which may introduce errors and ignore empty cells. On the other hand, cell-based methods focus on detecting cell bounding boxes. Its advantage is that once accurate cell bounding boxes are obtained, the table structure can be easily inferred due to the alignment properties of cells. Next, we provide an overview of each of these two categories.

Several research studies regard text blocks as nodes and construct table graphs. Qasim et al. [146] first applied GNN to table structure recognition. They extracted cell contents by employing OCR techniques and treated every content as a node within the table graph. A limitation of this work is that it cannot recognize any spanning cells. On the other hand, GraphTSR [26] can recognize spanning cells; however, it exclusively detects K-nearest neighbors when predicting cell relationships, which may not comprehensively represent the entire table structure. Additionally, GraphTSR requires cell content coordinates as input during both training and inference. In contrast, Liu et al. [105] proposed an end-to-end FLAG-Net without the need for OCR techniques and extra metadata. More specifically, FLAG-NET employs an object detection model to detect text blocks as table elements, which are then fed into novel **FLexible context AGgregator (FLAG)** modules to predict relationships from the cell, row, and column perspectives. NCGM [103] detects text segments as table elements and further leverages multi-modal features from content, appearance, and geometry perspectives.

Additionally, researchers also investigate considering cells as table elements. GTE-Cell [200] trains an attribute network to classify the presence of graphical ruling lines in a table and subsequently employs a corresponding cell detection network. Following cell detection, GTE-Cell merges cells using a specific rule: when the content of a cell begins with a lowercase character, it is merged with the cell above it, which begins with a capital character. TGRNet [182] adopts an instance segmentation model to detect cells, and simultaneously predict cell logical relations by formulating it as a node classification task. Similarly, TabStruct-Net [149] employs Mask R-CNN to detect cell bounding boxes and learns table structure using graphs. More specifically, it leverages the DGCNN architecture [145] to model the interaction between geometrically neighboring detected cells. One of the limitations of TabStruct-Net is that it cannot deal with tables containing a large amount of empty cells. In addition, Qiao et al. [147] developed an approach that can accurately detect cell bounding boxes and capture empty cells. The authors first generated the aligned bounding box annotations according to the maximum box height/width in each row/column. Then, they refined the detected aligned bounding boxes using the **local branch (LPMA)** and the **global branch (GPMA)**. Through visible texture perceptron, the LPMA learns more reliable text region information, whereas the GPMA learns the global information of cell range.

The advantages and disadvantages of bottom-up models are:

— Advantages
  - Compared to sequence-based models that utilize markup language to represent table structure while ignoring cell locations, bottom-up models explicitly detect cell bounding boxes. This approach is easier for humans to interpret and correct, resulting in better performance.

— Disadvantages
  - Bottom-up methods always suffer from the "cell boundary ambiguity" problem and may report poor results on tables containing multiple empty cells.
  - GNN is a prevalent architecture in bottom-up models for predicting cell relations. However, it is inefficient due to the more expensive training cost, e.g., training time and data volume.

*Sequence-based models.* Sequence-based models typically take a table image as the input of the encoder, and the decoder outputs a sequence of markup tags that indicate the table structure.

TableBank [96] provides a baseline model for table structure recognition based on the image-to-markup model [40]. Additionally, He et al. [60] employed MASTER [114], which consists of a multi-aspect global context attention-based encoder module and a transformer-based decoder module to generate LaTeX code for table images. Zhong et al. [201] proposed an **encoder-dual decoder (EDD)** architecture that reconstructs whole tables, including table content. In this work, the structure decoder generates HTML code to reproduce the table structure, whereas the cell decoder recognizes cell content. Ly et al. [118] employed a similar architecture to EDD, containing a structure decoder for generating table structure and a cell decoder to predict cell contents. Moreover, they constructed WikiTableSet, the largest publicly available table recognition dataset in three languages derived from Wikipedia. This initiative addresses the limitations of existing datasets, which predominantly focus on English tables. In another work, Ly et al. additionally introduced a local attention mechanism within decoders, which demonstrates effectiveness in big tables. Instead of decoding text content from images, TableFormer [134] predicts the bounding box of table cells and extracts content from PDFs. VAST [67] follows a similar structure, leveraging a coordinate sequence decoder for cell bounding box prediction. Additionally, a visual-alignment loss is introduced to generate more accurate bounding boxes.

Overall, sequence-based models have both advantages and disadvantages, which can be summarized below:

— Advantages
  - The computational cost of sequence prediction is much lower than relation prediction based on GNN. Sequence-based models demonstrate notable efficiency and fast computational speed.
  - Compared to two-stage models that employ either bottom-up or top-down strategies, sequenced-based models may exhibit reduced intermediate losses due to their end-to-end training process.

— Disadvantages
  - Sequence-based models usually do not generate explicit cell bounding boxes, which makes the results less interpretable and makes recovering from recognition errors or resolving ambiguities in cell recognition difficult.
  - This type of model relies heavily on large-scale end-to-end training, and the performance will degrade sharply in unseen data.
  - These methods worked well for simple tables but were not robust enough for dense and complex tables.

#### 4.1.3 *Evaluation Metrics.*

(1) `TEDS`
**Tree edit distance-based similarity (TEDS)** [201] regards the table structure as a tree structure and utilizes the tree distance to compare the similarity of two trees. This is how it is defined:

$$\mathrm{TEDS}\left(T_a, T_b\right) = 1 - \frac{\mathrm{EditDist}\left(T_a, T_b\right)}{\max\left(\left|T_a\right|, \left|T_b\right|\right)} \tag{8}$$

where T denotes the tree, EditDist represents the tree's editing distance, and T is the number of nodes in T. TEDS was first proposed along with the PubTabNet [201] dataset.

(2) `TEDS-Struct`
TEDS-Struct was proposed by Qiao et al. [147] and was modified from TEDS [201]. It ignores OCR errors and only focuses on table structure. The authors claimed that the performance difference between TEDS-Struct and TEDS is primarily due to recognition errors and annotation ambiguities.

(3) `BLEU`
**Bilingual Evaluation Understudy (BLEU)** [139] is an evaluation metric initially used for machine translation. The BLEU score is calculated by comparing the predicted text to the ground truth text. The BLEU measure assigns a number from 0 to 1, with 1 being the best result for the predicted text. The Table2LaTeX and TableBank datasets leverage the BLEU metric for evaluation.

(4) `Precision, Recall, and F1 score`
This metric was first proposed by the ICDAR 2013 table competition [53] and SciTSR employs it as well. The basic concept is to convert the table into a list of cell adjacencies and then use accuracy and recall measures to compare the predicted table to the ground true table. These scores are calculated separately for each table; the final result is macro and micro average scores.

*4.1.4 Performance.* Most studies compare their performance on PubTabNet, TableBank, and SciTSR, three popular table structure recognition datasets built from scientific documents. We summarize existing methods' results on these datasets, as shown in Table 6. As the table shows, the top two models within each model type exhibit commendable performance. For instance, in top-down methods, RobustTabNet [120] and TSRFormer [100] achieve the highest scores in the SciTSR-COMP dataset. Additionally, TSRFormer [100] achieves competitive results in both PubTabNet and SciTSR, closely approaching the state-of-the-art methods in these two datasets. FLAG-Net [105] represents the effectiveness of bottom-up models, reporting the highest precision and recall scores in the SciTSR and SciTSR-COMP datasets, respectively. The majority of sequence-based models are evaluated in the PubTabNet dataset, where the top two methods proposed by Ly and Takasu [117] demonstrate superior performance, outperforming other method types. Moreover, NCGM [104] proposes a novel neural collaborative graph machine, falling outside our predefined categories, and emerges as the winner in the TableBank and SciTSR benchmarks. Given the limited evaluation of models across all datasets and the absence of unified evaluation metrics, it is difficult to determine which type of method is best.

In addition, the IBM company and the IEEE ICDAR 2021 jointly organized the ICDAR 2021 Competition on Scientific Literature Parsing, Task-B, which employed PubTabNet as the competition dataset and aimed to drive the advances in scientific table recognition. The organizers further categorized the overall results (TEDS all) into simple and complex tables, as presented in Table 7. Notably, all models exhibit three to four percentage points lower scores on complex tables

Table 6. Competitive Models' Performances on PubTabNet, TableBank, and SciTSR Datasets

| Type | Method | PubTabNet | | Table Bank | SciTSR | | | SciTSR-COMP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TEDS | TEDS-Struct | BLEU | Precision | Recall | F1 | Precision | Recall | F1 |
| Top-down | DeepDeSRT [156] | – | – | – | 0.906 | 0.887 | 0.89 | 0.811 | 0.813 | 0.812 |
| | SPLERGE [170] | – | – | – | 0.922 | 0.915 | 0.918 | | | |
| | SEM [198] | – | – | – | **0.997** | 0.965 | 0.971 | 0.968 | 0.947 | 0.957 |
| | TRUST [54] | **0.962** | – | – | – | – | – | – | – | – |
| | RobustTabNet [120] | – | 0.97 | – | 0.994 | 0.991 | 0.993 | **0.99** | 0.984 | 0.987 |
| | TSRFormer [100] | – | **0.975** | – | 0.995 | **0.994** | **0.994** | 0.991 | **0.987** | **0.989** |
| Bottom-up | T2 [94] | – | – | – | 0.993 | 0.99 | 0.992 | 0.976 | 0.961 | 0.969 |
| | TabStruct-Net [149] | – | 0.901 | 0.916 | 0.927 | 0.913 | 0.92 | 0.909 | 0.882 | 0.895 |
| | GraphTSR | – | – | – | 0.959 | 0.948 | 0.953 | 0.964 | 0.945 | 0.955 |
| | GTE [200] | – | 0.93 | – | – | – | – | – | – | – |
| | TSR-Net [99] | – | 0.9564 | – | – | – | – | – | – | – |
| | LGPMA [147] | 0.946 | 0.967 | – | 0.982 | 0.993 | 0.988 | 0.973 | **0.987** | 0.98 |
| | FLAG-Net [105] | 0.951 | – | **0.939** | **0.997** | **0.993** | **0.995** | **0.984** | 0.986 | **0.985** |
| | SLANet [93] | **0.9589** | **0.9701** | – | – | – | – | – | – | – |
| Sequence-based | EDD [201] | 0.883 | – | | | | | | | |
| | MERT [175] | 0.9234 | 0.9571 | – | – | – | – | – | – | – |
| | TableFormer [134] | 0.936 | 0.9675 | – | – | – | – | – | – | – |
| | VAST [67] | 0.9631 | 0.9723 | – | – | – | – | – | – | – |
| | WSTabNet [118] | 0.9648 | **0.9774** | – | – | – | – | – | – | – |
| | Ly et al. [117] | **0.9677** | – | – | – | – | – | – | – | – |
| Others | NCGM [103] | 0.954 | – | **0.946** | **0.997** | **0.996** | **0.996** | – | – | – |
| | GridFormer [119] | 0.9584 | 0.97 | – | 0.9936 | 0.9904 | 0.992 | – | – | – |

The best model for each type of method is shown in **bold** font, and colored cells represent the best score for that dataset.

Table 7. ICDAR 2021 Competition on Scientific Literature Parsing, Task-B Results

| Team Name | TEDS Simple | TEDS Complex | TEDS all |
|---|---|---|---|
| **Davar-Lab-OCR**⋇ | 0.9788 | 0.9478 | **0.9636** |
| **VCGroup**⋇ | **0.9790** | 0.9468 | 0.9632 |
| **USTC-NELSLIP(SEM)**★ | 0.9760 | **0.9489** | 0.9627 |
| Ly et al.★ [117] | 0.9777 | 0.9458 | 0.9621 |
| YG | 0.9738 | 0.9479 | 0.9611 |
| WSTabNet* | 0.9751 | 0.9437 | 0.9597 |
| DBJ | 0.9739 | 0.9387 | 0.9566 |
| TAL | 0.9730 | 0.9393 | 0.9565 |
| PaodingAI⋇ | 0.9735 | 0.9379 | 0.9561 |
| anyone | 0.9695 | 0.9343 | 0.9523 |
| LTIAYN | 0.9718 | 0.9240 | 0.9484 |

⋇: bottom-up models, ★: top-down models, *: sequence-based models.

compared to simple tables, indicating a considerable scope for improvement in the model's robustness to handle complex table structures.

Inference efficiency has received attention from academics recently. However, there are no uniform evaluation metrics. Table 8 shows the inference efficiency experiments conducted by Liu et al. [105] and Guo et al. [54]. In Table 8, the units are million (M) for #Param, second(s) for GPU time, and second(s) for CPU time. The execution time is computed on one Nvidia Tesla V100 GPU and a 2.4 GHz Intel Xeon E5 CPU. We observe that TabStruct-Net [149] takes much longer to infer than FLAG-Net [105] because the former greedily exploits a large number of proposals. In contrast, the latter introduces a proposal filtering mechanism to avoid this. The reason for the inefficiency of

Table 8. The Inference Efficiency of Different Models [54, 105]

| Method | #Param | GPU | CPU | FPS |
|---|---|---|---|---|
| SPLERGE★ [170] | 0.37 | 0.95 | 24.25 | – |
| TabStruct-Net⋇ [149] | 68.63 | 22.63 | 76.52 | 0.77 |
| FLAG-Net⋇ [105] | 17 | 0.13 | 2.37 | – |
| EDD∗ [201] | – | – | – | 1 |
| SEM★ [198] | – | – | – | 1.94 |
| TRUST★ [54] | – | – | – | 10 |

Table 9. Available Datasets for Scientific Figure Structure Analysis

| Dataset | Source | Size | Cate. | Figure Type | Annotation | | | | | Year | Link |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Bbox | Caption | Text | CL | CR | | |
| FigureSeer [161] | CS Conferences | 999 | 7 | 🟠 | × | × | ✓ | ✓ | ✓ | 2016 | Link |
| Viziometrics [89] | PubMed | 2,881,372 | 5 | 🟠🟢🔵 | × | ✓ | × | × | × | 2016 | Link |
| ACA [142] | ACL repository | 332 | 5 | 🟠 | × | × | ✓ | ✓ | ✓ | 2017 | Link |
| ChartSense [74] | Google search | 5659 | 10 | 🟠 | × | × | × | × | × | 2017 | – |
| FigureQA [77] | Synthetic | 100,000 | 5 | 🟠 | × | × | × | ✓ | ✓ | 2018 | Link |
| DVQA [75] | Synthetic | 300,000 | 1 | 🟠 | × | × | × | ✓ | ✓ | 2018 | Link |
| MV Dataset [23] | IEEE conferences | 360 | 14 | 🟠 | ✓ | × | × | × | × | 2020 | Link |
| ICDAR2019 [34] | Synthetic | 202,550 | 10 | 🟠 | × | ✓ | ✓ | ✓ | ✓ | 2019 | Link |
| ICDAR2019 [34] | PubMed | 4,242 | 10 | 🟠 | × | ✓ | ✓ | ✓ | ✓ | 2019 | Link |
| PlotQA [126] | Synthetic | 224,377 | 3 | 🟠 | × | × | × | ✓ | ✓ | 2020 | Link |
| LEAF-QA [20] | Synthetic | 250,000 | 4 | 🟠 | × | × | ✓ | ✓ | ✓ | 2020 | – |
| VIS30K [21] | IEEE conferences | 30,000 | 4 | 🟢🔵 | × | × | × | × | × | 2021 | Link |
| VisImages [39] | IEEE conferences | 12,267 | 34 | 🟠🟢🔵 | ✓ | ✓ | ✓ | ✓ | ✓ | 2022 | Link |
| ChartQA [122] | Websites | 21,945 | 3 | 🟠 | × | ✓ | ✓ | ✓ | ✓ | 2022 | Link |
| MapQA [19] | Synthetic | 62,367 | 3 | 🔵 | × | ✓ | ✓ | × | × | 2022 | Link |
| ACL-Fig [80] | ACL repository | 112,052 | 19 | 🟠🟢🔵 | ✓ | ✓ | ✓ | × | × | 2023 | Link |
| GenPlot [8] | Synthetic | 500,000 | 5 | 🟠 | × | ✓ | ✓ | × | × | 2023 | Link |

The "Size" column represents the number of figures in the dataset. In the "Figure Type" column, 🟠 denotes chart, 🟢 is diagram and 🔵 is image.In the "Annotation" column, **bbox** is the bounding box of figure in document image, **CL** denotes component location while **CR** is component role.

EDD [201] may be that it uses LSTM, which cannot be computed in parallel as a cell decoder to generate HTML representations. The time-consuming part of SEM is the embedder, which contains **Region of Interest (RoI)** operations and context features extraction via BERT.

### 4.2 Figure Structure Analysis

In this survey, we categorize figures into three subtypes: charts, diagrams, and images, as outlined in Section 2. Since figures of different categories can vary widely, there is no unified structure analysis task similar to tables. Existing research on structure analysis primarily focuses on charts. For instance, Singh and Shekhar [164] argued that the main distinctions between statistical charts and natural images lie in the structure and set of chart elements. In this context, chart structure encompasses the types, positions, colors, and patterns of chart elements. Mishra et al. [129] extracted eight kinds of chart elements, including title, X-axis, X-axis values, Y-axis, Y-axis values, legend title, legend label, and data label. The authors constructed relationship graphs between chart elements, which could be another representation of chart structure. Additionally, as subfigures can be considered elements of a compound figure, we also encompass relevant work on subfigure separation within the scope of figure structure analysis. This subsection provides an overview of the literature on chart element extraction, subfigure separation, and relevant datasets.

*4.2.1 Datasets.* We outline several datasets which can be used for figure structure analysis from the perspectives of source, format, quantity, category, label type, and so on, as illustrated in Table 9. Some of these datasets are built from academic papers, while others are from Google or are synthetic statistical charts. Below we present some popular datasets built from scientific documents in detail.

**Viziometrics.** Lee et al. [89] collected the article files from the PMC FTP server and extracted the images into a figure corpus. Approximately 66% of these files have associated figure files. After some filtering steps, the authors classified 4.8 million images into five categories: equation, diagram, photo, plot, and table. Furthermore, there are multiple compound figures, and the authors use a customized approach to dismantle these compound figures.

**ACA.** The ACA dataset was proposed by Poco and Heer [142] and is composed of chart images extracted from scientific documents. The authors collected papers from the ACL Anthology repository and extracted figures using the pdffigures tool. This dataset contains 332 images divided into four categories: area charts, bar charts, line charts, and scatter plots. For each image, the authors annotated the position and role of the text in the image.

**MV Dataset.** MV Dataset [23] contains 360 images of multiple-view visualizations collected from the IEEE VIS, EuroVis, and PacificVis publications from 2011 to 2019. Annotators labeled these visualization images with fine-grained annotations of view types and layouts. Images were drawn from 1,976 publications, including 1,149 from IEEE VIS, 475 from EuroVis, and 352 from IEEE PacificVis.

**ICDAR 2019 CHART-Infographics.** The ICDAR 2019 CHART-Infographics [34] was the first competition on harvesting raw tables from infographics. This competition provided two datasets constructed from synthetic charts and scientific literature, respectively. The synthetic chart dataset is curated using data tables obtained from various online sources and encompasses 10 types of charts. The scientific chart dataset is built from the PubMedCentral Open Access repository and contains annotations including chart type, orientation, text location and role, axis, and so on.

**VIS30K.** VIS30k [21] comprises 29,689 images representing 30 years of figures and tables from each track of the IEEE Visualization conference series (Vis, SciVis, InfoVis, and VAST). Compared with other datasets, VIS30k contains a large number of diagrams and images. However, it does not provide fine-grained annotation results for each figure.

**VisImages.** VisImages [39] contains 12,267 images with 12,057 textual captions extracted from 1,397 VAST and InfoVis papers published between 1996 and 2018. The image components were divided into 13 categories by the authors, which included area, bar, circle, point, statistics, text, and so on. VisImages provides component-level information, such as the position and category of image components.

*4.2.2 Chart Elements Extraction.* Scientific charts encompass various types, such as bar charts, pie charts, line charts, and so forth. Several studies focus on element extraction tailored to a specific chart type. Cliche et al. [32] presented a system for extracting the numerical values of data points from scatter plots that depend on an OCR technique and regression model. VIEW, proposed by Gao et al. [49], provides category-specific solutions to extract the underlying data from bar charts, pie charts, and line graphs, and generates a data table for each chart. Nair et al. [132] explored extracting data from line plots, first extracting a dense set of points from a line plot, then representing the entire line plot as a sequence of trends, and finally implementing a Bayesian network for reasoning about the messages conveyed by the line plots and their trends.

There are a few works developing systems that could extract data across a variety of chart types. ChartSense [74] employs semi-automatic, interactive extraction algorithms optimized for each of the ten chart types. Poco and Heer [142] designed a suitable pipeline for different kinds of charts. First, the authors used OCR to get texts and their bounding boxes, then built an SVM classifier to determine the text element role based on their geometric features. The limitation is that it could not classify the non-text chart elements and relied on postprocessing to improve performance. REDEC [129] proposed a CNN-LSTM model to extract structural data from different kinds of charts. To further advance the field of chart recognition and understanding, ICDAR [34] and ICPR [36, 37] organized several competitions on harvesting raw tables from infographics. In these competitions, automatic chart recognition is divided into multiple tasks, including text role classification, plot element detection, and so on, overlapping in scope with the aforementioned studies.

*4.2.3 Subfigure Separation.* Scientific articles typically contain compound figures, which consist of several subfigures, researchers have investigated separating them into individual figures. Cheng et al. [25] utilized a hybrid clustering algorithm and decision tree to segment subfigure image panels automatically. Tsutsui and Crandall [173] trained a CNN model to separate compound figures in scientific documents using transfer learning and automatic synthesis training exemplars to overcome the lack of labeled data. Taschwer and Marques [169] proposed a two-stage system to detect compound figures and separate them. If interested, there are more works on separating composite graphs of the biomedical literature [88, 92, 158, 169].

### 4.3 Observation

We discover that the distribution of academic interests between table structure analysis and figure structure analysis is unbalanced. There are established methodologies, techniques, and datasets for table structure analysis. However, few research studies and datasets are available for figure structure analysis. Current figure structure analysis systems often depend on handcraft operation and complicated preprocessing or postprocessing techniques, which are labor-intensive and only effective for a specific type of figure or table. Although previous studies have shown promising results on multiple datasets, there are still some unresolved issues in table structure analysis, including inconsistency in table size and density, variation in table cell shapes and sizes, tables containing images or formulas, tables without separation lines, and tables with multiple empty cells or spanning cells. The efficiency of the model has also attracted attention. Several studies compare the inference time of the proposed model with previous studies and introduce some techniques to improve inference efficiency.

## 5 FIGURE AND TABLE INTERPRETATION

Interpreting figures and tables involves extracting meaningful information and understanding the semantics embedded within these visual elements. To achieve this goal, the intuitive way is to extract information from tables and charts in a structural way. We summarize the related research as *information extraction*. Additionally, figures, such as diagrams and images, lack data points as structured as charts and tables, leading researchers to explore *summary generation*. Furthermore, with the rapid development of **Large Language Models (LLMs)** and **Large Visual-Language Models (LVLMs)**, an increasing number of researchers are employing them for understanding figures and tables, and addressing various tasks within a single model. We categorize this kind of research as *visual-language reasoning*.

### 5.1 Information Extraction

*5.1.1 Table Information Extraction.* There have been several investigations into extracting table information from PDF files, which is the most common format for scientific papers. In 2005, Yildiz et al. [187] presented pdf2table, a heuristics-based system that recognizes and decomposes tables in PDF files and stores the extracted data in XML format. Milosevic et al. [127] explored extracting useful information from tables in the biomedical literature by template. More recently, researchers have increasingly turned to leveraging deep learning techniques for this task and constructing large datasets for model training. Desai et al. [42] proposed TabLeX, a benchmark for extracting structure and content information from scientific tables, encompassing the fields of physics, computer science, and mathematics. Several methods discussed in Section 4.1 extract table structure and table content simultaneously. For instance, EDD [201] takes table images as input and outputs structural table information in HTML format.

*5.1.2 Chart Information Extraction.* Charts serve as a visual representation of data tables, and the extraction of data from charts is crucial for comprehending chart semantics. In 2011, Mishchenko and Vassilieva [128] introduced an unsupervised model to extract numerical data from five types of charts and represented them in XML format. Al-Zaidy and Giles [6] leveraged image processing and text recognition techniques combined with various rules derived from chart properties to extract data values from bar charts. Chart Decoder [33], utilizing deep learning, computer vision, and text recognition techniques, takes a bar chart image as input and produces textual and numeric information as output. LineFormer [87] employs an instance segmentation model to extract data from line charts. In addition, ChartOCR [116] integrates deep learning techniques and rule-based methods to extract data from various types of charts. With the increasing interest of the academic community in chart data extraction, a range of related competitions and datasets has emerged. For instance, ICDAR [34] and ICPR [36, 37] have organized the CHART-Infographics competition over several years, focusing on harvesting raw tables from infographics. In the ICDAR 2023 competition,[9] CHART-Infographics introduced a new task centered around chart visual question answering, aiming at deepening the understanding of charts.

### 5.2 Summary Generation

According to Bhatia and Mitra [12], generating summaries for figures and tables helps users better understand retrieval results, hence improving search performances. Consequently, summary generation stands out as a viable approach for conveying the semantics of figures and tables. Abstractive summarization and extractive summarization represent the primary branches of current research in this domain.

Abstractive summarization has long been a question of great interest in automatic summarization. Carberry et al. [17] employed a Bayesian belief network to hypothesize the figure designer's intended message. Agarwal and Yu [3] proposed FigSum, which generates a structured text summary for each figure in an article that includes one sentence from each of the four rhetorical categories: **Introduction, Methods, Results, and Discussion (IMRaD)**. Saini et al. [154] proposed a novel unsupervised approach (FigSum++) for automatic figure summarization in biomedical scientific articles using a multi-objective evolutionary algorithm. Zhang et al. [197] built a new conversation-oriented, open-domain table summarization dataset. They experimented with three neural natural language generation models (CopyNet, CPT-2, and Text-to-Text Transfer Transformer) to generate summaries based on tables. Several researchers investigated how to extract text associated with figures in a document. Yu [191] assumed that abstract sentences might summarize figures in a full-text article. They invited the corresponding authors of several articles to identify abstract sentences that summarize the figure content in that article. They utilized the responses to build a corpus, which they then used to evaluate the NLP methodologies they proposed. Similarly, [16] was interested in the associations between figures and abstract sentences. They also

[9] https://chartinfo.github.io/index.html

Table 10. Large Vision-Language Models for Figure and Table Understanding

| Method | Figure/Table | Backbone | FT | VA | Task | Scope | Link |
|---|---|---|---|---|---|---|---|
| Chat2Vis [121] | chart | ChatGPT, etc. | × | × | chart generation | – | Link |
| ChartAssisstant [125] | chart | Sphinx, Donut | ✓ | ✓ | QA, etc. | – | Link |
| FinVis-GPT [178] | chart | LLaVA | ✓ | ✓ | QA, etc. | Financial | Link |
| ChartGPT [171] | chart | Flan-T5 | ✓ | × | chart generation | – | Link |
| MMCA [101] | chart | mPLUG-Owl | ✓ | ✓ | reasoning, etc. | – | Link |
| ChartLlama [55] | chart | LLaVA | ✓ | ✓ | QA, generation, editing | – | Link |
| CHOCOLATE [66] | chart | GPT-4V, etc. | × | ✓ | captioning | – | Link |
| ChatCAD [177] | image | ChatGPT | × | × | QA, etc. | Medical | Link |
| LLM-CXR [90] | image | dolly-v2-3b | ✓ | ✓ | QA, generation, etc. | Medical | Link |
| Tree-GPT [43] | image | ChatGPT | × | × | QA, etc. | Remote Sensing | – |
| ChartT5 [203] | chart, table | T5 | ✓ | ✓ | QA, summarization | – | Link |
| mPLUG-PaperOwl [65] | chart, table | LLaMA | ✓ | ✓ | QA, etc. | Scientific | Link |
| U-Reader [186] | chart, table, etc. | mPLUG-Owl | ✓ | ✓ | QA, etc. | – | Link |
| DiagrammerGPT [194] | diagram | Vicuna13B | × | × | diagram generation | – | Link |
| Chain-of-Table [179] | table | GPT-3.5, etc. | × | × | QA, etc. | – | – |
| mPLUG-DocOwl [185] | document | mPLUG-Owl | ✓ | ✓ | QA, etc. | – | Link |
| Hegde et al. [62] | document | Flan-T5 | ✓ | × | QA, etc. | – | – |

**FT** denotes fine-tuning, while **VA** represents vision alignment, indicating whether the method incorporates additional techniques to align the vision and text modalities, or if it solely relies on natural language to describe figures and tables.

implemented supervised approaches to train probabilistic language models, hidden Markov models, and conditional random fields to predict them. Bhatia and Mitra [12] employed naïve Bayes and support vector machine classifiers to select relevant sentences based on their similarity and proximity to the figure caption and sentences that refer to the document elements.

### 5.3 Visual-Language Reasoning

In contrast to the previously mentioned tasks, visual-language reasoning demands a deeper understanding of the semantics inherent in figures and tables, consistently presenting a formidable challenge. Addressing this task has prompted extensive efforts within the research community. Google researchers proposed TaPas [64], a model that extends BERT's architecture to encode tables and pre-trains on large-scale tables and texts from Wikipedia. STL-CQA [164] proposed a transformers-based framework that fully leverages the structural properties of charts. It defines novel pre-training tasks aimed at incorporating structural knowledge of charts into the model. As research interest in this area continues to grow, an increasing number of evaluation datasets have been proposed, as illustrated in Table 9.

The advancement of LLMs and VLLMs has brought visual language reasoning for tables and figures into a new era, showcasing promising performance across diverse disciplines and various types of figures or tables. We summarize the related research in Table 10, considering various aspects, such as data type, backbone model, tasks, and so on. From this table, we observe that research on image understanding spans diverse disciplines, notably in medical [90, 177] and remote sensing [43, 61]. Inspired by **Chain-of-Thought (CoT)** [180], Wang et al. [179] presented Chain-of-Table, which guides LLMs to generate operations and update the table step by step. ChartT5 [203] introduced a visual language pre-training task to enhance chart understanding. Specifically, given the input chart image and the extracted OCR tokens, ChartT5 predicts the masked values of the table in the output. In addition, mPLUG-PaperOwl [65] is an OCR-free **multimodal LLM (MLLM)** for scientific diagram analysis. The authors proposed M-Paper, a diagram understanding dataset constructed by aligning diagrams in scientific papers with related paragraphs, for fine-tuning the MLLM. FinVis-GPT [178] performs instruction tuning on financial charts and their corresponding description, enabling the model to generate chart descriptions, answer questions, and predict future market trends. In addition to charts and tables, several methods, like mPLUG-DocOwl [185] and UReader [186], demonstrated proficiency in handling diverse visual-language scenarios, such as documents, web screenshots, and so forth.

Beyond reasoning tasks, several researchers explored chart generation and editing [55, 121, 171]. For instance, ChartLamma introduced a novel instruction-tuning dataset and fine-tuned the LLaVA [102] model, resulting in an MLLM capable of addressing various complex tasks, such as text-to-chart and chart editing. ChartGPT [171] fine-tuned Flan-T5 to instruct the model to generate charts based on abstract natural language descriptions.

## 6 APPLICATIONS OF SCIENTIFIC TABLES AND FIGURES

Several downstream tasks have leveraged scientific figures and tables to improve performance. We summarized those findings as follows.

### 6.1 Academic Multimodal Search

The academic search may be the most practical application of scientific tables and figures. Sandusky et al. [155] conducted an experiment on user needs for scientific tables and figures and found that many users considered tables and figures essential to identify relevant articles.

Current research and benchmarks on scientific figure retrieval are mainly used in biomedical [7, 188, 189], medical [63, 168], clinical [131], and radiological [5, 76] images. Initially, image retrieval tasks were performed by annotating manually and retrieving by a text keyword-based search [7]. The disadvantages were the high cost of expert labeling and that the labels cannot adequately express visual semantics [124]. Therefore, an increasing number of studies focused on **content-based image retrieval (CBIR)**, which focuses on extracting image features and calculating the correlation between the query and the image. Müller et al. [131] presented a comprehensive survey on the research about CBIR in medical images. They observed that the most commonly used features were color, texture, shape, and the like. PathMaster, produced by Mattie et al. [124], extracted cytology-specific features using image segmentation techniques to generate binary isolation masks and identify cytoplasm, nucleus, and nucleolus. You et al. [189] noticed that authors usually used symbols, such as arrows and lines, to indicate the important content in images, and constructed a heuristic-based method to detect these symbols. The authors argued that extracting the features of image ROIs annotated by these symbols could facilitate biomedical image retrieval. Demner-Fushman et al. [38] acquired image features by MATLAB and trained an SVM classifier to tell if an image is relevant to a query. Yu et al. [193] described a hypothesis that figures could be ranked in terms of their bio-importance. Based on this hypothesis, they developed an unsupervised NLP approach to rank figures in bioscience articles automatically.

Studies also focus on other domains and other types of scientific figure retrieval. Choudhury et al. [28] constructed a chemical figure search engine by indexing figure captions and mentions. This method can be extended to other domains efficiently but does not utilize image features. In [24], the authors proposed DiagramFlyer, designed for searching statistical figures. This system extracted figure metadata, like axis labels, axis scale, title, and legend, and allowed users to query figures using them. FigExplorer [86] was the first general figure search engine, which provided various figure exploration functions, such as exploring figures with the same topics based on the citation network. In addition, the authors fed the caption and mentions of figures into an LSTM network to learn figure embedding, which was used for the figure re-ranking function. Yang et al. [184] provided a survey on diagram image retrieval and analysis, summarizing current scientific diagram retrieval research by the method.

Compared with scientific figure retrieval, there are few studies on scientific table search. Moreover, most of the research on table retrieval takes web tables as the research object. TableSeer

[107] was a system designed for academic table searches. Liu et al. [108] proposed a table ranking algorithm and embedded it into the TableSeer system to facilitate scientific table extracting and searching; experimental results demonstrated that TableSeer outperformed the widely used search engines, like Google Scholar, in searching for information in tables.

In addition to academic research, the retrieval of scientific tables and figures has entered the stage of practical use. For instance, search engines like CiteSeer,[10] Open-i,[11] BioText,[12] Academic Explorer,[13] and others, have introduced the table/figure search function.

### 6.2 Scientific Knowledge Graph

The science knowledge graph, which represents academic research in a machine-comprehensible way, can revolutionize scientific activity by allowing information and research results to be seamlessly integrated and better matched to complex information needs [48]. Initially, research on scientific knowledge graphs concentrated solely on textual information, neglecting figure and table data. To construct a survey articles knowledge graph, Fathalla et al. [48] introduced an ontology including the research problem, approach, implementation, and evaluation. It was the first step in shifting the paradigm of scholarly communication from document-based to knowledge-based. [9] and [72] both chose “research contribution” as the core concept of ontology. These works are limited by the fact that different disciplines have specialized concepts. The concept of “problem” in the natural sciences may be referred to as a “hypothesis” or “research topic” in engineering. As a result, an ontology designed for one domain may not work well in another. Luan et al. [115] solved this problem by developing a multi-task model to extract terms, relations, and co-reference in scientific documents without designing ontology or features manually.

Compared with unstructured text, the structured information provided by tables is inherently suitable for building knowledge graphs. Furthermore, the table is an effective tool for conveying the core concepts or knowledge in work. Several scholars have recently seen the potential value of scientific tables and integrated them into scientific knowledge graphs. Kruit et al. [85] presented Tab2Know, an end-to-end system for constructing a KB from scientific tables. This system can already answer some non-trivial questions, such as “What is the F1 of BERT on TACRED?”. The authors assumed that it could be used for various other purposes, such as categorizing papers and detecting inconsistencies or plagiarized content. In particular, [136] collected survey tables from literature review papers and then extracted knowledge from them to construct a scholarly knowledge graph. Apart from this, [79] presented an approach for extracting KGs from different modalities: text, architecture images, and source code.

### 6.3 Question Answering

Intuitively, the high-quality knowledge in academic papers benefits QA systems that require scientific information. Faldu et al. [46] partially demonstrated this by introducing KI-BERT, which infused knowledge context from ConceptNet and WordNet. Experiments revealed that it significantly outperformed BERT-Large for academic subsets of QQP, QNLI, and MNLI. Tab2Know [85], mentioned in the previous subsection, is an example of using academic tables for question answering. Recently, much attention has been placed on the problem of **visual question answering (VQA)**, and some datasets in scientific styles were proposed, such as FigureQA and PlotQA. The objective of VQA is to automatically predict the response to a natural language query given an

[10] https://citeseer.ist.psu.edu/
[11] https://openi.nlm.nih.gov/
[12] https://biosearch.berkeley.edu/
[13] http://figuresearch.web.illinois.edu

image. Masry and Prince [123] combined automatic chart data extraction and table parsing methods to boost chart question-answer performance. In [45], the authors fine-tuned CLIP based on PubMed articles and verified the effectiveness of PubMedCLIP for the task of **Medical Visual Question Answering (MedVQA)**. Experiments revealed that PubMedCLIP reported the best results, with overall accuracy increases of up to 3%.

### 6.4 Scientific Claim Verification

A significant challenge in natural language processing is determining whether a textual hypothesis is entailed or rejected by the information presented [81, 174]. TabFact [22], a dataset for table-based fact verification, shifted scholars' focus away from unstructured evidence and toward structured evidence. In 2021, SemEval introduced a task called Fact Verification, and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS) [176], which prompted the utilization of scientific tables in the fact verification field. The goal of sub-task A was to determine if a statement is supported, refuted, or unknown concerning a table. At the same time, sub-task B focused on identifying the specific cells of a table that provide evidence for the statement. King001 obtained the highest score for task A by Trained 20 instances of TAPAS, SAT, and Table-BERT for an ensemble of 60 models. BreakingBERT, proposed by Jindal et al. [73], won task B by building ensemble models with TAPAS and Table-BERT Transformers in a hierarchical two-step method for 3-way classification. These solutions bridged the gap with statement verification and evidence findings using tables from scientific articles.

## 7 CHALLENGES AND POSSIBLE FUTURE DIRECTIONS

Understanding scientific tables and figures has seen tremendous progress over the last few years with the help of deep learning. There have been several successful attempts at table detection and recognition, and some of these have already been put into practice. Furthermore, the rapid evolution of LLMs and VLLMs has ushered in a new era for the interpretation of tables and figures. To advance this field, we conclude with challenges and future directions from the data, models, performance, and application perspectives.

### 7.1 Data

Data is the basis for deep learning model training and testing. Existing datasets mainly focus on document layout and table structure. Based on the dozens of datasets summarized in this paper, we believe that the following aspects should be considered in the future while building datasets.

**Diversity.** Most datasets are based on data from PubMed and arXiv, and the articles are mainly in English and from the computer and medical areas. Models need to verify their generalization on multilingual and interdisciplinary tables and figures, so it is necessary to build a dataset containing papers in diverse languages, layout styles, and disciplines.

**Complexity.** The complexity of the dataset significantly influences the model's ability to robustly handle tables and figures in real-world scenarios. Existing datasets for the detection and structure analysis task may not comprehensively consider various complex tables and figures, such as tables containing images, formulas, and so on. Furthermore, in the interpretation task, a substantial portion of datasets primarily consist of chart question and answer pairs, yet there is a noticeable scarcity of datasets comprised of flow charts and subject-related images commonly encountered in academic papers.

**Completeness.** Table/figure and text descriptions in the literature tend to complement each other. Available datasets are incomplete due to the lack of captions and notes of tables and figures, as well as descriptions in the text, which are very important for some downstream tasks, such as retrieval and question answering. Moreover, it may contribute to mining the semantics of a table/figure and multi-modal learning based on academic papers. We can learn from the success of the CLIP model [148]. It can align the text and image well and get notable performance on multiple tasks. PubMedCLIP [45] is one of the successful attempts based on PubMed articles.

### 7.2 Models

**Interpretability.** Interpretability is always a non-negligible issue when building a deep learning model. Certain phenomena should be investigated; for instance, why does the performance of some models decline dramatically as the IoU threshold rises while the performance of others barely changes? Analyzing these questions allows us to understand models better and select the one that best meets the needs of the application.

**Trustworthiness.** As an increasing number of studies delve into harnessing the capabilities of LLMs and VLLMs for understanding tables and figures, concerns have arisen regarding the likelihood of LLMs producing hallucinations. LLMs occasionally generate inaccurate content or deviate from contextual logic, posing significant risks to scientific research. Therefore, addressing how to enhance the trustworthiness of LLMs in scientific table and figure understanding emerges as a crucial research direction for the future.

**End-to-end.** Existing models, particularly those for structural analysis tasks, sometimes rely on extensive preprocessing or postprocessing procedures or are made up of several sub-modules. The training objectives of each module are inconsistent, making it difficult for the trained system to achieve optimal performance in the end; another issue is the accumulation of errors, which means that the deviation produced by the previous module may affect a later module. The end-to-end model eliminates errors caused by intermediary processes and minimizes model complexity.

**Special design for scientific documents.** Scientific documents are different from ordinary documents in many ways. For example, knowledge extraction in academic papers imposes higher requirements on entirety. Models designed for scientific documents should take these characteristics into account.

### 7.3 Performance

**Accuracy in practice.** Although the model succeeded in public datasets, this may not remain true in practical applications. For example, in Semantic Scholar's table and figure preview function, the table image frequently contains a portion of the body text.

**Inference efficiency.** Most previous studies only compared evaluation metrics, such as precision and recall, ignoring model efficiency and computing resources. Inference efficiency is a crucial factor influencing practical applications. Therefore, reducing the time and computing resources required for inference while maintaining accuracy is a contemporary problem and hot topic.

**Generalization.** The performance of a model may be influenced by various factors, including discipline, layout style, font, language, and the content of tables and figures. Nevertheless, due to dataset limitations, comprehensive research has yet to examine the impact of these aspects on the model. Thus, further investigation is necessary to explore models' generalization capabilities in these contexts fully.

### 7.4 Application

Even though scientific tables and figures are used in numerous studies, the distribution of research topics and disciplines is uneven. In mining academic tables and figures, we can either integrate discipline characteristics and focus on discipline-specific knowledge or build interdisciplinary knowledge bases or pre-training models. For example, based on the descriptive text or data given by users, we can develop a scientific style figure pre-trained model to automatically generate or beautify figures or provide color matching and layout suggestions.

## 8 CONCLUSION

This paper presents a comprehensive and unifying survey on understanding the tables and figures of scientific documents. We review these studies by categorizing them into subtasks and summarizing current challenges and limitations. We observed that there has been extensive research on detecting tables and figures in papers with a significant number of benchmark datasets. We also present a summary of the experimental results of the state-of-the-art models on benchmark datasets. A thorough review of the practical applications that utilize scientific tables and figures is also provided. Finally, we highlight some potential directions for future research. Overall, we hope this survey will serve as a hands-on reference for a better understanding of the current research development on scientific tables and figures and assist readers in advancing this field.

## REFERENCES

[1] Abdelrahman Abdallah, Alexander Berendeyev, Islam Nuradin, and Daniyar Nurseitov. 2022. TNCR: Table net detection and classification dataset. *Neurocomputing* 473 (2022), 79–97.

[2] Madhav Agarwal, Ajoy Mondal, and C. V. Jawahar. 2021. CDeC-Net: Composite deformable cascade network for table detection in document images. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 9491–9498. https://doi.org/10.1109/ICPR48806.2021.9411922

[3] Shashank Agarwal and Hong Yu. 2009. FigSum: Automatically generating structured text summaries for figures in biomedical literature. *AMIA Annual Symposium Proceedings* 2009 (2009), 6–10.

[4] Md. Ajij, Sanjoy Pratihar, Diptendu Sinha Roy, and Thomas Hanne. 2022. Robust detection of tables in documents using scores from table cell cores. *SN Computer Science* 3, 2 (March 2022), 161. https://doi.org/10.1007/s42979-022-01041-z

[5] Ceyhun Burak Akgul, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and Burak Acar. 2011. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging* 24, 2 (Jan. 2011), 208–222. https://doi.org/10.1007/s10278-010-9290-9

[6] Rabah A. Al-Zaidy and C. Lee Giles. 2015. Automatic extraction of data from bar charts. (Oct. 2015), 30. https://doi.org/10.1145/2815833.2816956

[7] Sameer Antani, L. Rodney Long, and George R. Thoma. 2004. Content-based image retrieval for large biomedical image archives. In *MEDINFO 2004*. IOS Press, 829–833.

[8] Brendan Artley. 2023. GenPlot: Increasing the scale and diversity of chart derendering data. *arXiv preprint arXiv:2306.11699* (2023).

[9] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Anna Kasprzik, Markus Stocker, Maria-Esther Vidal, and Maria-Esther Vidal. 2018. Towards a knowledge graph for science. (June 2018), 1. https://doi.org/10.1145/3227609.3227689

[10] Filip Bajić and Josip Job. 2023. Review of chart image detection and classification. *International Journal on Document Analysis and Recognition (IJDAR)* (2023), 1–22.

[11] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. https://doi.org/10.48550/arXiv.1903.10676 arXiv:1903.10676 [cs]

[12] Sumit Bhatia and Prasenjit Mitra. 2012. Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems* 30, 1 (March 2012), 3:1–3:24. https://doi.org/10.1145/2094072.2094075

[13] Jwalin Bhatt, Khurram Azeem Hashmi, Muhammad Zeshan Afzal, and Didier Stricker. 2021. A survey of graphical page object detection with deep neural networks. *Applied Sciences* 11, 12 (2021), 5344.

[14] Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document layout analysis: A comprehensive survey. *Comput. Surveys* 52, 6 (Oct. 2019), 109:1–109:36. https://doi.org/10.1145/3355610

[15] Sanket Biswas, Ayan Banerjee, Josep Lladós, and Umapada Pal. 2022. DocSegTr: An instance-level end-to-end document image segmentation transformer. *arXiv preprint arXiv:2201.11438* (2022).

[16] Joseph P. Bockhorst, John M. Conroy, Shashank Agarwal, Dianne P. O'Leary, and Hong Yu. 2012. Beyond captions: Linking figures with abstract sentences in biomedical articles. *PLoS ONE* 7, 7 (July 2012), e39618. https://doi.org/10.1371/journal.pone.0039618

[17] Sandra Carberry, Stephanie Elzer, Nancy Green, Kathleen F. McCoy, and Daniel Chester. 2004. Extending document summarization to information graphics. In *Text Summarization Branches Out*. 3–9.
[18] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12346. Springer International Publishing, Cham, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
[19] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. MapQA: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545* (2022).
[20] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. LEAF-QA: Locate, encode & attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Snowmass Village, CO, USA, 3501–3510. https://doi.org/10.1109/WACV45572.2020.9093269
[21] Jian Chen, Meng Ling, Rui Li, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Torsten Möller, Robert S. Laramee, Han-Wei Shen, Katharina Wünsche, and Qiru Wang. 2021. VIS30K: A collection of figures and tables from IEEE visualization conference publications. *IEEE Transactions on Visualization and Computer Graphics* 27, 9 (Sept. 2021), 3826–3833. https://doi.org/10.1109/TVCG.2021.3054916
[22] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-Scale Dataset for Table-Based Fact Verification. https://doi.org/10.48550/arXiv.1909.02164 arXiv:1909.02164 [cs]
[23] Xi Chen, Wei Zeng, Yanna Lin, Hayder Mahdi AI-maneea, Jonathan Roberts, and Remco Chang. 2021. Composition and configuration patterns in multiple-view visualizations. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1514–1524. https://doi.org/10.1109/TVCG.2020.3030338
[24] Zhe Chen, Michael Cafarella, and Eytan Adar. 2015. DiagramFlyer: A search engine for data-driven diagrams. (May 2015), 183–186. https://doi.org/10.1145/2740908.2742831
[25] Beibei Cheng, Sameer Antani, R. Joe Stanley, and George R. Thoma. 2011. Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval. 7874 (Jan. 2011), 294–304. https://doi.org/10.1117/12.873685
[26] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition. (Aug. 2019). https://doi.org/10.48550/arXiv.1908.04729
[27] Sagnik Ray Choudhury, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, and C. Lee Giles. 2013. Figure metadata extraction from digital documents. In *2013 12th International Conference on Document Analysis and Recognition*. 135–139. https://doi.org/10.1109/ICDAR.2013.34
[28] Sagnik Ray Choudhury, Suppawong Tuarob, Prasenjit Mitra, Lior Rokach, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, and C. L. Giles. 2013. A figure search engine architecture for a chemistry digital library. (July 2013), 369–370. https://doi.org/10.1145/2467696.2467757
[29] Arnab Ghosh Chowdhury, Martin ben Ahmed, and Martin Atzmueller. 2022. Towards tabular data extraction from richly-structured documents using supervised and weakly-supervised learning. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 1–4.
[30] Christopher Clark and Santosh Divvala. 2016. PDFFigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 143–152.
[31] Christopher Clark and Santosh K. Divvala. 2015. Looking beyond text: Extracting figures, tables and captions from computer science papers. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
[32] Mathieu Cliche, David Rosenberg, Dhruv Madeka, and Connie Yee. 2017. Scatteract: Automated extraction of data from scatter plots. Vol. 10534. 135–150. https://doi.org/10.1007/978-3-319-71249-9_9 arXiv:1704.06687 [cs, stat]
[33] Wenjing Dai, Meng Wang, Zhibin Niu, and Jiawan Zhang. 2018. Chart decoder: Generating textual and numeric information from chart images automatically. *Journal of Visual Languages & Computing* 48 (Oct. 2018), 101–109. https://doi.org/10.1016/j.jvlc.2018.08.005
[34] Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju, Christopher Tensmeyer, Sumit Shekhar, and Ritwick Chaudhry. 2019. ICDAR 2019 competition on harvesting raw tables from infographics (CHART-infographics). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 1594–1599. https://doi.org/10.1109/ICDAR.2019.00203
[35] Kenny Davila, Srirangaraj Setlur, David Doermann, Bhargava Urala Kota, and Venu Govindaraju. 2020. Chart mining: A survey of methods for automated chart analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 3799–3819.
[36] Kenny Davila, Chris Tensmeyer, Sumit Shekhar, Hrituraj Singh, Srirangaraj Setlur, and Venu Govindaraju. 2021. ICPR 2020-competition on harvesting raw tables from infographics. In *International Conference on Pattern Recognition*. Springer, 361–380.
[37] Kenny Davila, Fei Xu, Saleem Ahmed, David A. Mendoza, Srirangaraj Setlur, and Venu Govindaraju. 2022. ICPR 2022: Challenge on harvesting raw tables from infographics (CHART-infographics). In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 4995–5001.

[38] Dina Demner-Fushman, Sameer Antani, and George R. Thoma. 2007. Automatically finding images for clinical decision support. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. 139–144. https://doi.org/10.1109/ICDMW.2007.12

[39] Dazhen Deng, Yihong Wu, Xinhuan Shu, Jiang Wu, Siwei Fu, Weiwei Cui, and Yingcai Wu. 2022. VisImages: A fine-grained expert-annotated visualization dataset. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–1. https://doi.org/10.1109/TVCG.2022.3155440

[40] Yuntian Deng, Anssi Kanervisto, and Alexander Rush. 2016. What you get is what you see: A visual markup decompiler. (Sept. 2016).

[41] Yuntian Deng, David Rosenberg, and Gideon Mann. 2019. Challenges in end-to-end neural scientific table recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 894–901. https://doi.org/10.1109/ICDAR.2019.00148

[42] Harsh Desai, Pratik Kayal, and Mayank Singh. 2021. TabLeX: A benchmark dataset for structure and content information extraction from scientific tables. In *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Vol. 12822. Springer International Publishing, Cham, 554–569. https://doi.org/10.1007/978-3-030-86331-9_36

[43] Siqi Du, Shengjun Tang, Weixi Wang, Xiaoming Li, and Renzhong Guo. 2023. Tree-GPT: Modular Large Language Model Expert System for Forest Remote Sensing Image Understanding and Interactive Analysis. https://doi.org/10.48550/arXiv.2310.04698 arXiv:2310.04698 [cs]

[44] David W. Embley, Matthew Hurst, Daniel Lopresti, and George Nagy. 2006. Table-processing paradigms: A research survey. *International Journal of Document Analysis and Recognition (IJDAR)* 8, 2-3 (June 2006), 66–86. https://doi.org/10.1007/s10032-006-0017-x

[45] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as It Does in the General Domain? https://doi.org/10.48550/arXiv.2112.13906 arXiv:2112.13906 [cs]

[46] Keyur Faldu, Amit Sheth, Prashant Kikani, and Hemang Akbari. 2021. KI-BERT: Infusing Knowledge Context for Better Language and Domain Understanding. https://doi.org/10.48550/arXiv.2104.08145 arXiv:2104.08145 [cs]

[47] Ali Mazraeh Farahani, Peyman Adibi, Alireza Darvishy, Mohammad Saeed Ehsani, and Hans-Peter Hutter. 2023. Automatic chart understanding: A review. *IEEE Access* (2023).

[48] Said Fathalla, Sahar Vahdati, Sören Auer, Christoph Lange, Christoph Lange, and Christoph Lange. 2017. Towards a knowledge graph representing research findings by semantifying survey articles. (Sept. 2017), 315–327. https://doi.org/10.1007/978-3-319-67008-9_25

[49] Jinglun Gao, Yin Zhou, and Kenneth E. Barner. 2012. View: Visual information extraction widget for improving chart images accessibility. In *2012 19th IEEE International Conference on Image Processing*. IEEE, 2865–2868.

[50] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. ICDAR 2019 Competition on table detection and recognition (cTDaR). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1510–1515.

[51] Andrea Gemelli, Emanuele Vivoli, and Simone Marinai. 2022. Graph Neural Networks and Representation Embedding for Table Extraction in PDF Documents. https://doi.org/10.48550/arXiv.2208.11203 arXiv:2208.11203 [cs]

[52] Azka Gilani, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. 2017. Table detection using deep learning. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Kyoto, 771–776. https://doi.org/10.1109/ICDAR.2017.131

[53] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. 2013. ICDAR 2013 table competition. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 1449–1453.

[54] Zengyuan Guo, Yuechen Yu, Pengyuan Lv, Chengquan Zhang, Haojie Li, Zhihui Wang, Kun Yao, Jingtuo Liu, and Jingdong Wang. 2022. TRUST: An Accurate and End-to-End Table Structure Recognizer Using Splitting-Based Transformers. arXiv:2208.14687 [cs]

[55] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. https://doi.org/10.48550/arXiv.2311.16483 arXiv:2311.16483 [cs]

[56] Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, Muhammad Adnan Afzal, Muhammad Ahtsham Afzal, and Muhammad Zeshan Afzal. 2021. Current status and performance analysis of table recognition in document images with deep neural networks. *arXiv:2104.14272 [cs]* (May 2021). arXiv:2104.14272 [cs]

[57] Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2021. CasTab-DetectoRS: Cascade network for table detection in document images with recursive feature pyramid and switchable atrous convolution. *Journal of Imaging* 7, 10 (Oct. 2021), 214. https://doi.org/10.3390/jimaging7100214

[58] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[60] Yelin He, Xianbiao Qi, Jiaquan Ye, Peng Gao, Yihao Chen, Bingcong Li, Xin Tang, and Rong Xiao. 2021. PingAn-VCGroup's solution for ICDAR 2021 Competition on scientific table image recognition to latex. *arXiv preprint arXiv:2105.01846* (2021).

[61] Yingxu He and Qiqi Sun. 2023. Towards Automatic Satellite Images Captions Generation Using Large Language Models. https://arxiv.org/abs/2310.11392v1

[62] Nidhi Hegde, Sujoy Paul, Gagan Madan, and Gaurav Aggarwal. 2023. Analyzing the Efficacy of an LLM-Only Approach for Image-Based Document Question Answering. https://arxiv.org/abs/2309.14389v1

[63] William R. Hersh, Henning Müller, and Jayashree Kalpathy-Cramer. 2009. The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging* 22, 6 (Dec. 2009), 648–655. https://doi.org/10.1007/s10278-008-9154-8

[64] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. TAPAS: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4320–4333. https://doi.org/10.18653/v1/2020.acl-main.398 arXiv:2004.02349 [cs]

[65] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2023. mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model. https://doi.org/10.48550/arXiv.2311.18248 arXiv:2311.18248 [cs]

[66] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. https://doi.org/10.48550/arXiv.2312.10160 arXiv:2312.10160 [cs]

[67] Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. 2023. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11134–11143.

[68] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. https://doi.org/10.48550/arXiv.2204.08387 arXiv:2204.08387 [cs]

[69] Yilun Huang, Qinqin Yan, Yibo Li, Yifan Chen, Xiong Wang, Liangcai Gao, and Zhi Tang. 2019. A YOLO-based table detection method. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 813–818. https://doi.org/10.1109/ICDAR.2019.00135

[70] Matthew Hurst. 2001. Layout and language: Challenges for table understanding on the web. In *Proceedings of the International Workshop on Web Document Analysis*. 27–30.

[71] Matthew Francis Hurst. 2000. *The Interpretation of Tables in Texts*. Ph. D. Dissertation. University of Edinburgh.

[72] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Kheir Eddine Farfar, Manuel Prinz, Jennifer D.'Souza, Jennifer D'Souza, Gábor Kismihók, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. (Sept. 2019), 243–246. https://doi.org/10.1145/3360901.3364435

[73] Aditya Jindal, Ankur Gupta, Jaya Srivastava, Preeti Menghwani, Vijit Malik, Vishesh Kaushik, and Ashutosh Modi. 2021. BreakingBERT@IITK at SemEval-2021 Task 9: Statement Verification and Evidence Finding with Tables. arXiv:2104.03071 [cs]

[74] Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. 2017. ChartSense: Interactive data extraction from chart images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6706–6717. https://doi.org/10.1145/3025453.3025957

[75] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5648–5656.

[76] Charles E. Kahn and Cheng Thao. 2007. GoldMiner: A Radiology Image Search Engine. *AJR. American Journal of Roentgenology* 188, 6 (June 2007), 1475–1478. https://doi.org/10.2214/AJR.06.1740

[77] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. https://doi.org/10.48550/arXiv.1710.07300 arXiv:1710.07300 [cs]

[78] Sampanna Yashwant Kahu, William A. Ingram, Edward A. Fox, and Jian Wu. 2021. ScanBank: A Benchmark Dataset for Figure Extraction from Scanned Electronic Theses and Dissertations. https://doi.org/10.48550/arXiv.2106.15320 arXiv:2106.15320 [cs]

[79] Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. 2020. Multimodal knowledge graph for deep learning papers and code. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3417–3420. https://doi.org/10.1145/3340531.3417439

[80] Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. ACL-Fig: A Dataset for Scientific Figure Classification. https://doi.org/10.48550/arXiv.2301.12293 arXiv:2301.12293 [cs]

[81] Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39, 2 (1963), 170–210. https://doi.org/10.2307/411200

[82] I. Kavasidis, C. Pino, S. Palazzo, F. Rundo, D. Giordano, P. Messina, and C. Spampinato. 2019. A saliency-based convolutional neural network for table and chart detection in digitized documents. In *Image Analysis and Processing – ICIAP 2019 (Lecture Notes in Computer Science)*, Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe (Eds.). Springer International Publishing, Cham, 292–302. https://doi.org/10.1007/978-3-030-30645-8_27

[83] Pratik Kayal, Mrinal Anand, Harsh Desai, and Mayank Singh. 2021. ICDAR 2021 competition on scientific table image recognition to latex. In *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Vol. 12824. Springer International Publishing, Cham, 754–766. https://doi.org/10.1007/978-3-030-86337-1_50

[84] Elvis Koci, Maik Thiele, Josephine Rehak, Oscar Romero, and Wolfgang Lehner. 2019. DECO: A dataset of annotated spreadsheets for layout and table recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1280–1285.

[85] Benno Kruit, Hongyu He, and Jacopo Urbani. 2020. Tab2Know: Building a knowledge base from tables in scientific papers. In *The Semantic Web – ISWC 2020*, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Vol. 12506. Springer International Publishing, Cham, 349–365. https://doi.org/10.1007/978-3-030-62419-4_20

[86] Saar Kuzi, ChengXiang Zhai, Yin Tian, and Haichuan Tang. 2020. FigExplorer: A system for retrieval and exploration of figures from collections of research articles. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2133–2136. https://doi.org/10.1145/3397271.3401400

[87] Jay Lal, Aditya Mitkari, Mahesh Bhosale, and David Doermann. 2023. LineFormer: Line chart data extraction using instance segmentation. In *International Conference on Document Analysis and Recognition*. Springer, 387–400.

[88] Po-Shen Lee and Bill Howe. 2015. Detecting and dismantling composite visualizations in the scientific literature. (Jan. 2015), 247–266. https://doi.org/10.1007/978-3-319-27677-9_16

[89] Po-Shen Lee, Jevin D. West, and Bill Howe. 2018. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data* 4, 1 (March 2018), 117–129. https://doi.org/10.1109/TBDATA.2017.2689038

[90] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2023. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation. https://arxiv.org/abs/2305.11490v4

[91] Shih-Hsiung Lee and Hung-Chun Chen. 2021. U-SSD: Improved SSD based on U-Net architecture for end-to-end table detection in document images. *Applied Sciences* 11, 23 (Jan. 2021), 11446. https://doi.org/10.3390/app112311446

[92] Sheng Long Lee, Mohammad Reza Zare, and Mohammad Reza Zare. 2018. Biomedical compound figure detection using deep learning and fusion techniques. *IET Image Processing* 12, 6 (Jan. 2018), 1031–1037. https://doi.org/10.1049/iet-ipr.2017.0800

[93] Chenxia Li, Ruoyu Guo, Jun Zhou, Mengtao An, Yuning Du, Lingfeng Zhu, Yi Liu, Xiaoguang Hu, and Dianhai Yu. 2022. PP-StructureV2: A Stronger Document Analysis System. arXiv:2210.05391 [cs]

[94] Huichao Li, Lingze Zeng, Weiyu Zhang, Jianing Zhang, Ju Fan, and Meihui Zhang. 2022. A two-phase approach for recognizing tables with complex structures. In *Database Systems for Advanced Applications*, Arnab Bhattacharya, Janice Lee Mong Li, Divyakant Agrawal, P. Krishna Reddy, Mukesh Mohania, Anirban Mondal, Vikram Goyal, and Rage Uday Kiran (Eds.). Vol. 13245. Springer International Publishing, Cham, 587–595. https://doi.org/10.1007/978-3-031-00123-9_47

[95] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-Supervised Pre-Training for Document Image Transformer. https://doi.org/10.48550/arXiv.2203.02378 arXiv:2203.02378 [cs]

[96] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2019. TableBank: A benchmark dataset for table detection and recognition. *arXiv preprint arXiv:1903.01949* (2019). arXiv:1903.01949

[97] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 1918–1925.

[98] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A Benchmark Dataset for Document Layout Analysis. https://doi.org/10.48550/arXiv.2006.01038 arXiv:2006.01038 [cs]

[99] Xiao-Hui Li. 2022. Table structure recognition and form parsing by end-to-end object detection and relation parsing. *Pattern Recognition* (2022).

[100] Weihong Lin. 2022. TSRFormer: Table structure recognition with transformers. (2022).

[101] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. MMC: Advancing Multimodal Chart Understanding with Large-Scale Instruction Tuning. https://doi.org/10.48550/arXiv.2311.10774 arXiv:2311.10774 [cs]

[102] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[103] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Neural collaborative graph machines for table structure recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4533–4542.

[104] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Neural collaborative graph machines for table structure recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4533–4542.

[105] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, Bo Ren, and Rongrong Ji. 2021. Show, read and reason: Table structure recognition with flexible context aggregator. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, New York, NY, USA, 1084–1092. https://doi.org/10.1145/3474085.3481534

[106] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. 2023. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics* 76 (2023), 100761.

[107] Ying Liu, Kun Bai, Prasenjit Mitra, and C. Lee Giles. 2007. TableSeer: Automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 2007 Conference on Digital Libraries - JCDL '07*. ACM Press, Vancouver, BC, Canada, 91. https://doi.org/10.1145/1255175.1255193

[108] Ying Liu, Kun Bai, Prasenjit Mitra, and C. Lee Giles. 1999. TableRank: A ranking algorithm for table search and retrieval. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 22. Menlo Park, CA, Cambridge, MA, London, AAAI Press, MIT Press. 317.

[109] Yan Liu, Xiaoqing Lu, Yeyang Qin, Zhi Tang, and Jianbo Xu. 2013. Review of chart recognition in document images. In *Visualization and Data Analysis 2013*, Vol. 8654. SPIE, 384–391. https://doi.org/10.1117/12.2008467

[110] Yingli Liu, Changkai Si, Kai Jin, Tao Shen, and Meng Hu. 2021. FCENet: An instance segmentation model for extracting figures and captions from material documents. *IEEE Access* 9 (2021), 551–564. https://doi.org/10.1109/ACCESS.2020.3046496

[111] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. 2021. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 944–952.

[112] Luis D. Lopez, Jingyi Yu, Cecilia N. Arighi, Hongzhan Huang, Hagit Shatkay, and Cathy Wu. 2011. An automatic system for extracting figures and captions in biomedical PDF documents. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*. 578–581. https://doi.org/10.1109/BIBM.2011.26

[113] Daniel Lopresti and George Nagy. 2000. A tabular survey of automated table processing. In *Graphics Recognition Recent Advances*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Atul K. Chhabra, and Dov Dori (Eds.). Vol. 1941. Springer Berlin, Berlin, 93–120. https://doi.org/10.1007/3-540-40953-X_9

[114] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. 2021. MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognition* 117 (Sept. 2021), 107980. https://doi.org/10.1016/j.patcog.2021.107980

[115] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3219–3232. https://doi.org/10.18653/v1/D18-1360

[116] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. ChartOCR: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1917–1925.

[117] Nam Tuan Ly and Atsuhiro Takasu. 2023. An end-to-end local attention based model for table recognition. In *International Conference on Document Analysis and Recognition*. Springer, 20–36.

[118] Nam Tuan Ly, Atsuhiro Takasu, Phuc Nguyen, and Hideaki Takeda. 2023. Rethinking image-based table recognition using weakly supervised methods. *arXiv preprint arXiv:2303.07641* (2023).

[119] Pengyuan Lyu, Weihong Ma, Hongyi Wang, Yuechen Yu, Chengquan Zhang, Kun Yao, Yang Xue, and Jingdong Wang. 2023. GridFormer: Towards accurate table structure recognition via grid prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7747–7757.

[120] Chixiang Ma, Weihong Lin, Lei Sun, and Qiang Huo. 2023. Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognition* 133 (Jan. 2023), 109006. https://doi.org/10.1016/j.patcog.2022.109006

[121] Paula Maddigan and Teo Susnjak. 2023. Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. *IEEE Access* 11 (2023), 45181–45193. https://doi.org/10.1109/ACCESS.2023.3274199

[122] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* (2022).

[123] Ahmed Masry and Enamul Hoque Prince. 2021. Integrating image data extraction and table parsing methods for chart question answering. *Chart Question Answering Workshop, in Conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR).* (2021), 5.

[124] Mark E. Mattie, Lawrence Staib, Eric Stratmann, Hemant D. Tagare, James Duncan, and Perry L. Miller. 2000. PathMaster: Content-based cell image retrieval using automated feature extraction. *Journal of the American Medical Informatics Association* 7, 4 (July 2000), 404–415. https://doi.org/10.1136/jamia.2000.0070404

[125] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssisstant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-Training and Multitask Instruction Tuning. https://doi.org/10.48550/arXiv.2401.02384 arXiv:2401.02384 [cs]

[126] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 1527–1536.

[127] Nikola Milosevic, Cassie Gregson, Robert Hernandez, and Goran Nenadic. 2019. A framework for information extraction from tables in biomedical literature. *International Journal on Document Analysis and Recognition (IJDAR)* 22 (2019), 55–78.

[128] Ales Mishchenko and Natalia Vassilieva. 2011. Chart image understanding and numerical data extraction. In *2011 Sixth International Conference on Digital Information Management.* IEEE, 115–120.

[129] Prerna Mishra, Santosh Kumar, and Mithilesh Kumar Chaube. 2022. Evaginating scientific charts: Recovering direct and derived information encodings from chart images. *Journal of Visualization* 25, 2 (April 2022), 343–359. https://doi.org/10.1007/s12650-021-00800-z

[130] Ajoy Mondal, Peter Lipps, and C. V. Jawahar. 2020. IIIT-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.* Springer, 216–230.

[131] Henning Müller, Nicolas Michoux, Nicolas Michoux, Nicolas Michoux, David Bandon, David Bandon, David Bandon, and Antoine Geissbuhler. 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics* 73, 1 (Feb. 2004), 1–23. https://doi.org/10.1016/j.ijmedinf.2003.11.024

[132] Rathin Radhakrishnan Nair, Nishant Sankaran, Ifeoma Nwogu, and Venu Govindaraju. 2016. Understanding line plots using Bayesian network. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS).* IEEE, 108–113.

[133] Marcin Namysł, Alexander M. Esser, Sven Behnke, and Joachim Köhler. 2023. Flexible hybrid table recognition and semantic interpretation system. *SN Computer Science* 4, 3 (2023), 246.

[134] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. TableFormer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4614–4623.

[135] Danish Nazir, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2021. HybridTabNet: Towards better table detection in scanned document images. *Applied Sciences* 11, 18 (Jan. 2021), 8396. https://doi.org/10.3390/app11188396

[136] Allard Oelen, Markus Stocker, and Sören Auer. 2020. Creating a scholarly knowledge graph from survey article tables. In *Digital Libraries at Times of Massive Societal Transition*, Emi Ishita, Natalie Lee San Pang, and Lihong Zhou (Eds.). Springer International Publishing, Cham, 373–389. https://doi.org/10.1007/978-3-030-64452-9_35

[137] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2018. Localization recall precision (LRP): A new performance metric for object detection. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Vol. 11211. Springer International Publishing, Cham, 521–537. https://doi.org/10.1007/978-3-030-01234-2_31

[138] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. 2020. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP).* 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

[139] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* 311–318.

[140] Hai-Hong Phan. 2021. An integrated approach for table detection and structure recognition. *Journal of Research and Development on Information and Communication Technology* 2021, 1 (May 2021), 41–50. https://doi.org/10.32913/mic-ict-research.v2021.n1.974

[141] Ihsin Tsaiyun Phillips. 1996. User's reference manual for the UW English/technical document image database III. *UW-III English/Technical Document Image Database Manual* (1996).

[142] Jorge Poco and Jeffrey Heer. 2017. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum* 36, 3 (June 2017), 353–363. https://doi.org/10.1111/cgf.13193

[143] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 2020. 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*. 572–573.

[144] Jay Pujara, Pedro Szekely, Huan Sun, and Muhao Chen. 2021. From tables to knowledge: Recent advances in table understanding. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event Singapore, 4060–4061. https://doi.org/10.1145/3447548.3470809

[145] Shah Rukh Qasim, Jan Kieseler, Yutaro Iiyama, and Maurizio Pierini. 2019. Learning representations of irregular particle-detector geometry with distance-weighted graph networks. *The European Physical Journal C* 79, 7 (2019), 1–11.

[146] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. 2019. Rethinking Table Recognition Using Graph Neural Networks. https://doi.org/10.48550/arXiv.1905.13391 arXiv:1905.13391 [cs]

[147] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. 2021. LGPMA: Complicated table structure recognition with local and global pyramid mask alignment. In *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Vol. 12821. Springer International Publishing, Cham, 99–114. https://doi.org/10.1007/978-3-030-86549-8_7

[148] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[149] Sachin Raja, Ajoy Mondal, and C. V. Jawahar. 2020. Table structure recognition using top-down and bottom-up cues. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12373. Springer International Publishing, Cham, 70–86. https://doi.org/10.1007/978-3-030-58604-1_5

[150] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).

[151] Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. 2019. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 122–127. https://doi.org/10.1109/ICDAR.2019.00028

[152] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[153] Ranajit Saha, Ajoy Mondal, and C. V. Jawahar. 2019. Graphical object detection in document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 51–58. https://doi.org/10/gngxg6

[154] Naveen Saini, Sriparna Saha, Pushpak Bhattacharyya, and Himanshu Tuteja. 2020. Textual entailment–based figure summarization for biomedical articles. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 1s (April 2020), 35:1–35:24. https://doi.org/10.1145/3357334

[155] Robert J. Sandusky, Carol Tenopir, and Margaret M. Casado. 2007. Figure and table retrieval from scholarly journal articles: User needs for teaching and research. *Proceedings of the American Society for Information Science and Technology* 44, 1 (2007), 1–13. https://doi.org/10.1002/meet.1450440390

[156] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. DeepDeSRT: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. 1162–1167. https://doi.org/10.1109/ICDAR.2017.192

[157] K. C. Shahira and A. Lijiya. 2021. Towards assisting the visually impaired: A review on techniques for decoding the visual data from chart images. *IEEE Access* 9 (2021), 52926–52943.

[158] Xiangyang Shi, Yue Wu, Yue Wu, Yue Wu, Huaigu Cao, Huaigu Cao, Gully A. P. C. Burns, and Prem Natarajan. 2019. Layout-aware subfigure decomposition for complex figures in the biomedical literature. (May 2019), 1343–1347. https://doi.org/10.1109/icassp.2019.8683824

[159] Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. 2019. DeepTabStR: Deep learning based table structure recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 1403–1409. https://doi.org/10.1109/ICDAR.2019.00226

[160] Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2018. DeCNT: Deep deformable CNN for table detection. *IEEE Access* 6 (2018), 74151–74161. https://doi.org/10/gf8qz9

[161] Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. FigureSeer: Parsing result-figures in research papers. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Vol. 9911. Springer International Publishing, Cham, 664–680. https://doi.org/10.1007/978-3-319-46478-7_41

[162] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 223–232. https://doi.org/10.1145/3197026.3197040

[163] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[164] Hrituraj Singh and Sumit Shekhar. 2020. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3275–3284. https://doi.org/10.18653/v1/2020.emnlp-main.264

[165] Brandon Smock, Rohith Pesala, and Robin Abraham. 2021. PubTables-1M: Towards comprehensive table extraction from unstructured documents. (Sept. 2021). https://doi.org/10.48550/arXiv.2110.00061

[166] Carlos Soto and Shinjae Yoo. 2019. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3464–3470. https://doi.org/10.18653/v1/D19-1348

[167] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2021. A Survey of Deep Learning Approaches for OCR and Document Understanding. https://doi.org/10.48550/arXiv.2011.13534 arXiv:2011.13534 [cs]

[168] Hemant D. Tagare, C. Carl Jaffe, and James Duncan. 1997. Medical image databases: A content-based retrieval approach. *Journal of the American Medical Informatics Association* 4, 3 (May 1997), 184–198. https://doi.org/10.1136/jamia.1997.0040184

[169] Mario Taschwer and Oge Marques. 2018. Automatic separation of compound figures in scientific articles. *Multimedia Tools and Applications* 77, 1 (Jan. 2018), 519–548. https://doi.org/10.1007/s11042-016-4237-x

[170] Chris Tensmeyer, Vlad I. Morariu, Brian Price, Scott Cohen, and Tony Martinez. 2019. Deep splitting and merging for table structure decomposition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 114–121. https://doi.org/10.1109/ICDAR.2019.00027

[171] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2023. ChartGPT: Leveraging LLMs to Generate Charts from Abstract Natural Language. https://arxiv.org/abs/2311.01920v1

[172] Dominika Tkaczyk, Pawel Szostek, and Lukasz Bolikowski. 2014. GROTOAP2-the methodology of creating a large ground truth dataset of scientific articles. *D-Lib Magazine* 20, 11/12 (2014).

[173] Satoshi Tsutsui and David J. Crandall. 2017. A data driven approach for compound figure separation using convolutional neural networks. (Nov. 2017), 533–540. https://doi.org/10.1109/icdar.2017.93

[174] Johan Van Benthem. 2008. A brief history of natural logic. (2008). https://eprints.illc.uva.nl/id/eprint/279/

[175] Honglin Wan, Zongfeng Zhong, Tianping Li, Huaxiang Zhang, and Jiande Sun. 2022. Contextual transformer sequence-based recognition network for medical examination reports. *Applied Intelligence* (Dec. 2022). https://doi.org/10.1007/s10489-022-04420-4

[176] Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). https://doi.org/10.48550/arXiv.2105.13995 arXiv:2105.13995 [cs]

[177] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257* (2023).

[178] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis. https://arxiv.org/abs/2308.01430v1

[179] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding. arXiv:2401.04398 [cs]

[180] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Richter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[181] Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. 2021. AI4VIS: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics* (2021).

[182] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. 2021. TGRNet: A table graph reconstruction network for table structure recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1295–1304.

[183] Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. 2023. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data* 10, 1 (2023), 110.

[184] Liping Yang, Ming Gong, and Vijayan K. Asari. 2020. Diagram image retrieval and analysis: Challenges and opportunities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 180–181.

[185] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding. https://arxiv.org/abs/2307.02499v1

[186] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. 2023. UReader: Universal OCR-Free Visually-Situated Language Understanding with Multimodal Large Language Model. https://doi.org/10.48550/arXiv.2310.05126 arXiv:2310.05126 [cs]

[187] Burcu Yildiz, Katharina Kaiser, and Silvia Miksch. 2005. pdf2table: A method to extract table information from pdf files. In *IICAI*, Vol. 2005. Citeseer, 1773–1785.

[188] Daekeun You, Emilia Apostolova, Sameer Antani, Dina Demner-Fushman, and George R. Thoma. 2009. Figure content analysis for improved biomedical article retrieval. In *Document Recognition and Retrieval XVI*, Vol. 7247. SPIE, 276–285.

[189] Daekeun You, Emilia Apostolova, Sameer Antani, Dina Demner-Fushman, and George R. Thoma. 2009. Figure content analysis for improved biomedical article retrieval. In *Document Recognition and Retrieval XVI*, Vol. 7247. SPIE, 276–285. https://doi.org/10.1117/12.805976

[190] Fengchang Yu, Jiani Huang, Zhuoran Luo, Li Zhang, and Wei Lu. 2023. An effective method for figures and tables detection in academic literature. *Information Processing & Management* 60, 3 (2023), 103286.

[191] Hong Yu. 2006. Towards answering biological questions with experimental evidence: Automatically identifying text that summarize image content in full-text articles. *AMIA Annual Symposium Proceedings* 2006 (2006), 834–838.

[192] Hong Yu and Minsuk Lee. 2006. Accessing bioscience images from abstract sentences. *Bioinformatics* 22, 14 (July 2006), e547–e556. https://doi.org/10.1093/bioinformatics/btl261

[193] Hong Yu, Feifan Liu, and Balaji Polepalli Ramesh. 2010. Automatic figure ranking and user interfacing for intelligent figure search. *PLOS ONE* 5, 10 (2010), e12983. https://doi.org/10.1371/journal.pone.0012983

[194] Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2023. DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning. arXiv:2310.12128 [cs]

[195] Richard Zanibbi, Dorothea Blostein, and James R. Cordy. 2004. A survey of table recognition: Models, observations, transformations, and inferences. *Document Analysis and Recognition* 7, 1 (March 2004). https://doi.org/10.1007/s10032-004-0120-9

[196] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. 2021. VSR: A unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition – ICDAR 2021 (Lecture Notes in Computer Science)*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Springer International Publishing, Cham, 115–130. https://doi.org/10.1007/978-3-030-86549-8_8

[197] Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020. Summarizing and exploring tabular data in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1537–1540.

[198] Zhenrong Zhang, Jianshu Zhang, Jun Du, and Fengren Wang. 2022. Split, embed and merge: An accurate table structure recognizer. *Pattern Recognition* 126 (June 2022), 108565. https://doi.org/10.1016/j.patcog.2022.108565

[199] Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2020. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. https://doi.org/10.48550/arXiv.2005.00589 arXiv:2005.00589 [cs]

[200] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. 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*. 697–706.

[201] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12366. Springer International Publishing, Cham, 564–580. https://doi.org/10.1007/978-3-030-58589-1_34

[202] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166

[203] Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pre-training on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1314–1326. https://doi.org/10.18653/v1/2023.findings-acl.85

[204] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4154–4164. https://doi.org/10.18653/v1/D18-1448

Received 21 March 2023; revised 26 January 2024; accepted 2 April 2024