



SEMv2: Table separation line detection based on instance segmentation

Zhenrong Zhang^{a,b}, Pengfei Hu^a, Jiefeng Ma^a, Jun Du^{a,*}, Jianshu Zhang^b, Baocai Yin^b, Bing Yin^b, Cong Liu^b

^a University of Science and Technology of China, PR China

^b iFLYTEK AI Research, PR China

ARTICLE INFO

Keywords:

Table structure recognition
Table separation line detection
Instance segmentation
Conditional convolution
Table structure dataset

ABSTRACT

Table structure recognition is an indispensable element for enabling machines to comprehend tables. Its primary purpose is to identify the internal structure of a table. Nevertheless, due to the complexity and diversity of their structure and style, it is highly challenging to parse the tabular data into a structured format that machines can comprehend. In this work, we adhere to the principle of the split-and-merge based methods and propose an accurate table structure recognizer, termed SEMv2 (SEM: Split, Embed and Merge). Unlike the previous works in the “split” stage, we aim to address the table separation line instance-level discrimination problem and introduce a table separation line detection strategy based on conditional convolution. Specifically, we design the “split” in a top-down manner that detects the table separation line instance first and then dynamically predicts the table separation line mask for each instance. The final table separation line shape can be accurately obtained by processing the table separation line mask in a row-wise/column-wise manner. To comprehensively evaluate the SEMv2, we also present a more challenging dataset for table structure recognition, dubbed iFLYTAB, which encompasses multiple style tables in various scenarios such as photos, scanned documents, etc. Extensive experiments on publicly available datasets (e.g. SciTSR, PubTabNet and iFLYTAB) demonstrate the efficacy of our proposed approach. The code and iFLYTAB dataset are available at <https://github.com/ZZR8066/SEMv2>

1. Introduction

In this era of knowledge and information, document is a significant source of information for numerous cognitive processes such as knowledge database creation, optical character recognition (OCR), document retrieval, etc. As a particular entity, the tabular structure is very commonly encountered in documents. These tabular structures convey important information in a concise form. They are highly prevalent in domains such as finance, administration, research, and even archival documents. Table structure recognition (TSR) aims to recognize the table internal structure to the machine readable data mainly presented in two formats: logical structure and physical structure [1]. More precisely, logical structure only contains every cell's row and column spanning information, while the physical one additionally contains bounding box coordinates of cells. Therefore, TSR as a precursor to contextual table understanding will be beneficial in a wide range of applications [2,3].

Limited by the training datasets [4–7] used for TSR, most previous works [3,8–10] focus on document images that are obtained from digital documents (e.g., PDF files). In such a scenario, the table images

are cropped under optimal imaging conditions and are often horizontally (or vertically) aligned with a clean background and distinct table structures. However, in some real-world applications, document images may be captured by mobile cameras. Many camera-captured document images are of poor image quality, and tables contained in them may be distorted (even curved) or contain noises, which makes TSR even more challenging. Although the WTW dataset proposed recently [11] contains table images from natural scenes, it only focuses on wired tables. Parsing wireless tables is a relatively more difficult task due to the lack of visual cues to delimit cells, columns and rows. To comprehensively evaluate the performance of TSR, we present a large-scale dataset in this paper, dubbed iFLYTAB. As shown in Fig. 1, the table images in the iFLYTAB dataset are collected from various scenarios, and contain both wired and wireless tables.

Considering that a table is composed of a set of table cells and each table cell is composed of one or more basic table grids, the recently proposed split-and-merge based methods [8,10,12] consider table grids as the fundamental processing units. These methods recognize the table

* Correspondence to: National Engineering Research Center of Speech and Language Information Processing (NERC-SLIP), University of Science and Technology of China, No. 96, Jinzhai Road, Hefei, Anhui, PR China.

E-mail addresses: zrz666@mail.ustc.edu.cn (Z. Zhang), jundu@ustc.edu.cn (J. Du).

<https://doi.org/10.1016/j.patcog.2024.110279>

Received 25 June 2023; Received in revised form 18 December 2023; Accepted 10 January 2024

Available online 15 January 2024

0031-3203/© 2024 Elsevier Ltd. All rights reserved.



Fig. 1. Some table samples in the iFLYTAB dataset. (a)–(b) are wired tables. (c)–(e) are wireless tables.

structure as the following pipeline: (1) split table into basic table grid pattern (2) merge grid elements to recover table cells that span multiple rows or columns. When the TSR is performed in this way, once the “split” stage predicts erroneous results, it is difficult for the “merge” stage to rectify them. Therefore, it is essential to make the model detect table grids more accurately. The previous methods [8,10,12] complete the first stage in a bottom-up manner. Specifically, they first apply semantic segmentation [13] to predict table row/column separation lines, and then represent the intersection of detected row/column separation lines as table grids. However, segmenting table row/column separation lines in a pixel-wise manner is imprecise due to the limited receptive field. In addition, it necessitates complex mask-to-line algorithms to extract the table separation lines from the predicted segmentation results.

In this work, we follow the split-and-merge based method SEM [10], and introduce an accurate table structure recognizer, termed SEMv2. Distinct from previous segmentation-based methods [3,8,10,12] in the “split” stage, we aim to distinguish each table separation line and formulate table separation line detection as an instance segmentation task. Specifically, the table separation line mask generation is decoupled into a mask kernel prediction and a mask feature learning, which are responsible for generating convolution kernels and the feature maps to be convolved with respectively. Accurate table row/column separation lines can be easily obtained by processing table row/column separation line masks in a column-wise/row-wise manner. Moreover, compared to the sequence decoder in the “merge” stage in [10], we propose a parallel decoder based on conditional convolution to process the merging of basic table grids, which increases the decoding speed. To comprehensively evaluate the SEMv2, we also introduce a new large-scale TSR dataset iFLYTAB, which contains multiple style tables in several scenes like photos, scanned documents, etc.

The main contributions of this paper are as follows:

- Following the split-and-merge based methods, we propose the SEMv2, which introduces a novel instance segmentation framework for the table separation line detection in the “split” stage, making the “split” more robust in various scenes.
- We release the iFLYTAB dataset, which is collected from various scenarios and manually annotated carefully, to the community for advancing related research.
- Based on our proposed method, we achieve state-of-the-art performance on publicly available datasets SciTSR, PubTabNet and iFLYTAB.

2. Related work

2.1. Existing datasets

Early datasets for addressing TSR include UW-3 [14], UNLV [15], ICDAR-2013 [4], ICDAR-2019 [16] and TabStructDB [17]. However, the magnitude of these datasets is limited. To meet the requirement of data-driven approaches for TSR, large-scale datasets such as Table2Latex [18], TableBank [19] and PubTabNet [6] are proposed, but incomplete annotations still impede their development. For instance, TableBank collects 145,463 training tables from the Word and Latex documents. Each table in TableBank solely presents its corresponding HTML tag sequence, devoid of any physical coordinate information. Recently, FinTabNet [7], SciTSR [5] and PubTables-1M [20] add the cell coordinates and row-column information to become relatively comprehensive datasets for TSR. Of particular significance is PubTables-1M, which collects nearly a million fully annotated tables sourced from scientific articles. These encompass comprehensive information for table detection, recognition and functional analysis (such as column headers, projected rows and table cells). Due to the inconsistency in

annotations among these datasets, the efficacy of the model is compromised. [21] also aligns these benchmark datasets through removing both errors and inconsistency between them, which improves model performance. Although dataset scale has been significantly increased, these datasets solely focus on digital documents (e.g., PDF files). Recently, the WTW [11] dataset is introduced, which contains tables in multiple real scenes. However, it mainly focuses on wired tables, ignoring the more challenging wireless ones. To comprehensively evaluate the TSR performance, we present a new large-scale dataset iFLYTAB. Different from WTW, tables in iFLYTAB encompass both wired and wireless tables in various scenarios.

2.2. Table structure recognition

Due to the rapid development of deep learning in documents, many deep learning-based TSR approaches [6,10,22,23] have been presented. These methods can be roughly divided into three categories: bottom-up methods, image-to-markup based methods and split-and-merge based methods.

One group of bottom-up methods [22,24–27] treat words or cell contents as nodes in a graph and use graph neural networks to predict whether each sampled node pair is in the same cell, row, or column. These methods rely on the assumption that the bounding boxes of words or cell contents are available as additional inputs, which are not easy to obtain from table images directly. To eliminate this assumption, another group of methods [9,23,28–31] proposed to detect the bounding boxes of table cells directly. After cell detection, they designed some rules to cluster cells into rows and columns. However, these methods regard the cells as bounding box, which is difficult to handle the cells in distorted tables. Other methods [32–34] detect cells through detecting the corner points of cells. They can more suitable for distorted cells, but they suffer from tables containing a lot of empty cells and wireless tables.

The image-to-markup based methods [6,18,35–37] treat table structure recognition as a task similar to image-to-markup generation and directly generate the markup tags that define the structure of the table through an attention-based structure decoder. These methods rely on a large amount of training data and are inefficient as the number of table cells increases.

The split-and-merge based methods [8,10] first split a table into the basic table grid pattern, and then merge grid elements to recover table cells. Previous methods [8,10] utilize semantic segmentation [13] for identifying rows, columns within tables in the “split” stage. However, segmenting table row/column separation lines in a pixel-wise manner is inaccurate due to the limited receptive field, and heuristic mask-to-line modules designed with strong assumptions in split stage make these methods work only on tables in digital documents. To more accurately split table grids even in distorted tables, RobustTabNet [12] uses a spatial CNN-based separation line predictor to propagate contextual information across the entire table image in both horizontal and vertical directions. TSRFormer with SepRETR [38] formulates the table separation line prediction as a line regression problem and regresses separation line by DETR, but it cannot regress too long separation line well. TSRFormer with DQ-DETR [39] progressively regresses separation lines, which further enhances localization accuracy for distorted tables. GrabTab [40] flexibly fuses multiple components to robustly predicate cell edges. TRACE [41] first segments the cell corners, explicit lines and implicit lines, then obtains the grid lines from the segmentation result by post-processing, but it needs detailed labels to supervise the train progress which are not available in public datasets. In our work, we formulate the table separation line detection as the instance segmentation task. The table separation line can be accurately obtained by processing the table separation line mask in a row-wise/column-wise manner.

Table 1

The comparison between our iFLYTAB dataset and the existing datasets for table structure recognition. In the last column, we report the total number of samples for all those datasets.

Dataset	Digital		Camera-captured		Num
	Wired	Wireless	Wired	Wireless	
ICDAR-2013 [4]	✓	✓	✗	✗	156
SciTSR [5]	✓	✓	✗	✗	15,000
TableBank [19]	✓	✓	✗	✗	145,000
PubTabNet [6]	✓	✓	✗	✗	568,000
FinTabNet [7]	✓	✓	✗	✗	113,000
PubTables-1M [20]	✓	✓	✗	✗	948,000
WTW [49]	✓	✗	✓	✗	14,581
iFLYTAB	✓	✓	✓	✓	17,291

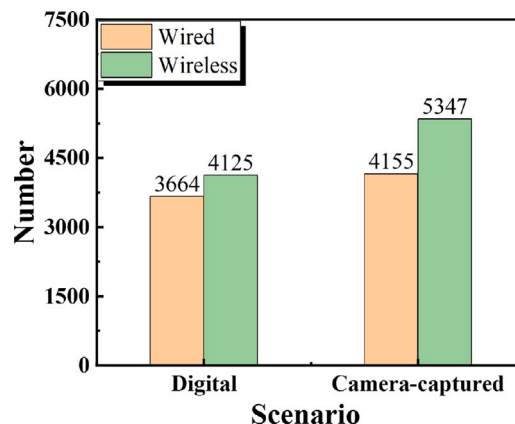


Fig. 2. Statistics of the iFLYTAB datasets.

2.3. Instance segmentation

Instance segmentation is a challenging task, as it necessitates instance-level and pixel-level predictions simultaneously. The dominant framework for instance segmentation is Mask R-CNN [42], which first detects the bounding boxes of objects and then segments the object in the box. Many works [43–45] with top performance are built on Mask R-CNN. Due to the slender shape of table separation lines, this widely utilized box-anchor based instance segmentation methods cannot be employed directly. Another approach to instance segmentation is based on dynamic filter network [46]. For example, SOLOv2 [47] and CondInst [48] learn instance-dependent convolutional kernels, which are applied to generate instance masks. Inspired by CondInst, we aim to resolve the table row/column instance-level discrimination problem, and propose the conditional table separation line detection strategy.

3. iFLYTAB

iFLYTAB collects table images of various styles from different scenarios. Specifically, as shown in Fig. 2, we collect both wired and wireless tables from digital documents, and camera-captured photos. As shown in Table 1, compared with existing datasets (e.g. SciTSR, PubTabNet, etc.) that are mainly derived from digital PDF files, the iFLYTAB includes table images captured by cameras, which contain complex image backgrounds and non-rigid image deformation. Although WTW provides table images in the photographic scenario, it ignores the more challenging wireless tables.

In terms of data labeling, we provide comprehensive annotation for each table image including physical coordinates and row/column information. Subsequently, we will present a detailed exposition on the annotation of physical coordinates and row/column information.

Physical Coordinates As illustrated in Fig. 3, the physical coordinates we have annotated comprise of both table cell and text line

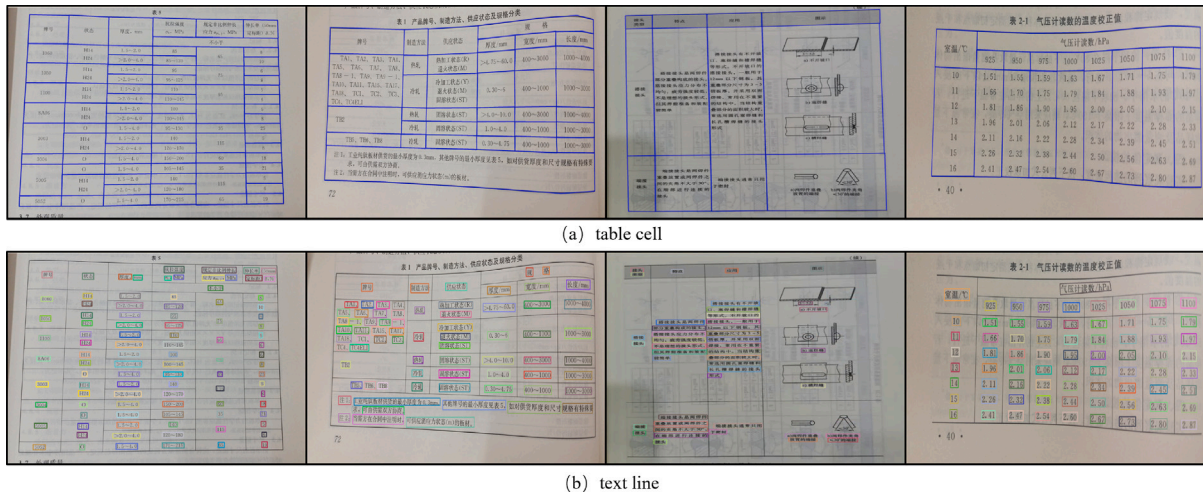


Fig. 3. The visualization of annotated physical coordinates. (a) refers to table cell polygons. (b) refers to text line polygons. Best view in zoom in.

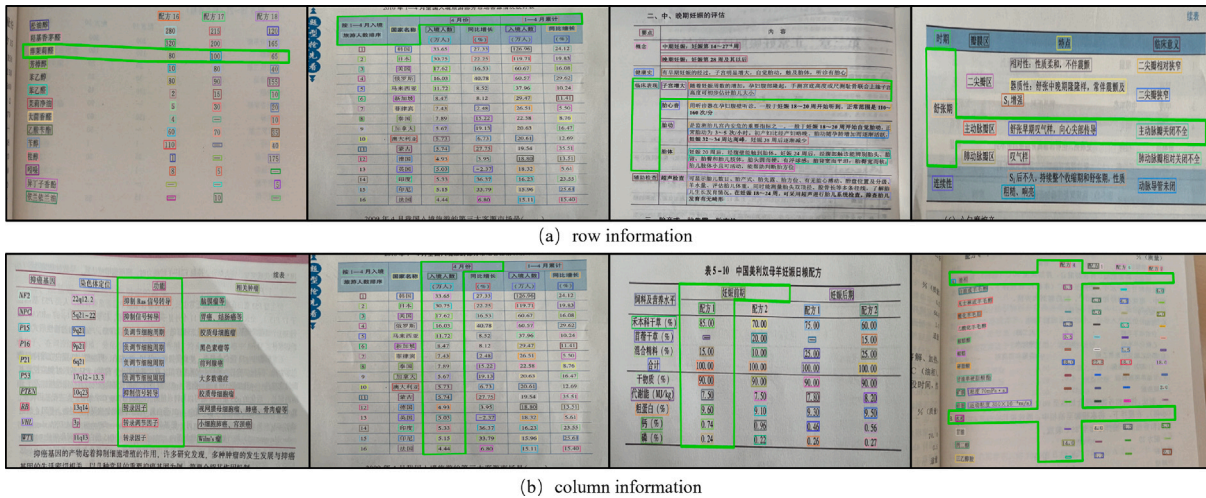


Fig. 4. The visualization of annotated row/column information. The green polygons are the annotated row/column information. (a) refers to row information. (b) refers to column information. Best view in zoom in.

polygons. Each polygon is labeled as $\{x^{lt}, y^{lt}, x^{rt}, y^{rt}, x^{rb}, y^{rb}, x^{lb}, y^{lb}\}$, representing the coordinates of the four vertices.

Row/column Information The row/column information is employed to ascertain which text lines are attributed to the same row/column in a table. Therefore, we additionally provide a series of polygons that envelop text lines belonging to the same row or column. As depicted in Fig. 4, the text lines enclosed in the green polygon indicate that they locate at the same row/column in a table.

We have manually annotated 735,781 polygons for table cells, 1,207,709 polygons for text lines, 207,972 polygons for row information, and 112,820 polygons for column information. We randomly select approximately 70% of the table images as the training set, and the rest data samples are used for testing. Finally, our iFLYTAB dataset has 12,104 training samples and 5187 testing ones.

4. Method

The schematic of our approach is depicted in Fig. 5. SEMv2 adheres to the split-and-merge based methodology of its predecessor, SEM, and is primarily comprised of three components: the splitter, the embedder, and the merger. The splitter takes the table image as input and predicts the fine grid structure of the table. The embedder extracts the grid-level feature representation of each basic table grid. The merger predicts which grids should be merged to recover the whole table structure. In the following sections, we will elucidate each component.

4.1. Splitter

Given an input table image $I \in \mathbb{R}^{H \times W \times 3}$, as illustrated in Fig. 5, the objective of the splitter is to predict the table grid structure with a set of grid bounding boxes $B \in \mathbb{R}^{M \times N \times 4}$, where M, N are the number of rows and columns occupied by the table grid structure respectively. Previous split-and-merge based methods apply the semantic segmentation to predict all table row/column separation lines in one mask and subsequently represent the intersection of detected row/column separation lines as grid bounding boxes B . In contrast to prior methods, we formulate table separation line detection as an instance segmentation task and endeavor to predict an individual mask for each table row/column separation line.

The overall architecture of our splitter is depicted in Fig. 5. The ResNet-34 [50] with FPN [51] is utilized to generate a feature pyramid with four feature maps $\{P_2, P_3, P_4, P_5\}$, whose scales are 1/4, 1/8, 1/16, 1/32 respectively. To amalgamate the information from all levels of the FPN pyramid into a single output F , we also propose a straightforward fuse operation as follows:

$$F = P_2 + Up_{1 \times} (P_3) + Up_{2 \times} (P_4) + Up_{3 \times} (P_5) \tag{1}$$

where $Up_{n \times}$ denotes the n times bilinear upsample operation. $F \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, where C denotes the number of feature channels.

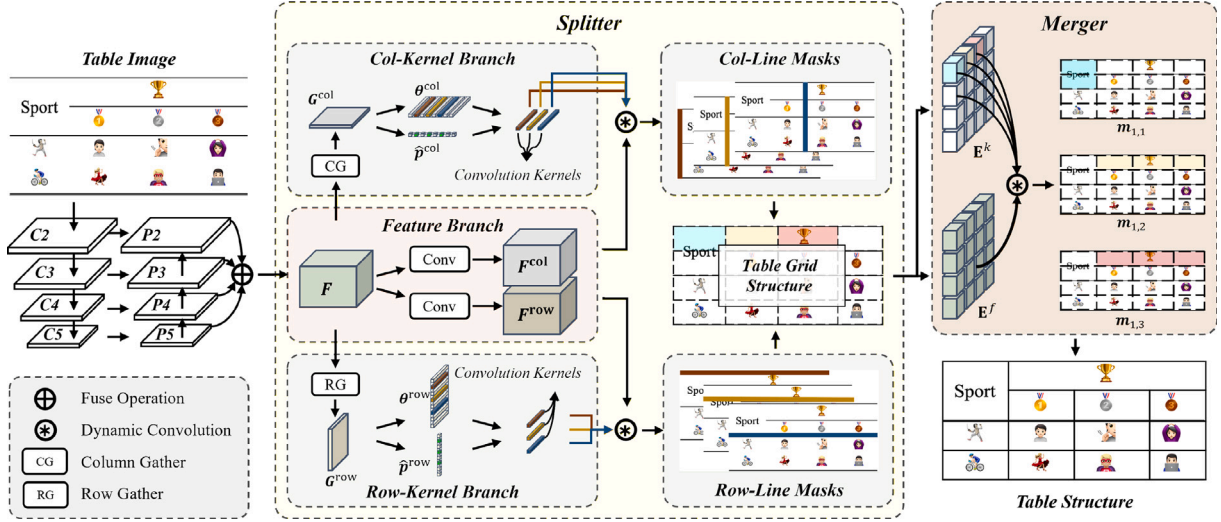


Fig. 5. The overall architecture of SEMv2. F is the feature map generated by fusing the FPN feature maps ($P2$ to $P5$). The *Splitter* module consists of *Kernel Branch* and *Feature Branch*, and predicts table separation lines between different columns or rows, which can be further processed to obtain the *Table Grid Structure*. The *Merger* module predicts the table cell to which each table grid belongs. We omit the *Embedder* module for simplicity.

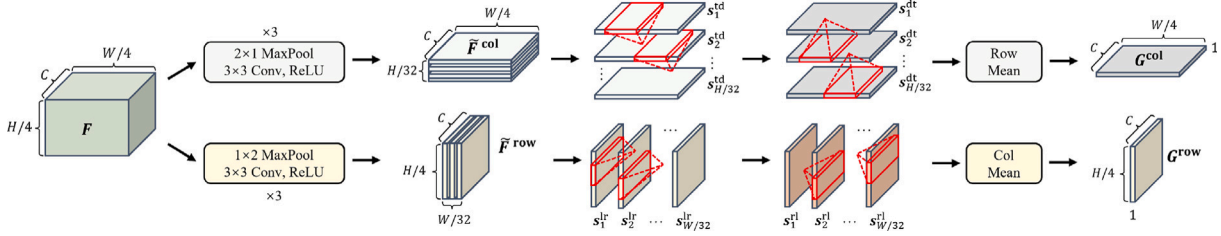


Fig. 6. Illustration of the Gather architecture. The upper part is the Column Gather. The lower part is the Row Gather.

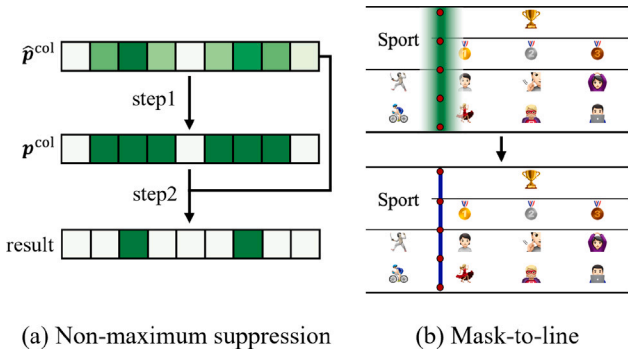


Fig. 7. The illustration of Post-processing.

Inspired by CondInst [48], we decouple the table separation line mask generation into a feature branch and a kernel branch. The feature branch contains two 1×1 convolution layers for generating $F^{\text{col}}/F^{\text{row}} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, which will be convoluted with convolution kernels from kernel branches to predict separation line masks. Since the table separation lines are usually slender and traverse the entire table image, it is necessary to design a kernel branch that has a broader receptive field. To address this issue, we propose the Gather module to capture the horizontal/vertical visual clues as shown in Fig. 6.

Taking the Column Gather as an example, we first conduct three repeated down-sampling operations on F , and each operation is composed of a sequence of a 2×1 max-pooling layer, a 3×3 convolutional layer and a ReLU activation function. The down-sampled feature map $\tilde{F}^{\text{col}} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{4} \times C}$ will be taken as the input of two

following spatial CNN modules [52]. The first spatial CNN module divides the feature map into $H/32$ slices, which are denoted as $S^{\text{td}} = \{s_i^{\text{td}} \in \mathbb{R}^{1 \times \frac{W}{4} \times C} \mid i = 1, 2, \dots, \frac{H}{32}\}$. Specifically, the topmost slice s_1^{td} is convolved by a 1×5 convolution layer, and its output feature map is merged with the next slice s_2^{td} by element-wise addition. This procedure is done iteratively so that the information can be propagated from the topmost to the bottommost effectively. The second spatial CNN module transmits information in a reversed direction. In this way, each pixel in the output feature map can leverage the structural information from both sides to enhance its feature representation ability. $G^{\text{col}} \in \mathbb{R}^{1 \times \frac{W}{4} \times C}$ is obtained by taking the row mean of the enhanced feature map. We add a linear transformation following the G^{col} to predict C -dimensional output $\theta^{\text{col}} \in \mathbb{R}^{1 \times \frac{W}{4} \times C}$. θ^{col} will be used as the weights of a 1×1 convolution layer to predict table column separation line masks. We also detect table column separation line instance by predicting $\hat{p}^{\text{col}} \in \mathbb{R}^{1 \times \frac{W}{4} \times 1}$ through a linear transformation. The loss function on \hat{p}^{col} is formulated as follows:

$$\mathcal{L}_{\text{inst}}^{\text{col}} = \sum_{i=1}^{W/4} \frac{L_{\text{bce}}(\hat{p}_i^{\text{col}}, \bar{p}_i^{\text{col}})}{W/4} \quad (2)$$

where L_{bce} is the binary cross-entropy loss, \bar{p}^{col} denotes the ground-truth distribution of starting points of table column separation lines on the x -axis. \bar{p}_i^{col} is 1 if the start point of a table column separation line is located in the i th column, otherwise 0. To eliminate the duplicated predictions of the starting point of a table column separation line in \hat{p}^{col} , as shown in Fig. 7(a), we perform non-maximum suppression as follows: (1) binarize the \hat{p}^{col} into p^{col} , (2) for the continuous pixels whose value equals 1 in p^{col} , the pixel with maximum score in \hat{p}^{col} will be selected to represent a table column separation line instance.

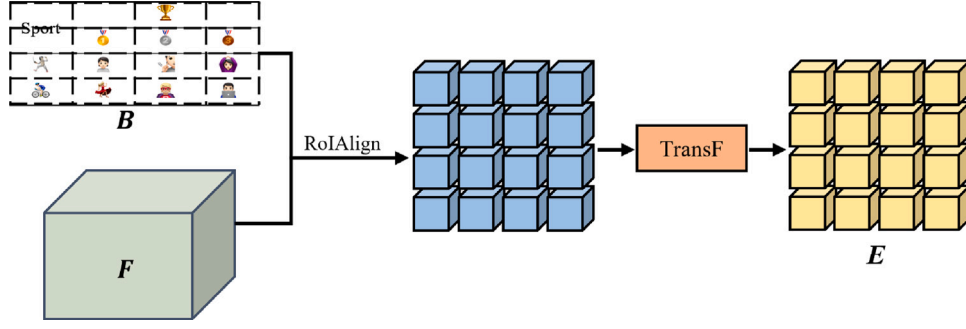


Fig. 8. The illustration of the embedder. The embedder extracts the grid-level features from F according to table grid boxes B . Then the transformer module is applied to the grid-level feature to obtain the final grid-level feature E . “TransF” denotes the transformer module.

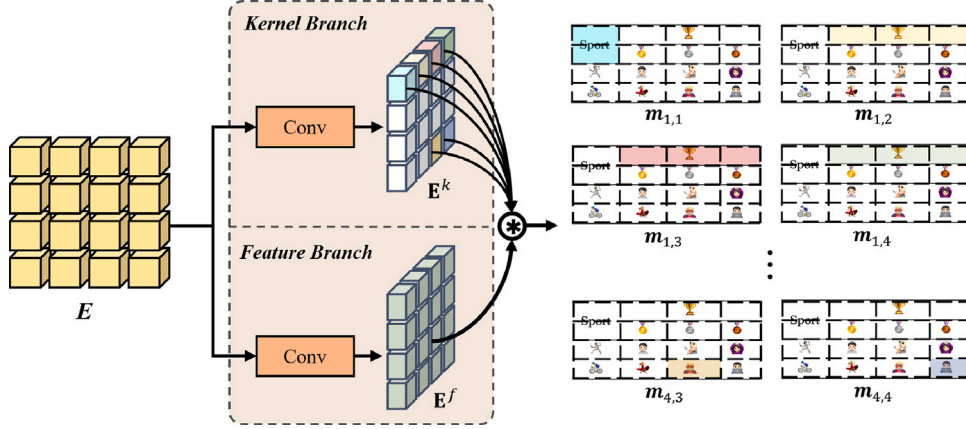


Fig. 9. The illustration of the merger. The merger first generates the feature maps E^f and kernel parameters E^k through a feature branch and a kernel branch, respectively. Then we convolute the feature map E^f with kernel parameter $e_{i,j}^k$, which is utilized as the weights of a 1×1 convolution layer, to obtain the merged map $m_{i,j}$ of the i th row, j th column grid.

According to the detected table separation line instance, we select convolution kernels from θ^{col} to conduct dynamic convolution with F^{col} to predict table column separation line masks $\hat{F}^{\text{col}} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times N^{\text{col}}}$, where N^{col} represents the number of detected table column separation line instances. The loss function on \hat{F}^{col} is defined as follow:

$$\mathcal{L}_s^{\text{col}} = \frac{1}{N^{\text{col}}} \sum_{k=1}^{N^{\text{col}}} \sum_{i=1}^H \sum_{j=1}^W L_f \left(\hat{F}_{i,j,k}^{\text{col}}, \tilde{F}_{i,j,k}^{\text{col}} \right) \quad (3)$$

in which

$$L_f(x, y) = \begin{cases} \alpha (1 - \sigma(x))^\gamma \log \left(\frac{1}{\sigma(x)} \right), & \text{if } y = 1 \\ (1 - \alpha) \sigma(x)^\gamma \log \left(\frac{1}{1 - \sigma(x)} \right), & \text{if } y = 0 \end{cases} \quad (4)$$

where \tilde{F}^{col} denotes the ground-truth of table column separation line masks. $\tilde{F}_{i,j,k}^{\text{col}}$ is 1 if the pixel in i th row, j th column and k th channel belongs to k th table column line, otherwise 0. The function L_f is actually the sigmoid focal loss [53] and the σ is the sigmoid function.

Considering the table column separation lines are typically spread through vertical direction, we process \hat{F}^{col} in a row-wise manner to obtain the final table column separation line. Specifically, as shown in Fig. 7(b), given a table column separation line mask, we first find the maximum score of each row, which is presented as a set of red dots as shown in Fig. 7(b). The table column separation line can be obtained by connecting these red dots together. The grid bounding boxes B can be derived from the intersection of table row/column separation lines.

4.2. Embedder

The embedder aims to extract the grid-level feature representations $E \in \mathbb{R}^{M \times N \times D}$, where D is the number of feature channels. As

shown in Fig. 8, we take the image-level feature map F and the well-divided table grids B obtained from the splitter as input, and apply the RoIAlign [42] to extract a fixed size $R \times R$ feature map $\hat{e}_{i,j} \in \mathbb{R}^{R \times R \times C}$ for each grid.

$$\hat{e}_{i,j} = \text{RoIAlign}_{R \times R}(F, b_{i,j}) \quad (5)$$

Then two linear transformations with a ReLU function are conducted on $\hat{e}_{i,j}$ to obtain the D -dimension output:

$$e_{i,j} = \max(0, \hat{e}_{i,j} W_1 + b_1) W_2 + b_2 \quad (6)$$

where W_1 and W_2 are learned projection matrices, b_1 and b_2 are learned biases. So far, the features of each basic table grid are still independent of each other. Therefore, we introduce the transformer [54] to capture long-range dependencies on table grid elements and utilize its output as the final grid-level features E .

4.3. Merger

The merger takes the grid-level features E as input and yields a set of merged maps, which can be formulated as: $M = \{m_{1,1}, \dots, m_{M,N}\}$, $m_{i,j} \in \{0, 1\}^{M \times N}$, $i \in \{1, \dots, M\}$, $j \in \{1, \dots, N\}$. Following the approach of the splitter, as illustrated in Fig. 9, we use a feature branch and a kernel branch to predict M jointly. Each branch contains only one 1×1 convolution layer to generate feature maps $E^f \in \mathbb{R}^{M \times N \times D}$ and kernel parameters $E^k \in \mathbb{R}^{M \times N \times D}$. We first convolute the feature map E^f with kernel parameter $e_{i,j}^k$, which is utilized as the weights of a 1×1 convolution layer, to predict the merged map $\tilde{m}_{i,j} \in [0, 1]^{M \times N}$. The loss function is formulated as follows:

$$\mathcal{L}_m = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N \frac{L_f(\tilde{m}_{i,j}, m_{i,j})}{\|\tilde{m}_{i,j}\|_1} \quad (7)$$

where $\|\cdot\|_1$ is the L1 norm, the function L_f has been defined in Eq. (4), $\tilde{m}_{i,j} \in \{0,1\}^{M \times N}$ denotes the merged map ground-truth of the i -th row, j -th column table grid. If the value of $\tilde{m}_{i,j}$ is equal to 1, then it indicates that the corresponding grid is associated with the identical table cell; otherwise, 0. The final merged map $m_{i,j}$ can be obtained by binarizing $\tilde{m}_{i,j}$.

5. Experiment

In this section, we perform comprehensive experiments on the SciTSR [5], PubTabNet [6], cTDaR [16], WTW [11] and the proposed iFLYTAB dataset to verify the effectiveness of SEMv2. We first introduce the relevant datasets and metrics, and then demonstrate the label generation and implementation details of our method. Additionally, we visualize the predicted results of our model and conduct ablation studies to analyze the effectiveness of our proposed splitter, gather and merger. Our model is compared with state-of-the-art methods on public benchmark datasets.

5.1. Datasets

SciTSR comprises of 12,000 training samples and 3,000 testing samples of axis-aligned tables extracted from the scientific literature. Furthermore, to reflect the model's ability of recognizing complex tables, it also selects all the 716 complex tables from the test set as a more challenging test subset, called SciTSR-COMP. It is worth noting that the test set of SciTSR encompasses the presence of annotation errors. To ensure a more accurate evaluation of the model's performance, we follow the RobusTabNet [12] and rectify the annotation errors present within the test set. Thus, it is pertinent to acknowledge that a comparison with other methods on the SciTSR dataset may entail a degree of unfairness, while the comparison on PubTabNet will emerge as a more equitable benchmark.

PubTabNet is a large-scale table recognition dataset, which contains 500,777 training samples and 9,115 validating samples. PubTabNet annotates each table image with information about both the structure of table and the text content with position of each non-empty table cell. All tables are also axis-aligned and collected from scientific articles.

cTDaR TrackB1-Historical [16] contains 600 training samples and 150 testing samples. It is worth noting that the table images utilized in this dataset are historical documents with handwritten. This dataset provides the physical coordinates and structure information for each table cell.

WTW contains 10,970 training images and 3,611 testing images collected from wild complex scenes. This dataset focuses on wired tabular objects only and provides the annotated information of tabular cell coordinates, and row/column information.

iFLYTAB is the proposed dataset in this paper, which contains 12,104 training samples and 5,187 testing samples. We provide comprehensive annotation for each table image including physical coordinates and structure information. However, it is worth noting that iFLYTAB does not provide annotations for the textual content within table images. In addition to the axis-aligned digital documents, the collected table images also include images taken by cameras, which are more challenging due to the complicated background and non-rigid image deformation.

5.2. Metric

We use F1-Measure [55], Tree-Edit-Distance-based Similarity (TEDS) [6], WAvg.F1 [16] and GriTS [56] metric, which are commonly adopted in table structure recognition literature and competitions, to evaluate the performance of our model for recognition of the table structure.

In order to use the F1-Measure, the adjacency relationships among the table cells need to be detected. F1-Measure measures the percentage

of correctly detected pairs of adjacent cells, where both cells are segmented correctly and identified as neighbors. When evaluating on the WTW dataset, we use the cell adjacent relationship metric [57]. This metric is a variant of F1-Measure that maps a groundtruth cell with a predicted cell according to the Intersection over Union (IoU). Here we use IoU=0.6.

TEDS measures the similarity of the tree structure of tables. While using the TEDS metric, we need to present tables as a tree structure in the HTML format. Finally, TEDS between two trees is computed as:

$$\text{TEDS}(T_a, T_b) = 1 - \frac{\text{EditDist}(T_a, T_b)}{\max(|T_a|, |T_b|)} \quad (8)$$

where T_a and T_b are the tree structure of tables in the HTML formats. EditDist represents the tree-edit distance [58], and $|T|$ is the number of nodes in T .

Since taking OCR errors into account may lead to an unfair comparison due to the different OCR models used by various TSR methods, we also employ a modified version of TEDS, called TEDS-Struct. The TEDS-Struct assesses the accuracy of table structure recognition, while disregarding the specific outcomes generated by OCR.

While using the WAvg.F1, the precision, recall and F1 value of cell detection and cell pair relation prediction need to be calculated at IoU thresholds of [0.6, 0.7, 0.8, 0.9]. The weighted average F1 (WAvg.F1) value of all IoU thresholds is defined as:

$$\text{WAvg.F1} = \frac{\sum_{i=1}^4 \text{IoU}_i \cdot \text{F1@IoU}_i}{\sum_{i=1}^4 \text{IoU}_i} \quad (9)$$

The recently proposed GriTS metric compares predicted tables and ground truth directly in matrix form and can be interpreted as an F-score over the correctness of predicted cells. Exact match accuracy considers the percentage of tables for which all cells, including blank cells, are matched exactly.

In our experiments, we align the official provided text contents to the predicted table cells according to the IOU metric. Ultimately, the output for each table image encompasses the physical coordinates of predicted cell bounding boxes, accompanied by spanning information its corresponding content. It is worth noting that the iFLYTAB dataset does not provide text content annotation. Consequently, during the evaluation of iFLYTAB, we assign a distinctive marker to each text line, signifying its individual content. The evaluation code has been made available to the public and can be accessed at the following link.¹

5.3. Label generation

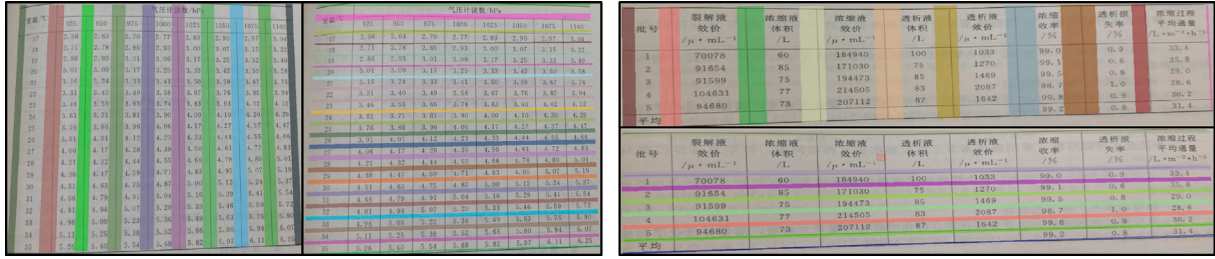
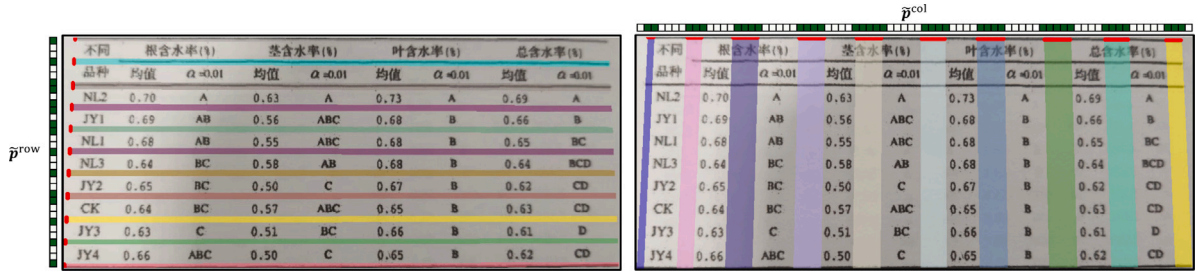
Label of Splitter Distinct from previous methods, we formulate table separation line detection as an instance segmentation task and endeavor to predict an individual mask for each table row/column separation line. Two labels necessitate generation to guide the training of the splitter, namely the table row/column separation line masks $\tilde{F}^{\text{row}}/\tilde{F}^{\text{col}}$ and table row/column line instance $\tilde{p}^{\text{row}}/\tilde{p}^{\text{col}}$.

Following the SEM [10], the table separation line mask $\tilde{F}^{\text{row}}/\tilde{F}^{\text{col}}$ are designed to maximize the size of the separator regions without intersecting any non-spanning cell content, as shown in Fig. 10.

The $\tilde{p}^{\text{row}}/\tilde{p}^{\text{col}}$ is utilized to distinguish different table row/column separation lines. As depicted in Fig. 11, we define $\tilde{p}^{\text{row}}/\tilde{p}^{\text{col}}$ as the projection of the start points of table row/column separation lines on the y-axis/x-axis.

Label of Merger Since we obtain the label of the splitter, we can partition the table into a series of basic table grids. According to the row/column information provided by the original table structure annotation, we can parse which table grids belong to the same table cell.

¹ <https://github.com/ZZR8066/SEMv2>

Fig. 10. The visualization of generated table row/column separation line mask labels $\bar{F}^{row}/\bar{F}^{col}$.Fig. 11. The visualization of generated table row/column line instance labels $\bar{p}^{row}/\bar{p}^{col}$.

5.4. Implementation details

The ResNet-34 [50] as our backbone is pre-trained on ImageNet [59]. The number of feature channels C and D is set to 256 and 512 respectively. The pool size $R \times R$ of RoAlign in the embedder is set to 3×3 . The hyperparameters α and γ of sigmoid focal loss L_f are set to 0.25 and 2. The threshold for binarization operations is set to 0.5.

The training objective of our model is to minimize the table row/column separation line segmentation loss ($\mathcal{L}_s^{row}/\mathcal{L}_s^{col}$), the table row/column separation line instance classification loss ($\mathcal{L}_{inst}^{row}/\mathcal{L}_{inst}^{col}$), and the cell merge loss (\mathcal{L}_m). The objective function for optimization is shown as follows:

$$O = \mathcal{L}_s^{row} + \mathcal{L}_s^{col} + \mathcal{L}_{inst}^{row} + \mathcal{L}_{inst}^{col} + \mathcal{L}_m \quad (10)$$

We employ the ADADELTA algorithm [60] for optimization, with the following hyper parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-9}$. We set the learning rate using the cosine annealing schedule [61] as follows:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \quad (11)$$

where η_t is the updated learning rate. η_{min} and η_{max} are the minimum learning rate and the initial learning rate, respectively. T_{cur} and T_{max} are the current number of iterations and the maximum number of iterations, respectively. Here we set $\eta_{min} = 10^{-6}$ and $\eta_{max} = 10^{-4}$.

In our experiments, we do not train the model using the training sets of all the datasets, but rather train it on the training sets of each dataset and test it on the test sets of each dataset. Following the SEM [10], SEMv2 is trained with table images in original size on the SciTSR and PubTabNet. However, in the case of the iFLYTAB dataset, we resize the table images using a randomly selected ratio within the range of [0.8, 1.2]. Different from the training strategy employed by TSRFormer [38], which follows a multi-stage process, we train the SEMv2 in an end-to-end manner. We initialize SEMv2 with the model trained on the iFLYTAB dataset and then fine-tune it on the cTDAr TrackB1-Historical dataset. We crop table regions from original images in the WTW dataset for both training and testing. Our training setup includes a single NVIDIA TESLA V100 GPU with 32 GB RAM memory and a batch size of 8 for the SciTSR, PubTabNet and cTDAr. For iFLYTAB and WTW, we utilize 8 NVIDIA TESLA V100 GPUs with 32 GB RAM memory and a batch size of 48. The whole framework was implemented using PyTorch.

Table 2

Comparison among systems from T1 to T6. ‘‘InstSeg’’ means the Instance Segmentation method which is the proposed splitter in this paper. ‘‘SemaSeg’’ means the Semantics Segmentation method which follows the splitter in SEM [10]. ‘‘ParaDec’’ means the Parallel Decoder which is the proposed merger in this paper. ‘‘SeqDec’’ means the Sequence Decoder which follows the merger in SEM [10].

System	Splitter		Gather		Merger	
	InstSeg	SemaSeg	ParaDec	SeqDec		
T1	✓	X	✓	✓		X
T2	X	✓	✓	✓		X
T3	✓	X	X	✓		X
T4	✓	X	✓	X		X
T5	✓	X	✓	✓		✓
T6	X	✓	X	X		✓

5.5. Visualization

In this section, we visualize the predicted results of both the splitter and the merger to show how SEMv2 recovers the table structure. In our work, we formulate table separation line detection as an instance segmentation task, as exemplified in Fig. 12(a–b), where different colors represent SEMv2’s predictions for distinct instances of table separation lines. By finding the maximum score on each column/row of the table row/column separation line mask, we can obtain the table row/column separation lines, as depicted in Fig. 12(c–d). As illustrated in Fig. 12(e), the table grid structure can be derived by intersecting the table row/column separation lines. Through the merger, the spanning cells structure can be restored, ultimately obtaining the final table structure, as shown in the red dashed boxes in Fig. 12(f).

5.6. Ablation study

To verify the effectiveness of each component, we conduct ablation experiments through several designed systems as shown in Table 2. The model is not modified except for the component being tested. As depicted in Table 2, T6 represents our baseline system, essentially representing SEM [10] without the textual branch. On the other hand, T1 corresponds to our proposed SEMv2. As shown in Table 3, by comparing T1 and T6, it is evident that our approach exhibits superiority over SEM in terms of both performance and efficiency.

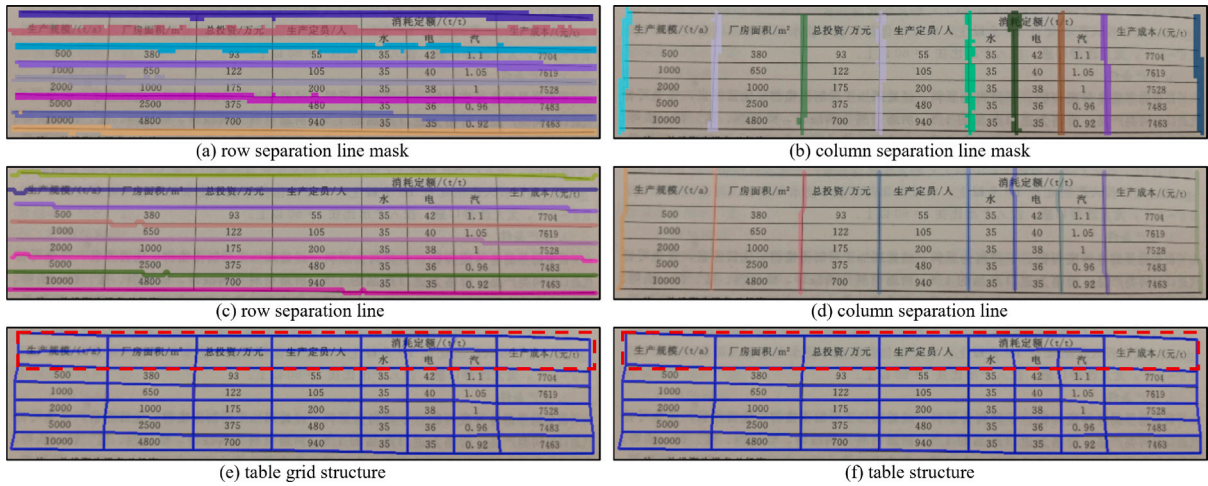


Fig. 12. The visualization of predicted results by SEMv2. The red dash boxes in (e–f) denote the difference between the table grid structure and table structure.

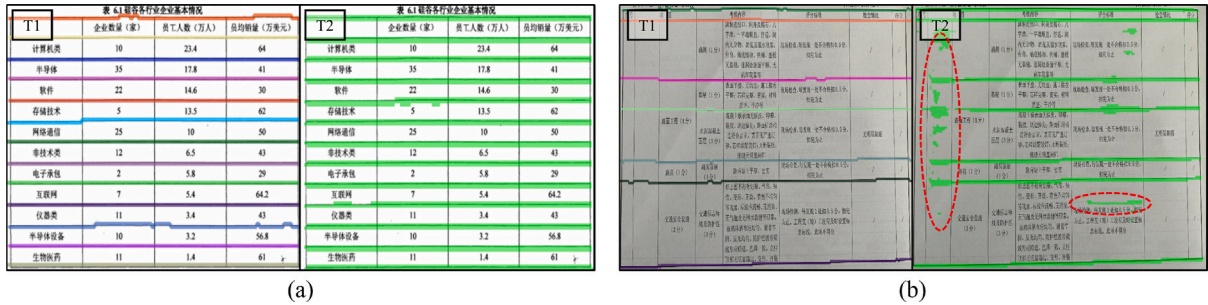


Fig. 13. The segmentation results from the splitters of the designed systems T1 and T2. (a) and (b) correspond to images extracted from digital documents and camera-captured documents, respectively.

Table 3

The performance by using different splitter modules on the SciTSR and iFLYTAB datasets. The system T6 actually is our baseline system SEM without the textual branch. Bold indicates the best result.

System	SciTSR			iFLYTAB			TEDS-Struct
	P	R	F1	P	R	F1	
T1	99.3	99.2	99.3	93.8	93.3	93.5	92.0
T2	99.1	98.5	98.8	81.9	72.7	77.0	75.8
T6	99.0	98.0	98.5	81.7	74.5	78.0	75.9

Table 4

The performance on the employment of the gather module on the SciTSR and iFLYTAB datasets. The system T6 actually is our baseline system SEM without the textual branch. Bold indicates the best result.

System	SciTSR			iFLYTAB			TEDS-Struct
	P	R	F1	P	R	F1	
T1	99.3	99.2	99.3	93.8	93.3	93.5	92.0
T3	98.6	98.2	98.4	91.1	89.8	90.4	89.3
T6	99.0	98.0	98.5	81.7	74.5	78.0	75.9

The effectiveness of the splitter. In contrast to the majority of previous “split-and-merge” based methods, we formulate table separation line detection in the “split” stage as an instance segmentation task rather than a semantic segmentation. To evaluate the efficacy of our proposed splitter, we devise the systems T1 and T2 as shown in Table 2. Specifically, T1 represents our proposed SEMv2, whereas T2 replaces the splitter with the one used in the previous state-of-the-art method, SEM [10]. As shown in Table 3, although the performance of T1 is only marginally better than T2 on datasets comprising axis-aligned scanned

Table 5

The performance and Frames Per Second (FPS) by using different merger modules on the SciTSR dataset. The system T6 actually is our baseline system SEM without the textual branch. Bold indicates the best result and underline indicates the second best.

System	SciTSR			
	P	R	F1	FPS
T1	99.3	99.2	99.3	<u>7.3</u>
T4	99.1	97.4	98.2	8.9
T5	99.3	99.2	<u>99.2</u>	2.9
T6	99.0	98.0	98.5	3.1

PDF documents (e.g., SciTSR), T1 exhibits a significantly superior performance on the iFLYTAB dataset. This is because the iFLYTAB dataset features camera-captured images with severe deformation, bending, or occlusions. Additionally, we present the segmentation results from splitters of both T1 and T2 in Fig. 13. It can be seen that both T1 and T2 can get high-quality masks on the digital document, but on the camera-captured document, the prediction result of T2 is relatively lower. It is very hard for the mask-to-line post-processing module to handle such low-quality masks well. In contrast, our instance segmentation based method can easily obtain the shape of the table separation line in a row-wise or column-wise manner, which is more robust to such challenging tables.

The effectiveness of the Gather. To illustrate the effectiveness of the Gather module, as shown in Table 2, we designed a system T3, which eliminates RowGather/ColGather, and obtains G^{row}/G^{col} by calculating the column/row mean value of F . As shown in Table 4, T1 outperforms T3 by a large margin on both SciTSR and iFLYTAB datasets, which demonstrates the effectiveness of the Gather module for capturing horizontal/vertical visual clues.

Table 6Comparison with state-of-the-art methods. Bold indicates the SOTA and underline indicates the second best.

Method	SciTSR			SciTSR-COMP			PubTabNet	
	P	R	F1	P	R	F1	TEDS	TEDS-Struct
EDD [6]	–	–	–	–	–	–	88.3	–
TabStructNet [23]	92.7	91.3	92.0	90.9	88.2	89.5	–	90.1
GraphTSR [5]	95.9	94.8	95.3	96.4	94.5	95.5	–	–
SEM [10]	97.7	96.5	97.1	96.8	94.7	95.7	93.7	96.3
LGPMA [9]	98.2	99.3	98.8	97.3	98.7	98.0	94.6	96.7
RobusTabNet [12]	99.4	99.1	<u>99.3</u>	99.0	98.4	<u>98.7</u>	–	<u>97.0</u>
TSRFormer [38]	99.5	99.4	<u>99.4</u>	99.1	98.7	98.9	–	97.5
SEMv2	99.3	99.2	<u>99.3</u>	98.7	98.6	<u>98.7</u>	–	97.5

Table 7

The performance comparison on the cTDaR TrackB1-Historical dataset. Bold indicates the SOTA.

Team	IoU=0.6		IoU=0.7		WAvg.F1
	P	R	P	R	
HCL IDORAN	25.0	5.0	23.0	4.0	6.0
NLPR-PAL	76.0	83.0	69.0	75.0	48.0
SEMv2	89.2	82.8	88.3	79.2	67.5

Table 8

The performance comparison on the WTW dataset. Bold indicates the SOTA.

Method	P	R	F1
Cycle-CenterNet [11]	93.3	91.5	92.4
TSRFormer [38]	93.7	93.2	93.4
SEMv2	93.8	93.4	93.6

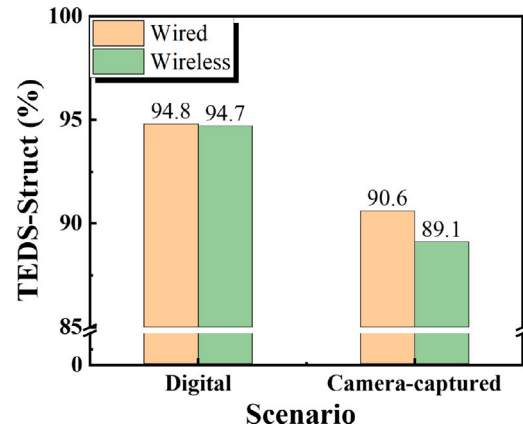
The efficiency of the merger. As shown in Table 2, we design the systems T1, T4 and T5 that employ different mergers. T4 eliminates the merger, while T5 substitutes it with the one utilized in SEM [10]. As shown in Table 5, though T4 exhibits slightly higher Frames Per Second (FPS) than T1, its performance deteriorates significantly as it disregards table cells that span multiple rows or columns. The comparison between T1 and T4 also illustrates the indispensability of the merger. The merger in SEM predicts the merging of grids in a step-by-step manner. As the number of table cells increases, the costed time in the decoding stages of T5 rise, causing the FPS of T5 to be much lower than T1 and T4.

5.7. Comparison with state-of-the-art methods

We compare our method with other state-of-the-art methods on five TSR datasets, including SciTSR, PubTabNet, cTDaR, WTW and iFLYTAB. The results are shown in Tables 6–10. Furthermore, we visualize the prediction results of our method as shown in Fig. 15. Finally, we discuss the differences between SEMv2 and other methods that adhere to the “split-and-merge” principle.

SciTSR and PubTabNet As shown in Table 6, our method achieves competitive performance compared to state-of-the-art methods. The test dataset of SciTSR exhibits a few annotation errors. Following the RobusTabNet [12], we manually rectify these annotation errors. However, it may result in an unfair comparison with other methods. The more equitable comparison can be observed on the PubTabNet. It is worth noting that the LGPMA [9] emerged as the winner of the ICDAR 2021 Competition on Scientific Literature Parsing, Task B.

cTDaR TrackB1-Historical To verify the effectiveness of our approach on tabular objects in various scenes, we conduct experiments on the cTDaR TrackB1-Historical dataset. The table images in this dataset are historical documents with handwritten. As shown in Table 7, we compare the SEMv2 with the participant teams of ICDAR 2019 Competition on Table Detection and Recognition, TrackB2 [16]. Our method outperforms other teams by a large margin.

**Fig. 14.** The performance of SEMv2 on a series of iFLYTAB subsets.

WTW To verify the effectiveness of our approach on wired distorted/curved tabular objects in wild scenes, we also conduct experiments on the WTW dataset. The results in Table 8 show that our method achieves comparable performance with state-of-the-art methods. To investigate the performance of SEMv2 in table structure recognition across various scenarios, we conducted tests on a series of subsets divided by WTW as shown in Table 9. For polygon detection methods like Cycle-CenterNet [11], SEMv2 performs better in scenarios such as curved, occluded and blurred, and extreme aspect ratio. However, in overlaid scenes, we found that SEMv2’s performance was inferior. In the overlaid subset, most of the table images exhibit significant angular rotation, which reduces the overall performance of table structure recognition. How to enhance the performance of SEMv2 in scenarios with obvious angle rotation is also a direction for our future research.

iFLYTAB The iFLYTAB dataset is distinct from SciTSR and PubTabNet in that it encompasses table images taken by cameras. These images are typically accompanied by intricate backgrounds and non-rigid deformations, which makes them more challenging. It is worth noting that detection-based methods (e.g. TabStructNet [23], LGPMA [9]) are subject to the constraints that tables are free of visual rotation or perspective transformation. This condition is difficult to satisfy when table images are camera-captured. Therefore, as shown in Table 10, we reimplement the closest method, SEM [10], for a fair comparison on the iFLYTAB dataset. Since the iFLYTAB does not provide the text content annotation, we remove the textual feature in our reimplemented SEM. It can be seen that SEMv2 outperforms SEM by a large margin. As depicted in Fig. 14, we also evaluate the performance of SEMv2 on different subsets of the iFLYTAB dataset. Notably, the model’s performance is comparatively lower in the camera-captured scenarios due to the suboptimal image qualities. Furthermore, the model performance is less satisfactory on wireless tables than on wired ones, as the former lacks crucial visual information.

Split-and-merge Previous works such as TSRFormer [38], SEM [10], SPLERGE [8] and RobusTabNet [12] all follow the “split-and-merge” principle. The main difference between these methods is found

Table 9
The F1 score comparison on different categories of the WTW dataset. Bold indicates the SOTA.

Method	Simple	Inclined	Curved	Occluded and blurred	Extreme aspect ratio	Overlaid	Multi color and grid
Cycle-CenterNet	99.3	97.7	76.1	77.4	91.9	84.1	93.7
SEMv2	97.9	96.0	87.1	85.7	96.1	75.1	92.1



Fig. 15. Some structure recognition results of SEMv2. a* refer to SciTSR, b* refer to PubTabNet, c* refer to cTDAr, d* refer to WTW, e* refer to iFLYTAB. (e1) refers to the digital wired table, (e2)~(e3) refer to camera-captured wired tables, (e4) refers to the digital wireless table, (e5)~(e6) refer to camera-captured wireless tables.

Table 10
The performance comparison with SEM on the iFLYTAB dataset. Bold indicates the SOTA.

Method	P	R	F1	TEDS-Struct	GriTS _{Con}	GriTS _{Top}
SEM	81.7	74.5	78.0	75.9	79.8	82.3
SEMv2	93.8	93.3	93.5	92.0	94.2	94.6

in their individual split stages. SEM, SPLERGE and RobusTabNet design the splitter based on the semantic segmentation, which requires a complex mask-to-line algorithm to extract table row/column separation lines from the predicted masks. In contrast, SEMv2 designs an instance segmentation-based method in the split stage, which predicts a mask for each table row/column separation line. Considering that most table row/column separation lines are spread through horizontal/vertical direction, we propose a simple “mask-to-line” algorithm, as shown in Fig. 7, that can accurately extract table row/column separation lines. TSRFormer formulates separation line prediction as a line regression problem instead of an image segmentation problem. Specifically, TSRFormer predicts reference points and using a DETR decoder to regress line coordinates. As shown in Table 8, TSRFormer and SEMv2 are comparable in performance.

6. Error analysis

WTW In this section, we present the table structure recognition results obtained by TSRFormer [38] and SEMv2 on the iFLYTAB dataset, as shown in Fig. 16. TSRFormer follows the “split-and-merge” approaches [10,12] and proposes a line regression method to detect table separation lines in the split stage. As shown in Table 8, TSRFormer performs similarly to SEMv2 in terms of performance. As shown in the first row of Fig. 16, both methods can handle scanned documents well. In the second and third rows of Fig. 16, SEMv2 performs slightly better in some natural scenes. For example, in the third row, SEMv2 can achieve more accurate results for some curved table separation lines.

Additionally, as shown in Fig. 17, we also present three scenarios where SEMv2 performs poorly on the WTW test set. As demonstrated in Fig. 17(a-b), for text outside the table area, SEMv2 mistakenly judges it as a row of the table, resulting in lower accuracy. For tables with obvious angle rotation, although SEMv2 has a certain capability to handle it, there is still a gap compared to polygon detection methods [11]. How to improve the model’s performance in scenarios with significant angle rotation will be a direction for our future research.

iFLYTAB In this section, we present erroneous table structure recognition results obtained by SEMv2 on the iFLYTAB dataset, as depicted



Fig. 16. Some table structure recognition results obtained by TSRFormer and SEMv2 on the WTW dataset. **First Column** the table structure predictions from the TSRFormer. **Second Column** the table structure predictions from the SEMv2. **Third Column** the ground truth table structure. The red dash boxes denote the different prediction regions between TSRFormer and SEMv2.

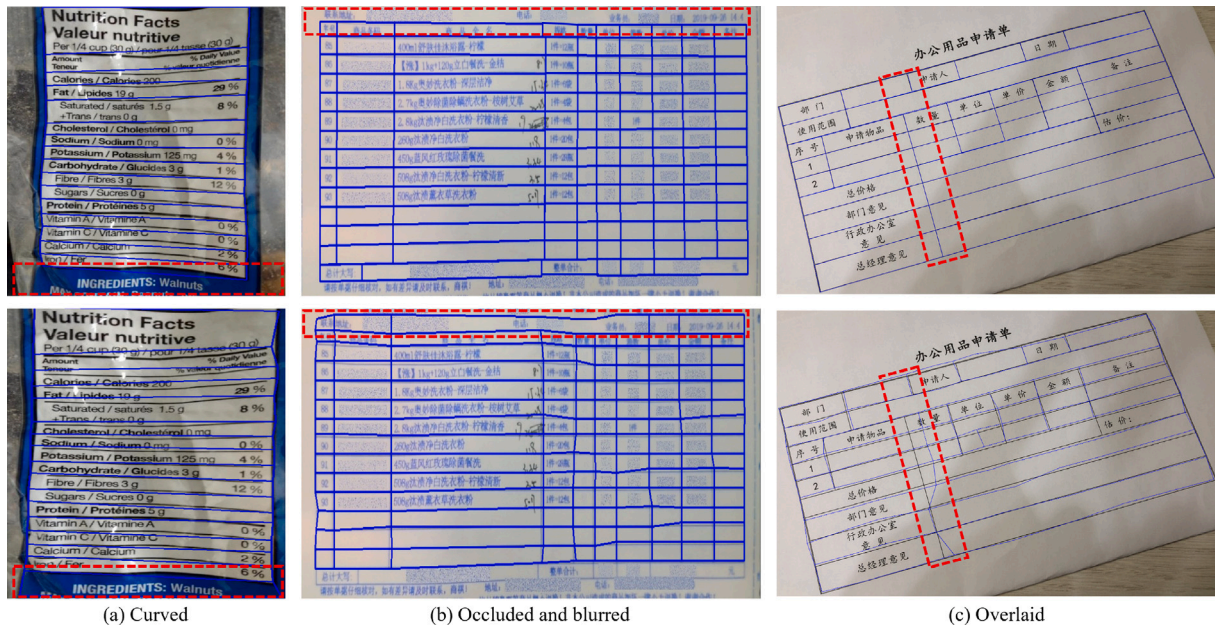


Fig. 17. Some table structure recognition results obtained by SEMv2 on different categories of the WTW dataset. **First Row** the ground truth table structure. **Second Row** the table structure predictions from the SEMv2. The red dash boxes denote the different prediction regions between ground truth and SEMv2.

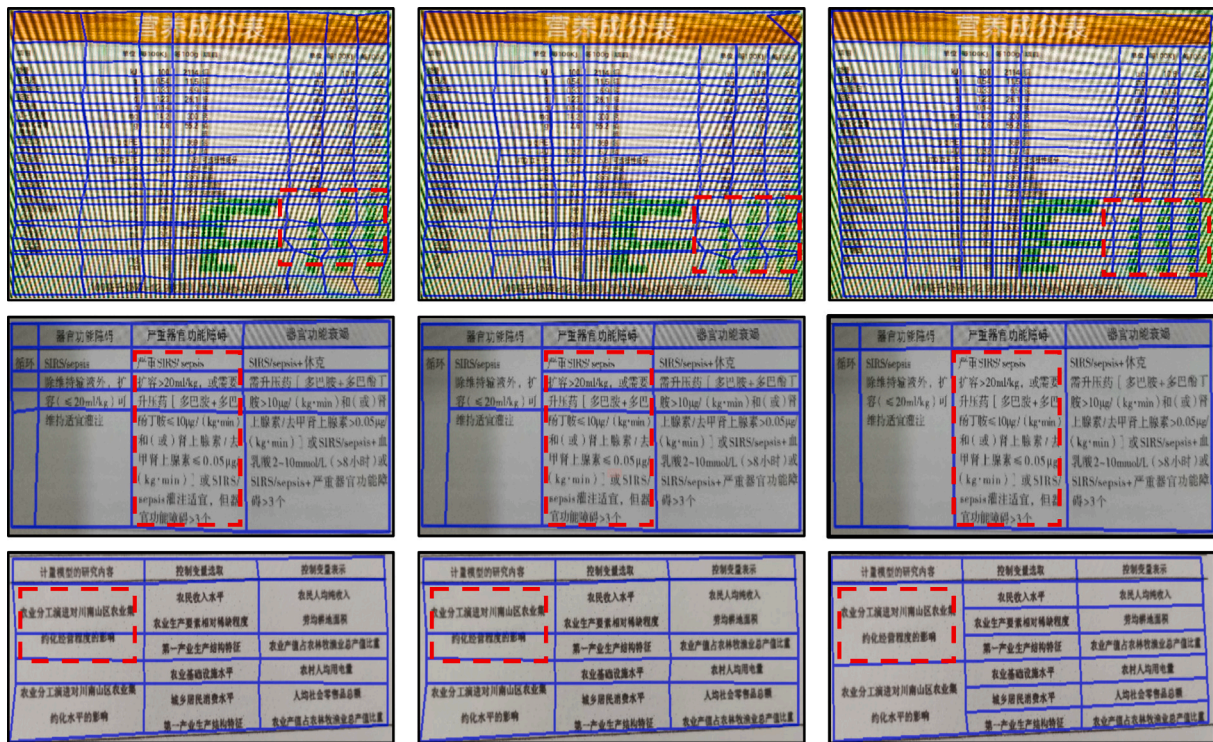


Fig. 18. Some incorrect table structure recognition results by SEMv2 on the iFLYTAB dataset. **First Column** the predictions of the table grid structure from the splitter. **Second Column** the final predictions of the table structure from the merger. **Third Column** the ground truth table structure. The red dash boxes denote the incorrect prediction regions.

in Fig. 18. As illustrated in the first row of Fig. 18, the iFLYTAB dataset comprises screen-captured table images, characterized by pronounced moiré patterns, which adversely affect the performance of our approach. The second and third rows of Fig. 18 showcase table cells within the iFLYTAB dataset that harbor multi-line content. As explicated by SEM [10], the model necessitates a comprehensive comprehension of the textual elements within the table images to facilitate more precise predictions. To enhance the efficiency of the model, SEMv2 refrains from introducing the textual branch, in contrast to its predecessor SEM, which consequently results in underperformance when confronted with table cells containing multi-line content. Devising efficient strategies to enable the model to understand the textual content in the table images and predict accurate table structure will be our future work.

7. Conclusion and future work

In this paper, we propose a novel method for tackling the problem of table structure recognition, SEMv2. It mainly contains three components including splitter, embedder and merger. Distinct from previous methods in the “split” stage, SEMv2 aims to distinguish each table line and formulate the table line detection as an instance segmentation task. The ablation experiment also demonstrates that our proposed splitter is more robust to tables in various scenarios. Moreover, we propose a parallel decoder based on conditional convolution for the merger, which significantly boosting the model’s efficiency. To comprehensively evaluate the SEMv2, we also present a more challenging dataset for table structure recognition, named iFLYTAB. We collect and annotate 17,291 tables (both wired and wireless type tables) in various scenarios such as camera-captured photos, scanned documents, etc. Some table images in iFLYTAB have intricate backgrounds and non-rigid deformations. The comprehensive experiments on the publicly available datasets (e.g. SciTSR, PubTabNet) and the proposed iFLYTAB dataset illustrate that the SEMv2 achieves a state-of-the-art performance for table structure recognition. We hope our proposed

iFLYTAB dataset can further advance future research on table structure recognition.

SEMv2 achieves state-of-the-art performance on table structure recognition, but it still has some limitations. As discussed in the error analysis section, the model’s recognition capabilities are notably compromised when dealing with images that not only exhibit significant rotational distortion but also suffer from poor image quality. Additionally, SEMv2 performs poorly in predicting the structure of cells with multi-line text due to the lack of textual information. To make SEMv2 more robust, we will study an efficient multi-modal table structure recognition scheme in our future work. This scheme will fully utilize the textual, visual, and layout information in table images to complete the prediction of table structure recognition. Furthermore, we will also study the application of computer vision techniques in table structure recognition to eliminate moiré patterns, severe image distortion, and other challenges, thereby improving the model’s prediction accuracy.

CRediT authorship contribution statement

Zhenrong Zhang: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. Pengfei Hu: Methodology, Software. Jiefeng Ma: Visualization, Writing – original draft. Jun Du: Methodology, Writing – original draft, Writing – review & editing. Jianshu Zhang: Methodology. Baocai Yin: Methodology. Bing Yin: Resources. Cong Liu: Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- [1] R. Zanibbi, D. Blostein, R. Cordy, A survey of table recognition: models, observations, transformations, and inferences, *IJDAR* (2004).
- [2] S.A. Siddiqui, M.I. Malik, S. Agne, A. Dengel, S. Ahmed, Decnt: deep deformable CNN for table detection, *IEEE Access* (2018).
- [3] S. Schreiber, S. Agne, I. Wolf, A. Dengel, S. Ahmed, DeepDeSRT: Deep learning for detection and structure recognition of tables in document images, in: *ICDAR*, 2017.
- [4] M. Göbel, T. Hassan, E. Oro, G. Orsi, *ICDAR 2013 table competition*, in: *ICDAR*, 2013.
- [5] Z. Chi, H. Huang, H.-D. Xu, H. Yu, W. Yin, X.-L. Mao, Complicated table structure recognition, 2019, arXiv.
- [6] X. Zhong, E. ShafieiBavani, A. Jimeno Yepes, Image-based table recognition: Data, model, and evaluation, in: *ECCV*, 2020.
- [7] X. Zheng, D. Burdick, L. Popa, X. Zhong, N.X.R. Wang, Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context, in: *WACV*, 2021.
- [8] C. Tensmeyer, V.I. Morariu, B. Price, S. Cohen, T. Martinez, Deep splitting and merging for table structure decomposition, in: *ICDAR*, 2019.
- [9] L. Qiao, Z. Li, Z. Cheng, P. Zhang, S. Pu, Y. Niu, W. Ren, W. Tan, F. Wu, LGPMA: Complicated table structure recognition with local and global pyramid mask alignment, in: *ICDAR*, 2021.
- [10] Z. Zhang, J. Zhang, J. Du, F. Wang, Split, embed and merge: An accurate table structure recognizer, *Pattern Recognit.* (2022).
- [11] R. Long, W. Wang, N. Xue, F. Gao, Z. Yang, Y. Wang, G.-S. Xia, Parsing table structures in the wild, in: *ICCV*, 2021.
- [12] C. Ma, W. Lin, L. Sun, Q. Huo, Robust table detection and structure recognition from heterogeneous document images, *Pattern Recognit.* (2023).
- [13] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: *CVPR*, 2015.
- [14] F. Shafait, D. Keysers, T.M. Breuel, Performance evaluation and benchmarking of six-page segmentation algorithms, *IEEE TPAMI* (2008).
- [15] A. Shahab, F. Shafait, T. Kieninger, A. Dengel, An open approach towards the benchmarking of table structure recognition systems, in: *International Workshop on Document Analysis Systems*, 2010.
- [16] L. Gao, Y. Huang, H. Déjean, J. Meunier, Q. Yan, Y. Fang, F. Kleber, E.M. Lang, *ICDAR 2019 competition on table detection and recognition (ctdar)*, in: *ICDAR*, 2019.
- [17] S.A. Siddiqui, I.A. Fateh, S.T.R. Rizvi, A. Dengel, S. Ahmed, DeepTabStR: Deep learning based table structure recognition, in: *ICDAR*, 2019.
- [18] Y. Deng, D. Rosenberg, G. Mann, Challenges in end-to-end neural scientific table recognition, in: *ICDAR*, 2019.
- [19] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, TableBank: Table benchmark for image-based table detection and recognition, in: *LREC*, 2020.
- [20] B. Smock, R. Pesala, R. Abraham, PubTables-1M: Towards comprehensive table extraction from unstructured documents, in: *CVPR*, 2022.
- [21] B. Smock, R. Pesala, R. Abraham, Aligning benchmark datasets for table structure recognition, 2023, arXiv.
- [22] Z. Chi, H. Huang, H.-D. Xu, H. Yu, W. Yin, X.-L. Mao, Complicated table structure recognition, 2019, arXiv.
- [23] S. Raja, A. Mondal, C.V. Jawahar, Table structure recognition using top-down and bottom-up cues, in: *ECCV*, 2020.
- [24] X. Li, F. Yin, H. Dai, C. Liu, Table structure recognition and form parsing by end-to-end object detection and relation parsing, *Pattern Recognit.* (2022).
- [25] S.R. Qasim, H. Mahmood, F. Shafait, Rethinking table recognition using graph neural networks, in: *ICDAR*, 2019.
- [26] W. Xue, Q. Li, D. Tao, Res2TIM: Reconstruct syntactic structures from table images, in: *ICDAR*, 2019.
- [27] H. Liu, X. Li, B. Liu, D. Jiang, Y. Liu, B. Ren, Neural collaborative graph machines for table structure recognition, in: *CVPR*, 2022.
- [28] S. Raja, A. Mondal, C. Jawahar, Visual understanding of complex table structures from document images, in: *WACV*, 2022.
- [29] Z. Li, Y. Li, Q. Liang, P. Li, Z. Cheng, Y. Niu, S. Pu, X. Li, End-to-end compound table understanding with multi-modal modeling, in: *ACM MM*, 2022.
- [30] H. Liu, X. Li, B. Liu, D. Jiang, Y. Liu, B. Ren, R. Ji, Show, read and reason: Table structure recognition with flexible context aggregator, in: *ACM MM*, 2021.
- [31] X. Zheng, D. Burdick, L. Popa, X. Zhong, N.X.R. Wang, Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context, in: *WACV*, 2021.
- [32] H. Xing, F. Gao, R. Long, J. Bu, Q. Zheng, L. Li, C. Yao, Z. Yu, LORE: Logical location regression network for table structure recognition, 2023, arXiv.
- [33] W. Xue, B. Yu, W. Wang, D. Tao, Q. Li, Tgrnet: A table graph reconstruction network for table structure recognition, in: *ICCV*, 2021.
- [34] R. Long, W. Wang, N. Xue, F. Gao, Z. Yang, Y. Wang, G.-S. Xia, Parsing table structures in the wild, in: *ICCV*, 2021.
- [35] A. Nassar, N. Livathinos, M. Lysak, P. Staar, TableFormer: Table structure understanding with transformers, in: *CVPR*, 2022.
- [36] N.T. Ly, A. Takasu, P. Nguyen, H. Takeda, Rethinking image-based table recognition using weakly supervised methods, in: *ICPRAM*, 2023.
- [37] Y. Huang, N. Lu, D. Chen, Y. Li, Z. Xie, S. Zhu, L. Gao, W. Peng, Improving table structure recognition with visual-alignment sequential coordinate modeling, in: *CVPR*, 2023.
- [38] W. Lin, Z. Sun, C. Ma, M. Li, J. Wang, L. Sun, Q. Huo, Tsrformer: Table structure recognition with transformers, in: *ACM MM*, 2022.
- [39] J. Wang, W. Lin, C. Ma, M. Li, Z. Sun, L. Sun, Q. Huo, Robust table structure recognition with dynamic queries enhanced detection transformer, *Pattern Recognit.* (2023).
- [40] H. Liu, X. Li, M. Gong, B. Liu, Y. Wu, D. Jiang, Y. Liu, X. Sun, Grab what you need: Rethinking complex table structure recognition with flexible components deliberation, 2023, arXiv.
- [41] Y. Baek, D. Nam, J. Surh, S. Shin, S. Kim, TRACE: Table reconstruction aligned to corner and edges, 2023, arXiv.
- [42] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: *ICCV*, 2017.
- [43] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: *CVPR*, 2018.
- [44] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C.C. Loy, D. Lin, Hybrid task cascade for instance segmentation, in: *CVPR*, 2019.
- [45] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring R-CNN, in: *CVPR*, 2019.
- [46] B. De Brabandere, X. Jia, T. Tuytelaars, L. Van Gool, Dynamic filter networks, in: *NIPS*, 2016.
- [47] X. Wang, R. Zhang, T. Kong, L. Li, C. Shen, SOLOv2: Dynamic and fast instance segmentation, *NIPS* (2020).
- [48] Z. Tian, C. Shen, H. Chen, Conditional convolutions for instance segmentation, in: *ECCV*, 2020.
- [49] R. Long, W. Wang, N. Xue, F. Gao, Z. Yang, Y. Wang, G.-S. Xia, Parsing table structures in the wild, in: *ICCV*, 2021.
- [50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *CVPR*, 2016.
- [51] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: *CVPR*, 2017.
- [52] X. Pan, J. Shi, P. Luo, X. Wang, X. Tang, Spatial as deep: Spatial cnn for traffic scene understanding, in: *AAAI*, 2018.
- [53] T. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *ICCV*, 2017.
- [54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *NIPS*, 2017.
- [55] M. Hurst, A constraint-based approach to table structure derivation, in: *ICDAR*, 2003.
- [56] B. Smock, R. Pesala, R. Abraham, GriTS: Grid table similarity metric for table structure recognition, 2022, arXiv.
- [57] M.C. Göbel, T. Hassan, E. Oro, G. Orsi, A methodology for evaluating algorithms for table understanding in PDF documents, in: *ACM Symposium on Document Engineering, DocEng '12, Paris, France, September 4–7, 2012*, 2012.
- [58] M. Pawlik, N. Augsten, Tree edit distance: Robust and memory-efficient, *Inf. Syst.* (2016).
- [59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, ImageNet: A large-scale hierarchical image database, in: *CVPR*, 2009.
- [60] M.D. Zeiler, ADADELTA: An adaptive learning rate method, 2012.
- [61] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, 2017.



Zhenrong Zhang received his B.Eng. degree from the Department of Computer Science and Engineering, North-eastern University of China, in 2020. He is currently a Ph.D. candidate at the University of Science and Technology of China (USTC). His current research area includes table recognition, document analysis and natural language generation.



Pengfei Hu received his B.Eng. degree from the Department of Computer Science and Technology, University of Science and Technology of China (USTC) in 2021. He is currently a Ph.D. candidate at the University of Science and Technology of China (USTC). His current research area includes OCR and document analysis.



Jiefeng Ma received his B.Eng. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2020. He is currently a Ph.D. candidate at the University of Science and Technology of China (USTC). His current research area includes natural language generation, information extraction, and document analysis.



Jun Du received his B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above years, he worked as an Intern for two 9-month periods at Microsoft Research Asia (MSRA), Beijing. In 2007, he worked as a Research Assistant for 6 months in the Department of Computer Science, University of Hong Kong. From July 2009 to June 2010, he worked at iFLYTEK Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.