IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 27 November 2023, accepted 19 December 2023, date of publication 22 December 2023, date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3346326

APPLIED RESEARCH

DAFA: A Dual-Awareness Feature Aggregator for Table Structure Recognition on Medical Examination Reports

XUANRUI HONG[1], KAI ZHA[1], ZIMING FENG[2], ZHEXUAN CHEN[1], XIAODONG DU[1], AND MIN LIU[3]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Credit Card Center, China Merchants Bank, Shanghai 201201, China
[3]Department of Integrative Medicine, Obstetrics and Gynecology Hospital, Fudan University, Shanghai 200011, China

Corresponding author: Min Liu (liumin7325@fckyy.org.cn)

ABSTRACT Table structure recognition (TSR) is crucial for document analysis, particularly for medical examination report tables (MERTs), impacting efficiency and decision-making in healthcare. Most models for TSR utilize either Graph Convolution Neural Networks (GCNs) or Transformers with html sequences for structure recognition. These methods, however, face challenges with graph inductive bias and instability in training, respectively. We observe that cells within the same row or column of a table not only are closely aligned in their vertical and horizontal coordinates, respectively, but also exhibit highly similar features. In previous work, the spatial feature of coordinates was often only used for concatenation with image features, text features, etc. We believe that explicitly utilizing the unique spatial properties of tables can better encode table features. In this paper, we introduce a novel structure named Dual-Awareness Feature Aggregator (DAFA) for table, which leverages attention mechanisms to effectively extract table features. Based on it, we design an end-to-end model called DAFA-Net requiring only images as input, without the need for additional information such as texts. In addition, we try to address the prevalent challenge of recognizing cross-row and cross-column cells in TSR — a scenario frequently encountered in medical examination reports — by introducing a modified focal loss known as CRCC loss. We conduct extensive experiments on four popular datasets. This includes a dataset specifically dedicated to medical data and others that mirror the complexity typically encountered in medical tables. Experimental results show the effectiveness and potential of our DAFA-Net for TSR within the healthcare sector.

INDEX TERMS Table structure recognition, graph attention, focal loss, medical examination reports.

I. INTRODUCTION

Tabular form is powerful in data organization and information representation. Huge amounts of tabular data are disclosed in medical and many other scenarios. Enterprises extract data from tables for disease diagnosis and analysis. And there are many downstream applications for tabular data extraction on medical examination report tables. For example, it is the base to transform the medical document data into other formats, such as JSON. After that, it can be easier to store the information in tables into data warehouses and perform complicated data analysis [1]. This data analysis can be employed to enhance clinical decision-making, predict diseases, manage patients, and support medical research, ultimately contributing to the improvement of healthcare and treatment. However, it is difficult to extract tabular data automatically from medical examination reports which are always unstructured files such as PDF files or images. Besides, tables in medical examination reports are typically without lines and contain cells that span across rows or columns, which further complicates table recognition. Furthermore, table data extraction is a tedious and error-prone task when doing it manually, so we need a more intelligent and automatic method.

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

There are three sub-tasks for table data extraction: table region detection, table structure recognition, and table content extraction. Table region detection is to detect the region of the table in the images or PDFs. Table structure recognition is to extract table layout information from the table image and rebuild its structure. Table content extraction is to extract information with its structure from the table, which can be provided to downstream tasks. These three sub-tasks can be carried out one by one, or be treated as a whole process. Table structure recognition, which is the topic in this paper, lies between two other processes and plays a vital role in table data extraction.

Table structure recognition has been researched for many years, and a lot of methods are proposed. The existing methods can be categorized in the following three main categories: heuristic-based approach, traditional machine learning approach, and deep learning approach.

1) Heuristic-based approach is to specify a set of rules and apply the rules on table structure recognition, such as PDF2TABLE [2], PDF-TREX [3] and an automatic table metadata extraction algorithm [4].

2) There are shallow learning models for table recognition, such as SVM [5], the Naive Bayes classifier [6], [7], and the decision trees [8].

3) Deep learning, especially GNN models, is recently adopted in table structure recognition, whereby a node in GNN refers to a cell in the table. Chi et al. [9] propose a model GraphTSR, which predicts edge relationship by extract nodes and edges feature. Xuen et al. [10] propose ReS2TIM to predict relationships(up/down/left/right) of adjacent cells. Besides, some Transformer-based models have been used in this area. Liu et al. [11] propose FLAG-Net, in which a Flexible Context Aggregator (FLAG) combining CNN and Transformer is presented. Ye et al. [12] fulfill the task of recognizing table structure from image to HTML based on MASTER [13] and use multi Transformer layers.

In current research, we observe that methods based on Graph Neural Networks (GNNs) commonly employ the K-Nearest Neighbors (KNN) algorithm to construct a graph representation of the table [14], [15], effectively viewing the relationships between table cells as static connections. Such hardcoded connectivity undeniably restricts the model's understanding of the inherent heterogeneity and complex structure of tables. Therefore, for tables with variable structures, especially those common in the medical field that lack lines and contain cells spanning across rows or columns, this approach proves to be inadequate. On the other hand, although Transformer-based sequence models have shown potential in understanding tabular data, their dependency on large-scale training samples and the instability during training pose significant barriers to their application [11]. In light of this, We incorporate Graph Attention Networks (GATs) as the graph feature attention mechanism into our proposed Dual-Awareness Feature Aggregator (DAFA) for table. The advantage is that GAT is capable of capturing complex interactions between nodes and dynamically learning the relationships between cells through an adaptive attention mechanism.

Furthermore, we observe that cells in the same row or column of a table not only share similarities in their vertical or horizontal coordinates but also exhibit strong consistency in their features. For instance, in a medical examination report, a column may represent different values of the same test indicator. These values are not only consistent in their spatial layout but also similar in their magnitude or numerical characteristics, so we propose a graph spatial attention mechanism (GSAT) based on spatial distance as a weighting factor. Although this weighting factor is static in form, it reflects the inherent spatial relationship between cells, providing a location-based principle for feature aggregation. The GSAT grounded in fixed spatial relationships simplifies the learning process of the model and offers a robust strategy for handling complex table layouts, such as those with cells spanning multiple rows or columns. Based on the motivations outlined, we introduce the Dual-Awareness Feature Aggregator (DAFA) for table, which unifies the adaptive learning capabilities of GAT with the spatial insights of GSAT.

The main contributions of our work are as follows:

- We propose a Dual-Awareness Feature Aggregator (DAFA) for table to aggregate and extract the feature of cells in the table, based on which we present an end-to-end network DAFA-Net for table structure recognition with the table image as input;
- To address the problem of recognizing cross-row or cross-column cells correctly, which often occurs in medical tables, we design a modified focal loss called CRCC loss;
- Experiment results show our method's superior performance over the GCN-based baseline TGRNet [16].

In this paper, we first introduce some related works in Sec. II, including table cells detection and table structure recognition, to show the current studies and the main challenge. Then, we show the proposed method in Sec. III and display the experimental results in Sec. IV. Finally, we conclude the whole paper and suggest some future works in Sec. V.

II. RELATED WORK

This section introduces the prior works related to our method. We first introduce some methods about table cells detection task in Sec. II-A. Then we introduce the efficient table structure recognition methods including methods based on graph in Sec. II-B, based on sequence in Sec. II-C and based on region in Sec. II-D.

A. TABLE CELLS DETECTION

The table cells detection branch plays a crucial role in the entire table structure recognition as it provides the coordinate information of cells for the subsequent structure reasoning tasks. Many detection models have been adapted for cell detection, including object detection models such as Faster R-CNN [17], Mask R-CNN [18], as well as text line detection models like MASTER [13] and convolutional recurrent neural network (CRNN) [19]. Wei et al. [20] integrated a table projection module (TPM) into the framework of the Faster R-CNN model, projecting the entire table horizontally and vertically to obtain grid division features. This model directly provides the position coordinates of cells. Raja et al. [21], in their end-to-end table structure recognition model, employed Mask R-CNN for table cell recognition. They incorporated dilated convolutions into the Region Proposal Network (RPN) to capture features in both horizontal and vertical directions and combined pyramid features in the Feature Pyramid Network (FPN) to extract features of table cells at different sizes. In the TableMaster model [12], they used the MASTER [13] to recognize the positions of cells in the HTML sequence. Additionally, they employed PSENet [22] to detect text lines and match the coordinates of cell boxes obtained from MASTER. Qiao et al. proposed the LGPMA [23], which primarily identifies the positions of table cells and then uses a post-processing algorithm to reconstruct the table's structure. This model achieves cell detection using a soft pyramid mask learning mechanism for both local and global image features. Their work focuses on detecting and identifying reliable table cells, serving as the foundation for table structure recognition.

B. TABLE STRUCTURE RECOGNITION BASED ON GRAPH

Tables and graphs exhibit structural similarities. Graph-based methods for table structure recognition treat cells as nodes to construct a graph. Nodes are typically represented using cell visual, textual, and positional features that have been processed through embedding. Subsequently, Graph Neural Networks (GNNs) are used for feature extraction, resulting in two types of outputs. One type represents the relationships between cells [9], [10], [11], [14], [15], [24], [25], such as being in the same row, same column, or neither. The other type represents the logical positions of cells, such as the row and column indices of each cell [16], [26].

C. TABLE STRUCTURE RECOGNITION BASED ON SEQUENCE

Table structures can be represented in sequential text formats such as LaTeX or HTML. Consequently, a category of table structure recognition methods has emerged, known as "image to sequence" methods [12], [27], [28], [29], [30]. These methods typically employ an encoder-decoder network architecture, where the encoder is used to extract table features, and the decoder is used to generate the table structure sequence. For example, EDD [31] utilizes the CNN as the encoder, employs the RNN as the structural decoder to produce table HTML tags, and uses a text decoder to output the table's textual content.

D. TABLE STRUCTURE RECOGNITION BASED ON REGION

Tables are essentially grid structures formed by horizontal and vertical lines within a region, but the merged cells and missing lines add complexity to table structure recognition. In region-based methods, the process typically involves two steps: split and merge [32], [33], [34], [35], [36]. During the split phase, grids are formed by identifying lines; in the merge phase, grids are combined based on the features of adjacent grids. Additionally, TableStrRec [37] adopts a different approach. It proposes marking rows and columns as regular or irregular while identifying them, and then applying post-processing to restore the table structure.

III. METHOD

A. OVERALL ARCHITECTURE

As shown in Fig. 1, the DAFA-Net proposed in this paper mainly consists of four modules: backbone(Sec. III-B), cell detection(Sec. III-C), table feature extractor(Sec. III-D) and table structure reasoning(Sec. III-E).

Firstly, the table image is input into backbone for visual feature extraction, and then cell detection module outputs bounding box coordinates of the cells. Given visual features and bounding boxes, table feature extractor containing two stacked DAFAs extracts and aggregates cell features. Finally, table structure reasoning module reasons each cell's logical position, namely row index of start, row index of end, column index of start, and column index of end.

B. BACKBONE MODULE

The backbone module is based on ResNet50 [38] and FPN [39] networks. Given a table image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$, four feature maps are obtained through backbone network and they are $\boldsymbol{f}_1 \in \mathbb{R}^{256\times 128\times 128}$, $\boldsymbol{f}_2 \in \mathbb{R}^{256\times 64\times 64}$, $\boldsymbol{f}_3 \in \mathbb{R}^{256\times 32\times 32}$, $\boldsymbol{f}_4 \in \mathbb{R}^{256\times 16\times 16}$. After feature fusion, we get $\boldsymbol{f}_{\text{up-to-down}} \in \mathbb{R}^{1024\times 128\times 128}$.

C. CELL DETECTION MODULE

Inspired by the work of Xue et al. [16], we also use the method of masking cells in table images to mask the pixels in cell areas. After processing the feature map extracted from backbone, the text box coordinates are extracted. After these text box coordinates are aligned, they will be used to build graph structure nodes in GNN models. The loss of cell detection is formed with 3 parts in Eq. (1):

$$\mathcal{L}_{\text{cell\_det}} = \mathcal{L}_{\text{row\_seg}} + \mathcal{L}_{\text{col\_seg}} + \mathcal{L}_{\text{cell\_seg}} \tag{1}$$

$\mathcal{L}_{\text{row\_seg}}$ here is the cross entropy loss of row segmentation coordinates, while $\mathcal{L}_{\text{col\_seg}}$ and $\mathcal{L}_{\text{cell\_seg}}$ are column and cell segmentation respectively. The weights are equal by default for their relative similar implication.

D. DUAL-AWARENESS TABLE FEATURE AGGREGATOR

In this section, we introduce the core structure, Dual-Awareness Feature Aggregator (DAFA) for table in the proposed network of this paper.

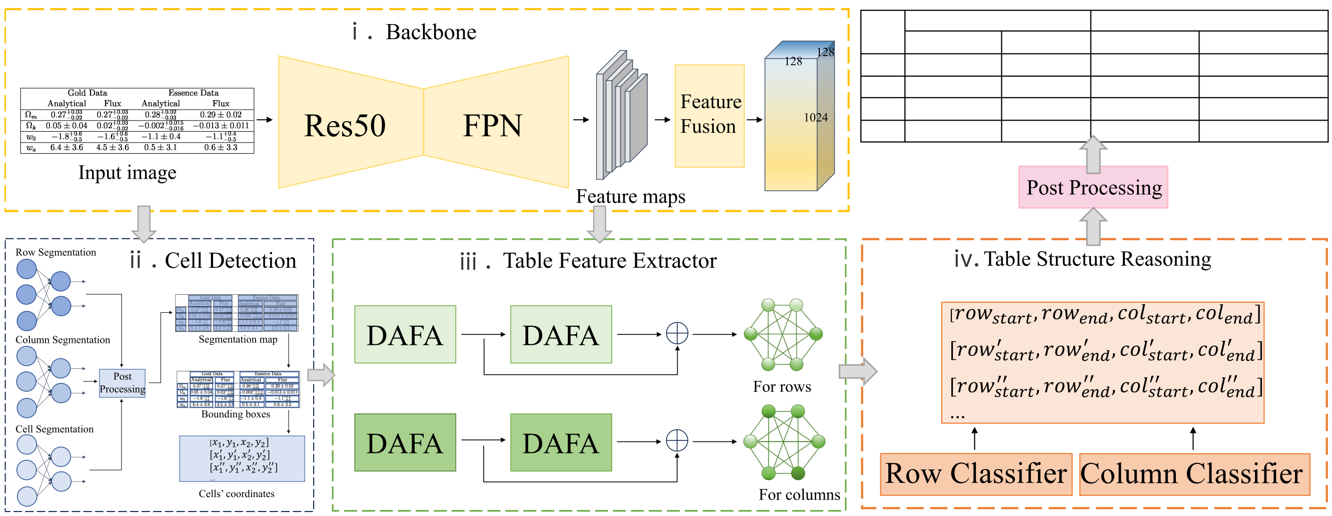


FIGURE 1. The architecture of DAFA-Net proposed in the paper. The input, a table image, sequentially passes through four modules: backbone, cell detection, table feature extractor, and table structure reasoning. The output is row index of start, row index of end, column index of start, column index of end for each cell in the table, effectively representing the table structure. The detailed introductions of each module can be found in Sec. III-B, Sec. III-C, Sec. III-D, Sec. III-E.

It is very natural to consider a table's cells as a graph's nodes. Therefore, a table is modeled as a graph in DAFA. We set each cell in the table as a node v, so N cells constitutes a node set $\boldsymbol{V} \in \mathbb{R}^N$. The feature of node v_i is represented as $\boldsymbol{x}_i \in \mathbb{R}^d$, where d is the feature dimension. So the features of N cells constitute a feature set $\boldsymbol{X} \in \mathbb{R}^{N\times d}$. Using K-Nearest Neighbor to construct a graph will make the graph carry the strong inductive bias, and it will constrain feature aggregation from cells. So we construct a fully connected graph G and make use of the attention mechanism for feature extraction. The edge between node v_i and v_j is $e_{i,j}$, and the edge set is $\boldsymbol{E} \in \mathbb{R}^{N\times(N-1)/2}$. Based on the above, we get the graphical description of the table, $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{E})$.

1) FEATURE INITIALIZATION

The feature $\boldsymbol{x}_i$ of the node v_i consists of two parts as shown in Eq. (2), Eq. (3) and Eq. (4)

$$\boldsymbol{x}_i = \boldsymbol{f}_i^{\text{visual}} || \boldsymbol{f}_i^{\text{spatial}} \tag{2}$$

$$\boldsymbol{f}_i^{\text{visual}} = \text{Roi\_Align}(\boldsymbol{f}_{\text{up-to-down}}, \boldsymbol{b}_i) \tag{3}$$

$$\boldsymbol{f}_i^{\text{spatial}} = \left(\frac{x_i^c}{W}, \frac{y_i^c}{H}, \frac{w_i}{W}, \frac{h_i}{H}\right) \tag{4}$$

where $\boldsymbol{b}_i = (x_1^i, y_1^i, x_2^i, y_2^i)$ and (x_i^c, y_i^c) represent the bounding box and the center coordinates of the node v_i respectively. And w_i, h_i, W, H are the width and height of the node v_i and the entire table. Besides, we use Roi_Align from Mask R-CNN [18] for obtaining the feature of the node v_i from the feature map.

2) GRAPH FEATURE ATTENTION

Introducing the attention mechanism from Graph Attention Network (GAT) [40], we input $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{E})$ and get $\boldsymbol{G}' = (\boldsymbol{X}', \boldsymbol{E})$ as output, where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{N\times d}$ and $\boldsymbol{X}' = (\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_N) \in \mathbb{R}^{N\times d}$, d is the feature dimension. Attention coefficient $\alpha_{i,j}$ between node v_i and v_j is calculated by Eq. (5).

$$\alpha_{i,j} = \frac{\exp(\text{LeakyReLU}(\boldsymbol{a}^\top[\boldsymbol{W}\boldsymbol{x}_i || \boldsymbol{W}\boldsymbol{x}_j]))}{\sum_{k\in N}\exp(\text{LeakyReLU}(\boldsymbol{a}^\top[\boldsymbol{W}\boldsymbol{x}_i || \boldsymbol{W}\boldsymbol{x}_k]))} \tag{5}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{a} \in \mathbb{R}^{2d}$. Attention coefficient $\alpha_{i,j}$ indicates the relative importance of node v_j to node v_i from the perspective of feature similarity. Then we update the node features according to the $\alpha_{i,j}$ as Eq. (6).

$$\boldsymbol{x}'_i = \sigma\left(\sum_{j\in N}\alpha_{i,j}\boldsymbol{W}\boldsymbol{x}_j\right) \tag{6}$$

where σ represents activation functions.

3) GRAPH SPATIAL ATTENTION

Inspired by the Graph Attention Network (GAT) [40], we propose Graph Spatial Attention (GSAT) mechanism for table structure recognition task. As a method of organizing content, tables exhibit a unique characteristic. For instance, when we consider columns, the horizontal coordinates of nodes within the same column are remarkably similar, indicating a strong correlation among their features. Suppose the column header indicates "Financial Amount," then each cell's text will represent a specific monetary value, such as "\$1,000," "\$500," or "\$10,500." Visually, these values may share common features like the currency symbol prefix, similar digit formatting, and decimal alignment. Given this correlation, leveraging spatial coordinates to guide feature aggregation can enhance the performance of machine learning models in processing tabular data. This method utilizes the inherent spatial properties of tables to direct the learning process of the model, offering an informed approach to understand and integrate data effectively. Based on it, our GSAT aggregates node features by measuring the distance among nodes. Because in the final table structure reasoning module, the logical position (row_{start}, row_{end}, col_{start}, col_{end}) of the nodes are predicted for row and column respectively, we aggregate features of nodes for row and column respectively. For this reason, it is reasonable to use Euclidean distance in the GSAT.

In the GSAT, we input $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{E})$, (x_i^c, y_i^c) and get $\boldsymbol{G}'' = (\boldsymbol{X}'', \boldsymbol{E})$ as output, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{N \times d}$, $\boldsymbol{X}'' = (\boldsymbol{x}_1'', \ldots, \boldsymbol{x}_N'') \in \mathbb{R}^{N \times d}$, d is the feature dimension, and (x_i^c, y_i^c) is the center coordinates of the node v_i's bounding box. Attention coefficient $\alpha_{i,j}$ between node v_i and v_j is calculated by Eq. (7) and Eq. (8).

$$\alpha_{i,j}^{row} = \frac{\exp\left(\tanh(a(y_i^c - y_j^c)^2)\right)}{\sum_{k \in N} \exp\left(\tanh(a(y_i^c - y_k^c)^2)\right)} \tag{7}$$

$$\alpha_{i,j}^{col} = \frac{\exp\left(\tanh(a(x_i^c - x_j^c)^2)\right)}{\sum_{k \in N} \exp(\tanh(a(x_i^c - x_k^c)^2))} \tag{8}$$

where $a \in \mathbb{R}$. Attention coefficient $\alpha_{i,j}$ indicates the relative importance of node v_j to node v_i from the perspective of position similarity for row and column respectively. Then we update the node features according to the $\alpha_{i,j}$ as:

$$\boldsymbol{x}_{i_{\text{row}}}'' = \sigma\left(\sum_{j \in N} \alpha_{i,j}^{\text{row}} \boldsymbol{W} \boldsymbol{x}_j\right) \tag{9}$$

$$\boldsymbol{x}_{i_{\text{col}}}'' = \sigma\left(\sum_{j \in N} \alpha_{i,j}^{\text{col}} \boldsymbol{W} \boldsymbol{x}_j\right) \tag{10}$$

where σ represents activation functions and $\boldsymbol{W} \in \mathbb{R}^{d \times d}$. It is worth noting that in the table feature extractor module as shown in Fig. 1, we use a pair of feature extractors stacked by DAFA to extract features for row and column respectively in fact. For convenience, we call $\boldsymbol{X}''$ uniformly for $\boldsymbol{X}_{\text{row}}''$ and $\boldsymbol{X}_{\text{col}}''$ in the following content.

The overall structure of DAFA is shown in Fig. 2. The updated feature matrix $\boldsymbol{X}'$ based on GAT and the updated feature matrix $\boldsymbol{X}''$ based on GSAT are concatenated in the last dimension $\boldsymbol{X}_{\text{new}} = \boldsymbol{X}' || \boldsymbol{X}''$. We adopt the residual connection and layer normalization like Transformer [41] for avoiding vanishing gradient and improving the training speed as shown in Eq. (11) and Eq. (12):

$$\boldsymbol{A} = \text{Add\&Norm}(\boldsymbol{X}, \boldsymbol{X}_{\text{new}}) \tag{11}$$

$$\boldsymbol{Y} = \text{Add\&Norm}(\text{FFN}(\boldsymbol{A}), \boldsymbol{A}) \tag{12}$$

where the FFN$(\cdot)$ represents the feed forward network and Add&Norm$(\cdot)$ represents the element-wise and layer normalization. The graph $\boldsymbol{G} = (\boldsymbol{Y}, \boldsymbol{E})$ is the final output of the Dual-Awareness Feature Aggregator (DAFA).

E. TABLE STRUCTURE REASONING MODULE

The node feature matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{N \times d}$ obtained by the table feature extractor is used for structure reasoning in this module. Referring to TGRNet [16], the logical position of the cell in the table can be represented as $\boldsymbol{s_i} = (row_i^{start}, row_i^{end}, col_i^{start}, col_i^{end}) \in \mathbb{R}^4$, so the final table structure recognized can be represented as $\boldsymbol{S}_{\text{table}} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N) \in \mathbb{R}^{N \times 4}$, N is the total number of cells in the table. We define the maximum index of the row for the table as $R \in \mathbb{R}$, the row index of start for the cell i as $row_i^{start} = r \in \{0, 1, 2, \ldots, R-2\}$. And then we turn multi-classification into binary-classification problem by Eq. (13). For simplicity, row_i^{end}, col_i^{start}, col_i^{end} are not repeated here.

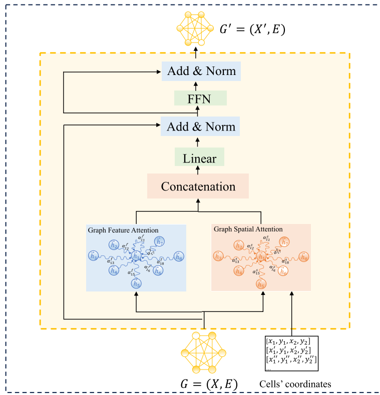


FIGURE 2. The architecture of DAFA proposed in the paper. DAFA builds a fully connected graph using cells as nodes. Subsequently, it leverages Graph Feature Attention and Graph Spatial Attention to aggregate features based on feature similarity and spatial similarity. Finally, integrating a structure akin to a Transformer encoder yields updated node representations.

$$u_i^j = \begin{cases} 0 & \text{if } (j \geq r) \\ 1 & \text{if } (j < r) \end{cases} \quad \text{subject to } j = 0, 1, 2, \ldots, R-2 \tag{13}$$

F. DESIGN OF LOSS FUNCTION

The DAFA-Net proposed in this paper mainly consists of two tasks: cell detection, with $\mathcal{L}_{\text{cell\_det}}$, and structure recognition, with $\mathcal{L}_{\text{table\_recog}}$. They are trained in an end-to-end way. The overall loss is as follows, where $\mathcal{L}_{\text{cell\_det}}$ for cell detection module refers to Eq. (1).

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{cell\_det}} + \mathcal{L}_{\text{table\_recog}} \\ \mathcal{L}_{\text{table\_recog}} &= \mathcal{L}_{r^{\text{start}}} + \mathcal{L}_{r^{\text{end}}} + \mathcal{L}_{c^{\text{start}}} + \mathcal{L}_{c^{\text{end}}} \end{aligned} \tag{14}$$

The recognition for cross-row or cross-column cells is one of the difficulties in table structure recognition (TSR) task. Inspired by focal loss [16], [42], we propose the focal loss for TSR called CRCC (Cross-Row-Cross-Column) loss. We use $\mathcal{L}_{r^{\text{start}}}$ as the example:

$$\mathcal{L}_{r^{\text{start}}}(\boldsymbol{X}, \Theta) = -\frac{1}{N} \sum_{i=1}^{N} \psi(\boldsymbol{x}_i, \Theta) \tag{15}$$

$$\psi(\boldsymbol{x}_i, \Theta) = \sum_{j=0}^{r-1} (1 - p_i^j)^{\gamma_i^{\text{span}}} \log(p_i^j)$$

TABLE 1. Performance of GAE-Net and TGRNet [16].

Methods	cell detection			table structure recognition				
CMDD								
	P	R	H	A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
TGRNet [16]	**0.998**	0.993	0.990	0.996	0.995	0.997	0.996	0.991
GAE-Net	0.997	**0.997**	**0.997**	**0.999**	**0.999**	**1.000**	**1.000**	**0.999**
ICDAR2013								
	P	R	H	A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
TGRNet [16]	0.682	0.652	0.667	0.445	0.445	0.700	0.692	0.275
GAE-Net	**0.897**	**0.890**	**0.893**	**0.546**	**0.544**	**0.723**	**0.720**	**0.390**
TableGraph-24K								
	P	R	H	A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
TGRNet [16]	**0.916**	0.895	**0.906**	0.917	0.916	0.919	**0.923**	0.832
GAE-Net	0.909	**0.899**	0.904	**0.952**	**0.951**	**0.923**	0.920	**0.876**
SciTSR								
	P	R	H	A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
TGRNet [16]	0.898	0.885	0.891	0.917	0.918	0.949	0.948	0.870
GAE-Net	**0.910**	**0.894**	**0.902**	**0.948**	**0.947**	**0.957**	**0.955**	**0.903**

$$+\sum_{j=r}^{R-2}(p_i^j)^{\gamma_i^{span}}\log(1-p_i^j) \tag{16}$$

$$\gamma_i^{span} = \min(2, -\lambda_i^{span}\log(1-\lambda_i^{span})+1) \tag{17}$$

where N is the total number of cells in the table, p_i^j is the probability for predicting $u_i^j = 1$ of the cell i. We set the number of rows crossed by the cell i to be s, and then the proportion of cells which cross s rows in the dataset is λ_i^{span}. Cross-row and cross-column cells account for a small proportion of all cells. By setting γ^{span} parameter, they will be highlighted in training. The ablation experiment results in Sec. IV-E show the effectiveness of our CRCC loss.

IV. EXPERIMENTS

A. DATASETS

We evaluate the DAFA-Net proposed in the paper on four datasets which are widely used for table structure recognition task, containing CMDD [43], ICDAR-2013 [44], TableGraph-24K [16], and SciTSR [9].

The CMDD dataset [43] consists of 357 images of medical experiment report tables, some of which come from scanners, while the others are taken by mobile phones. The dataset includes 476 tables, with 372 in the training set and 104 in the test set. The tables have a maximum of 24 rows and 5 columns. These MERT images fully reflect the complex borderless table structure recognition scenario in the medical field.

The ICDAR-2013 dataset [44] consists of a total of 150 tables: 75 of them are excerpts from 27 European documents, and the other 75 are excerpts from U.S. government documents, containing English content from a total of 238 pages in 76 PDF files.

The TableGraph-24K dataset [16] is collected from the TABLE2LATEX-450K dataset [27]. It contains 24,000 tables, with 20,000 for training, 2,000 for validation, and 2,000 for testing.

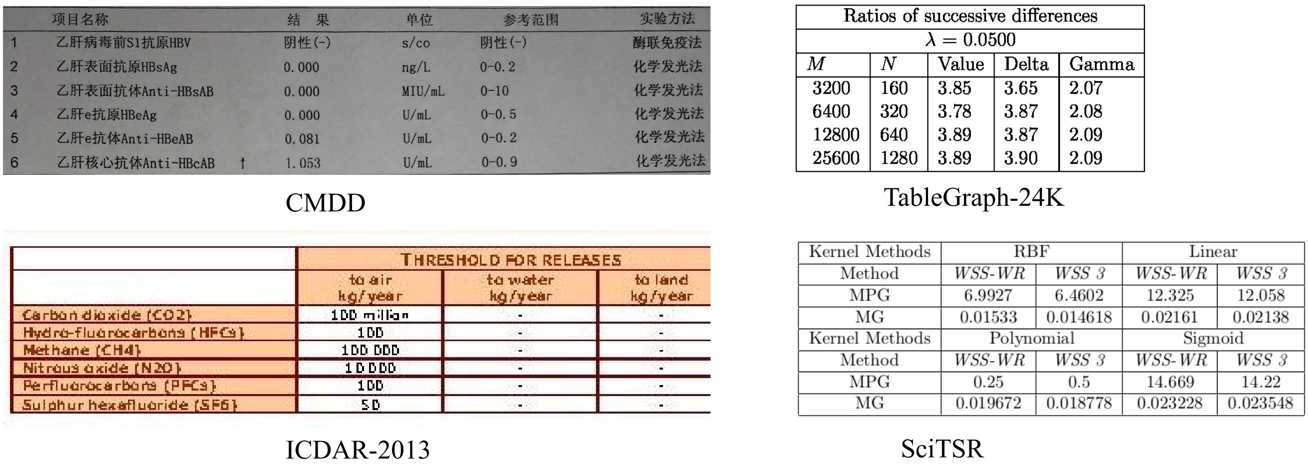
The SciTSR dataset [9] is a comprehensive dataset. The dataset consists of 15,000 PDF-formatted tables, images of the table regions, their corresponding structural labels, and bounding boxes for each cell. It is divided into 12,000 training instances and 3,000 testing instances. In addition, a complex table list known as SciTSR-COMP is also provided.

B. EVALUATION METRICS

We use the same metrics with the TGRNet [16]. For cell detection module, we use the $Precision(P)$, $Recall(R)$, and $Hmean(H)$ with the IoU threshold 0.5. For table structure recognition, we use the accuracy of four indices(row_{start}, row_{end}, col_{start}, col_{end}) based on the detected table cells, which are represented as A_{rowSt}, A_{rowEd}, A_{colSt}, A_{colEd}. We also use A_{all} for correctness evaluation of all four indices. Specially, A_{all} represents the ratio of correctly predicted cells to the total number of predicted cells, wherein a cell's prediction is considered correct if and only if the other four indices are predicted correctly.

C. EXPERIMENTAL SETTINGS

Our code is based on PyTorch, and all our experiments were carried out on GeForce RTX 3090 GPUs. We used the Adam optimizer with a learning rate set to 10^{-4} for all models. Before training the complete end-to-end DAFA-Net, we separately pre-trained the cell detection module and the table structure recognition module for 30 epochs each. Subsequently, we conducted a total of 50 epochs of overall training. To speed up the training process, we incorporated a pre-trained ResNet-50 model into the backbone module. It's also important to note that as part of our data preprocessing, each input image was resized to 480 × 480 pixels.



	项目名称	结　果	单位	参考范围	实验方法
1	乙肝病毒前S1抗原HBV	阴性(-)	s/co	阴性(-)	酶联免疫法
2	乙肝表面抗原HBsAg	0.000	ng/L	0-0.2	化学发光法
3	乙肝表面抗体Anti-HBsAB	0.000	MIU/mL	0-10	化学发光法
4	乙肝e抗原HBeAg	0.000	U/mL	0-0.5	化学发光法
5	乙肝e抗体Anti-HBeAB	0.081	U/mL	0-0.2	化学发光法
6	乙肝核心抗体Anti-HBcAB ↑	1.053	U/mL	0-0.9	化学发光法

CMDD

Ratios of successive differences				
$\lambda = 0.0500$				
M	N	Value	Delta	Gamma
3200	160	3.85	3.65	2.07
6400	320	3.78	3.87	2.08
12800	640	3.89	3.87	2.09
25600	1280	3.89	3.90	2.09

TableGraph-24K

	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Carbon dioxide (CO2)	100 million	-	-
Hydro-fluorocarbons (HFCs)	100	-	-
Methane (CH4)	100 000	-	-
Nitrous oxide (N2O)	10 000	-	-
Perfluorocarbons (PFCs)	100	-	-
Sulphur hexafluoride (SF6)	50	-	-

ICDAR-2013

Kernel Methods	RBF		Linear	
Method	*WSS-WR*	*WSS 3*	*WSS-WR*	*WSS 3*
MPG	6.9927	6.4602	12.325	12.058
MG	0.01533	0.014618	0.02161	0.02138
Kernel Methods	Polynomial		Sigmoid	
Method	*WSS-WR*	*WSS 3*	*WSS-WR*	*WSS 3*
MPG	0.25	0.5	14.669	14.22
MG	0.019672	0.018778	0.023228	0.023548

SciTSR

FIGURE 3. Sample images from the four datasets.

TABLE 2. Performance of table feature extractors on CMDD.

Exp.	extractor	the performance of table structure recognition				
		A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
1	GCN	0.134	0.134	0.221	0.221	0.030
2	GAT	0.876	0.873	0.601	0.601	0.536
3	DAFAs	**0.999**	**0.999**	**1.000**	**1.000**	**0.999**

TABLE 3. Performance of different loss on ICDAR2013.

Exp.	loss function	the performance of table structure recognition				
		A_{rowSt}	A_{rowEd}	A_{colSt}	A_{colEd}	A_{all}
1	TGRNet loss [16]	**0.444**	0.443	0.705	0.703	0.298
2	CRCC loss	**0.444**	**0.446**	**0.726**	**0.731**	**0.301**

D. PERFORMANCE EVALUATION

We compare our method DAFA-Net with the TGRNet [16] on CMDD, ICDAR2013, TableGraph-24K and SciTSR datasets. To our best knowledge, TGRNet [16] is an advanced method based on GCN for end-to-end table structure recognition. The results are shown in Table 1 and our DAFA-Net outperforms the strong baseline TGRNet [16] on most metrics, especially on A_{rowSt}, A_{rowEd} and A_{all}. The reason is that GCN relies too much on the structure of the graph when extracting cell features, but we use the attention mechanism to capture important context information from the data. Besides, DAFA-Net based on attention doesn't suffer from instability during training, and achieves better results than GCN-based method in the case of consistent data volume, which demonstrate that our method overcomes the problem of Transformer in the application of table structure recognition.

E. ABLATION STUDY

To evaluate the Dual-Awareness Feature Aggregator (DAFA) for table and the proposed CRCC loss, we conduct ablation studies on CMDD and ICDAR2013 datasets.

We use different models as table feature extractors for TSR: Graph Convolutional Network (GCN), Graph Attention Network(GAT), and the stacked DAFAs. As shown in Table 2, our structure DAFAs achieves satisfactory performance.

TGRNet [16] proposes a modified focal loss function for addressing the long-tailed distribution problem in TSR while we design the CRCC loss focusing on the difficulty of recognizing cross-row or cross-column cells. The results are shown in Table 3. Our CRCC loss achieves better scores, especially on the accuracy of column prediction. According to the statistics of cross-row and cross-column cells in ICDAR2013, we find that the cross-row cells only span 2 or 3 rows, whereas the cross-column cells span 2, 3, 4, 5, 6, 7 or 9 columns. It means the column prediction is more challenging than the row prediction, which is the reason why in Table 3 CRCC loss outperforms in column prediction.

V. CONCLUSION AND OUTLOOK

We studied the problem of how to recognize the complex table structure using attention mechanism. In this paper, we propose a Dual-Awareness Feature Aggregator (DAFA) as the table feature extractor. It makes full use of the feature correlation and position correlation to aggregate and extract the cell's feature in the table. Based on it, we propose the end-to-end solution for table structure recognition, called DAFA-Net. To address the difficulty of recognizing cross-row or cross-column cells in this task, we design a modified focal loss focusing on these special cells. It's worth mentioning that this challenge exists in the majority of table recognition scenarios, especially in medical examination reports. Extensive experiments on CMDD which is released from the Chinese medical laboratory reports, and three other public datasets show our method achieves competitive performance.

Regarding the future study, we may expand the scale of the table and reduce the complexity of the graph model [45], [46]. One important direction is to model the tables using graph structure, and accordingly it calls for effective ways of measuring the distance between graphs e.g. using the graph edit distance [47]. Also, it could be interesting to explore the retrieval of similar tables via graph matching techniques [48], ranging from the traditional non-learning approaches (including both pairwise matching [49] to joint matching of multiple graphs [50], [51] to the more recent learning-based models either by supervised training [52] or unsupervised training [53].

Another practical direction is to develop automatic neural architecture search approaches [54] to develop more tailored architecture for solving the table recognition problem, whereby different techniques such as sampling [55], zero-order decent search [56], operation re-merging [57] etc. have been well developed.

REFERENCES

[1] E. Koci, M. Thiele, O. Romero, and W. Lehner, "A machine learning approach for layout inference in spreadsheets," in *Proc. 8th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage.* Setúbal, Portugal: SciTePress, 2016, pp. 77–88.

[2] B. Yildiz, K. Kaiser, and S. Miksch, "pdf2table: A method to extract table information from pdf files," in *Proc. 2nd Indian Int. Conf. Artif. Intell.*, Jan. 2005, pp. 1773–1785.

[3] E. Oro and M. Ruffolo, "PDF-TREX: An approach for recognizing and extracting tables from PDF documents," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 906–910.

[4] Y. Liu, P. Mitra, C. L. Giles, and K. Bai, "Automatic extraction of table metadata from digital documents," in *Proc. 6th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2006, pp. 339–340.

[5] Y. Wang and J. Hu, "A machine learning based approach for table detection on the web," in *Proc. 11th Int. Conf. World Wide Web*, 2002, pp. 242–250.

[6] W. W. Cohen, M. Hurst, and L. S. Jensen, "A flexible learning system for wrapping tables and lists in HTML documents," in *Proc. 11th Int. Conf. World Wide Web*, May 2002, pp. 232–241.

[7] E. Oro and M. Ruffolo, "XONTO: An ontology-based system for semantic information extraction from PDF documents," in *Proc. 20th IEEE Int. Conf. Tools Artif. Intell.*, vol. 1, Nov. 2008, pp. 118–125.

[8] A. Costa e Silva, "New metrics for evaluating performance in document analysis tasks_application to the table case," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Sep. 2007, pp. 481–485.

[9] Z. Chi, H. Huang, H.-D. Xu, H. Yu, W. Yin, and X.-L. Mao, "Complicated table structure recognition," 2019, *arXiv:1908.04729*.

[10] W. Xue, Q. Li, and D. Tao, "ReS2TIM: Reconstruct syntactic structures from table images," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 749–755.

[11] H. Liu, X. Li, B. Liu, D. Jiang, Y. Liu, B. Ren, and R. Ji, "Show, read and reason: Table structure recognition with flexible context aggregator," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1084–1092.

[12] J. Ye, X. Qi, Y. He, Y. Chen, D. Gu, P. Gao, and R. Xiao, "PingAn-VCGroup's solution for ICDAR 2021 competition on scientific literature parsing task B: Table recognition to HTML," 2021, *arXiv:2105.01848*.

[13] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao, and X. Bai, "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107980.

[14] Y. Li, Z. Huang, J. Yan, Y. Zhou, F. Ye, and X. Liu, "GFTE: Graph-based financial table extraction," in *Proc. Pattern Recognit. ICPR Int. Workshops Challenges*. Cham, Switzerland: Springer, Jan. 2021, pp. 644–658.

[15] S. R. Qasim, H. Mahmood, and F. Shafait, "Rethinking table recognition using graph neural networks," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 142–147.

[16] W. Xue, B. Yu, W. Wang, D. Tao, and Q. Li, "TGRNet: A table graph reconstruction network for table structure recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1275–1284.

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[19] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[20] D. Wei, H. Lu, Y. Zhou, and K. Chen, "Image-based table cell detection: A novel table structure decomposition method with new dataset," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1–7.

[21] S. Raja, A. Mondal, and C. Jawahar, "Visual understanding of complex table structures from document images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2543–2552.

[22] Y. Li, Z. Wu, S. Zhao, X. Wu, Y. Kuang, Y. Yan, S. Ge, K. Wang, W. Fan, X. Chen, and Y. Wang, "PSENet: Psoriasis severity evaluation network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 800–807.

[23] L. Qiao, Z. Li, Z. Cheng, P. Zhang, S. Pu, Y. Niu, W. Ren, W. Tan, and F. Wu, "LGPMA: Complicated table structure recognition with local and global pyramid mask alignment," in *Proc. Int. Conf. document Anal. Recognit.* Cham, Switzerland: Springer, 2021, pp. 99–114.

[24] H. Liu, X. Li, B. Liu, D. Jiang, Y. Liu, and B. Ren, "Neural collaborative graph machines for table structure recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4523–4532.

[25] Z. Li, Y. Li, Q. Liang, P. Li, Z. Cheng, Y. Niu, S. Pu, and X. Li, "End-to-end compound table understanding with multi-modal modeling," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4112–4121.

[26] D. Lohani, A. Belaïd, and Y. Belaïd, "An invoice reading system using a graph convolutional network," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2019, pp. 144–158.

[27] Y. Deng, D. Rosenberg, and G. Mann, "Challenges in end-to-end neural scientific table recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 894–901.

[28] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, "TableBank: Table benchmark for image-based table detection and recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1918–1925.

[29] Y. Huang, N. Lu, D. Chen, Y. Li, Z. Xie, S. Zhu, L. Gao, and W. Peng, "Improving table structure recognition with visual-alignment sequential coordinate modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11134–11143.

[30] P. Kayal, M. Anand, H. Desai, and M. Singh, "Tables to LaTeX: Structure and content extraction from scientific tables," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 26, no. 2, pp. 121–130, Jun. 2023.

[31] X. Zhong, E. ShafieiBavani, and A. J. Yepes, "Image-based table recognition: Data, model, and evaluation," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2020, pp. 564–580.

[32] Z. Guo, Y. Yu, P. Lv, C. Zhang, H. Li, Z. Wang, K. Yao, J. Liu, and J. Wang, "TRUST: An accurate and end-to-end table structure recognizer using splitting-based transformers," 2022, *arXiv:2208.14687*.

[33] W. Lin, Z. Sun, C. Ma, M. Li, J. Wang, L. Sun, and Q. Huo, "TSRFormer: Table structure recognition with transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6473–6482.

[34] C. Ma, W. Lin, L. Sun, and Q. Huo, "Robust table detection and structure recognition from heterogeneous document images," *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 109006.

[35] C. Tensmeyer, V. I. Morariu, B. Price, S. Cohen, and T. Martinez, "Deep splitting and merging for table structure decomposition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 114–121.

[36] Z. Zhang, J. Zhang, J. Du, and F. Wang, "Split, embed and merge: An accurate table structure recognizer," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108565.

[37] J. Fernandes, B. Xiao, M. Simsek, B. Kantarci, S. Khan, and A. A. Alkheir, "TableStrRec: Framework for table structure recognition in data sheet images," *Int. J. Document Anal. Recognit. (IJDAR)*, Sep. 2023, pp. 1–19, doi: 10.1007/s10032-023-00453-8.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

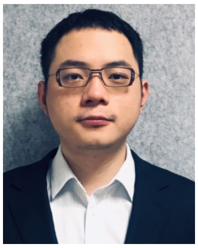
[40] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liõ, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[43] W. Xue, Q. Li, Z. Zhang, Y. Zhao, and H. Wang, "Table analysis and information extraction for medical laboratory reports," in *Proc. IEEE 16th Intl. Conf. Dependable, Autonomic Secure Comput., 16th Intl. Conf. Pervasive Intell. Comput., 4th Intl. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 193–199.
[44] M. Göbel, T. Hassan, E. Oro, and G. Orsi, "ICDAR 2013 table competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1449–1453.
[45] Q. Wu, W. Zhao, Z. Li, D. Wipf, and J. Yan, "NodeFormer: A scalable graph structure learning transformer for node classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–15.
[46] Q. Wu, C. Yang, W. Zhao, Y. He, D. Wipf, and J. Yan, "DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion," 2023, *arXiv:2301.09474*.
[47] R. Wang, T. Zhang, T. Yu, J. Yan, and X. Yang, "Combinatorial learning of graph edit distance via dynamic embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5237–5246.
[48] J. Yan, X.-C. Yin, W. Lin, C. Deng, H. Zha, and X. Yang, "A short survey of recent advances in graph matching," in *Proc. 2016 ACM Int. Conf. Multimedia Retr.*, 2016, pp. 167–174.
[49] J. Yan, C. Li, Y. Li, and G. Cao, "Adaptive discrete hypergraph matching," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 765–779, Feb. 2018.
[50] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1228–1242, Jun. 2016.
[51] Z. Jiang, T. Wang, and J. Yan, "Unifying offline and online multi-graph matching via finding shortest paths on supergraph," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3648–3663, Oct. 2021.
[52] R. Wang, J. Yan, and X. Yang, "Neural graph matching network: Learning Lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5261–5279, Sep. 2022.
[53] R. Wang, J. Yan, and X. Yang, "Unsupervised learning of graph matching with mixture of modes via discrepancy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10500–10518, Aug. 2023.
[54] X. Wang, Z. Lian, J. Lin, C. Xue, and J. Yan, "DIY your EasyNAS for vision: Convolution operation merging, map channel reducing, and search space to supernet conversion tooling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13974–13990, Nov. 2023.
[55] C. Xue, X. Wang, J. Yan, Y. Hu, X. Yang, and K. Sun, "Rethinking bi-level optimization in neural architecture search: A Gibbs sampling perspective," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 12, pp. 10551–10559.
[56] X. Wang, W. Guo, J. Su, X. Yang, and J. Yan, "ZARTS: On zero-order optimization for neural architecture search," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 12868–12880.
[57] X. Wang, C. Xue, J. Yan, X. Yang, Y. Hu, and K. Sun, "MergeNAS: Merge operations into one for differentiable architecture search," in *Proc. 29th Int. Conf. Int. Joint Artif. Intell.*, 2021, pp. 3065–3072.

XUANRUI HONG received the M.E. degree from Shanghai Jiao Tong University, in 2021, where she is currently pursuing the master's degree. Her research interests include deep learning and computer vision.

KAI ZHA received the B.E. degree from Tongji University and the M.E. degree in computer science and engineering from Shanghai Jiao Tong University, in 2021. He is currently a Senior Engineer with AutCAD. His research interests include deep learning and computer vision.



ZIMING FENG received the B.S. and M.S. degrees in computer science from Shanghai Jiao Tong University, in 2011 and 2014, respectively. He is currently a Senior Engineer with the Credit Card Center, China Merchant Bank. His main research interests include computer vision and machine learning.

ZHEXUAN CHEN received the bachelor's degree from Shanghai International Studies University, in 2012, and the master's degree from Kent State University, in 2013. In 2013, she joined the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), where she is currently an Assistant Engineer. Her research interests include software engineering and applied AI.

XIAODONG DU received the bachelor's and master's degrees from Beijing Foreign Studies University, in 2010 and 2013, respectively. In 2013, she joined the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), where she is currently an Assistant Engineer. Her research interests include software engineering and applied AI.

MIN LIU received the master's degree in traditional Chinese medicine and the Ph.D. degree in gynecology from the Nanjing University of Chinese Medicine, in 2010 and 2013, respectively. She is currently a Physician with the Department of Integrative Medicine, Obstetrics and Gynecology Hospital, Fudan University. Her research interests include the combination of traditional Chinese medicine and Western medicine for the treatment of infertility, menstrual disorders, low ovarian function, and chronic pelvic pain.