Neural Collaborative Graph Machines for Table Structure Recognition

Hao Liu[†*] Xin Li[*] Bing Liu Deqiang Jiang Yinsong Liu Bo Ren
Tencent YouTu Lab
{ivanhliu, fujikoli, billbliu, dqiangjiang, jasonysliu, timren}@tencent.com

Abstract

Recently, table structure recognition has achieved impressive progress with the help of deep graph models. Most of them exploit single visual cues of tabular elements or simply combine visual cues with other modalities via early fusion to reason their graph relationships. However, neither early fusion nor individually reasoning in terms of multiple modalities can be appropriate for all varieties of table structures with great diversity. Instead, different modalities are expected to collaborate with each other in different patterns for different table cases. In the community, the importance of intra-inter modality interactions for table structure reasoning is still unexplored. In this paper, we define it as heterogeneous table structure recognition (Hetero-TSR) problem. With the aim of filling this gap, we present a novel Neural Collaborative Graph Machines (NCGM) equipped with stacked collaborative blocks, which alternatively extracts intra-modality context and models inter-modality interactions in a hierarchical way. It can represent the intra-inter modality relationships of tabular elements more robustly, which significantly improves the recognition performance. We also show that the proposed NCGM can modulate collaborative pattern of different modalities conditioned on the context of intra-modality cues, which is vital for diversified table cases. Experimental results on benchmarks demonstrate our proposed NCGM achieves state-of-the-art performance and beats other contemporary methods by a large margin especially under challenging scenarios.

1. Introduction

Table structure recognition (TSR) aims to recognize the table internal structure to the machine readable data mainly presented in two formats: *logical structure* [18, 46] and *physical structure* [2, 13, 20, 22, 27, 30, 31, 34, 35, 40, 45]. More concretely, logical structure only focuses on whether two table elements belong to the same row, column or cells (*i.e.*, logical relationships), while the physical one contains

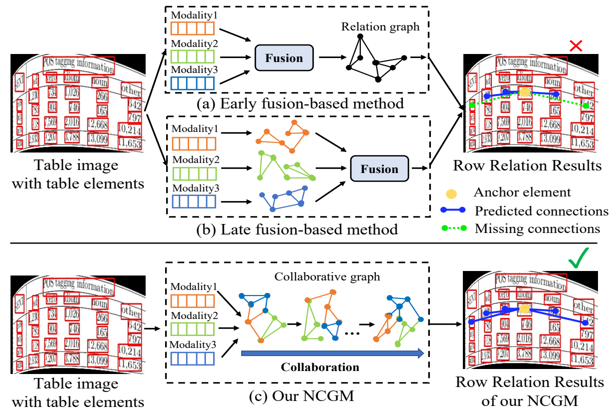


Figure 1. Illustration of motivation of the proposed NCGM. (a) Early fusion-based method. The multiple modalities of table elements are fused before modeling their relationships. (b) Late fusion-based method. The multiple modalities are modeled on their intra-modality relationships which are then fused for final results prediction. Due to lack of collaboration, for a distorted table case, previous methods cannot well extract the row relations (connected by blue lines) for an anchor element (yellow) with some true relation lost (green dotted line). (c) Our proposed NCGM. Different modalities are built into graphs with collaboration, which well accommodate the distorted table case.

not only logical relationships but also physical coordinates of cell boxes. The recognized tabular structure is essential to many downstream applications [12, 17]. Although many previous algorithms [2, 13, 18, 20, 22, 30, 31, 34, 35, 40, 45, 46] have achieved impressive progress in the community, TSR is still a challenging task due to two factors of complicated tables. The interior factor is complex table structure where spanning cell occupies at least two columns or rows, while exterior one is table distortion incurred by capture device.

Intuitively, *table elements* (text segment bounding boxes or table cells) commonly have inherent relationships and natural graph structure. Therefore, recent methods [2, 30, 34] attempt to attack the problem via constructing visual cues of table elements as graphs and applying the deep graph model, such as Graph Convolutional Networks (GCN) [15] to reason their relationships. To introduce richer table information, several methods [20, 30, 34] con-

[*]Equal contribution. [†]Contact person.

catenate the visual features with other modalities of features, such as geometry features, as a whole input to the graph model, as shown in Fig. 1 (a). Nevertheless, the relational inductive biases of different modalities would be highly discrepant, which makes naively early-fused modalities unable to deal with all table structures of great diversity. Besides, the intra-modality relationships would negatively affect each other when reasoning specific table structures. For example, the coordinates of table would dominate when recognizing a regular table, but they would become unreliable when processing distorted table cases. Instead, another alternative way is to individually model intra-modality relationships between table elements and combine them by a late-fusion strategy (Fig. 1 (b)). Unfortunately, the disentangled reasoning in terms of intra-modality interactions would introduce the curtailment of inter-modality interactions. This dilemma leads to the following question: *can different modalities collaborate with each other rather than interfering under different table scenarios?* We define this practical problem as heterogeneous table structure recognition (Hetero-TSR), which still lacks investigation.

In this work, we propose a novel Neural Collaborative Graph Machines (NCGM) tailored for this problem, as illustrated in Fig. 1 (c). Concretely, we adopt text segment bounding boxes as table elements in our method and extract their multi-modality feature embeddings from appearance, geometry and content dimensionality separately. To obtain the corresponding graph context and explore their interactions, we go beyond the standard attention model and propose a basic collaborative block with two successive modules, *i.e.,* Ego Context Extractor (ECE) and Cross Context Synthesizer (CCS). Among, ECE plays a role that dynamically generates graph context for the samples of each modality while the subsequent CCS is in charge of fusing and modulating inter-modality interactive information for different table cases. We stack this elemental block multiple times. Through this way, the intra-modality context generation and inter-modality collaboration can be conducted alternatively in a hierarchical way, which enables intra-inter modality interactions to be generated constantly from the low layer to the top one. In other words, the low-level contextual information in multiple modalities and the high-level one can collaborate with each other throughout the whole network, which is similar to the human perception process [1, 26]. The yielded collaborative graph embeddings enable our method to achieve better performance compared to other TSR methods, especially under more challenging scenarios, as clearly validated by extensive experimental results. To sum up, our contributions are in the four folds:

- We investigate the importance of collaboration between different modalities in TSR and propose the Hetero-TSR problem. To our best knowledge, we are the first to research the collaborative patterns between modality interaction for predicting table structure.
- We coin a novel NCGM tailored for Hetero-TSR problem, which consists of collaborative blocks alternatively conducting intra-modality context extraction and inter-modality collaboration in a hierarchical way.
- Experimental results on public benchmarks demonstrate that our method significantly outperforms the state-of-the-arts.
- We release a synthesizing method to augment existing benchmarks to more challenging ones. Under more challenging scenarios, our method can achieve at most 11% improvement than the second best method.

2. Related Work

2.1. Table Structure Recognition

Before the flourishing of deep learning, traditional table structure recognition methods rely on pre-defined rules and hand-crafted features [9–11, 14, 41]. With the development of deep learning, table structure recognition methods have recently advanced substantially on performance, which can be classified into three categories: boundary extraction-based [13, 22, 27, 35, 40], generative model-based [18, 46], and graph-based [2, 20, 30, 34] methods.

Boundary extraction-based methods. To extract cell boundaries, DeepDeSRT [35] and TableNet [27] are proposed by utilizing semantic segmentation. Besides, another technique [13] exploits bi-directional GRUs to establish row and column boundaries in a context driven manner. However, these methods are struggled when identifying cells spanning multiple rows and columns. SPLERGE [40] splits the table into grid elements in which adjacent ones are merged to restore spanning cells, whereas it still suffers from boundary ambiguity problem. To tackle this issue, the hierarchical GTE [45] leverages clustering algorithm for cell structure recognition. Cycle-CenterNet [22] exploits the cycle-pairing module to simultaneously detect and group tabular cells into structured tables, which focuses on the precision of cell boundary of the wired table in the wild. In the similar spirit, LGPMA [31] applies soft pyramid mask learning mechanism on both the local and global feature maps. Nevertheless, the subsequently heuristic structure recovery pipeline cannot achieve decent performance in complex scenarios.

Generative model-based methods. The method [18] utilizes the encoder-decoder framework, which generates an HTML tag sequence that represents the arrangement of rows and columns as well as the type of table cells. Moreover, another generative algorithm [46], termed EDD, consists of an encoder, a structure decoder and a cell decoder.

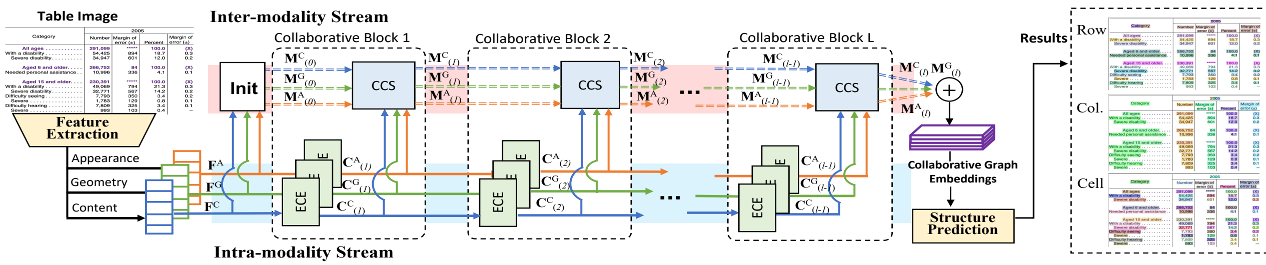


Figure 2. The architecture of our proposed method. Best viewed in color.

The encoder captures visual features of input table images, while the structure decoder reconstructs table structure and helps the cell decoder to recognize cell content.

Graph-based methods. GraphTSR [2] employs graph attention blocks to learn the vertex and edge representations in the latent space, and classifies edges as horizontal, vertical or unrelated. The method [30] introduces DGCNN to predict the relationship between words represented by the appearance and geometry features. Also based on DGCNN, TabStruct-Net [34] proposes an end-to-end network training cell detection and structure recognition networks in a joint manner. Besides, FLAG-Net [20] leverages the modulatable dense and sparse context of table elements. However, the above graph-based works are mostly designed for the interaction between table elements but lack the cues of the collaborative pattern of different modalities. In contrast to these works, our proposed NCGM leverages modality interaction to boost the multimodal representation for complex scenarios.

2.2. Transformer-based Multimodal Fusion

Transformer [42] architecture not only achieves significant performance gains in NLP community [5, 16, 21, 32, 39], but also gives birth to several pre-training methods [19, 23, 44] fusing various modalities for multimodal tasks.

Multiple embeddings fusion. VL-BERT [38] inheriting from BERT [5] introduces additional visual feature embeddings for visual-linguistic representations. LayoutLM [44] is a document understanding pre-trained model, which jointly models the interactions between text and layout information across scanned document images. However, the above algorithms simply take early-fused multiple embeddings as inputs, which may ignore the interactions between different modalities and result in discretization error and important details missing.

Co-attentional fusion. To better utilize visual-linguistic representations, ViLBERT [23] processes both visual and textual inputs in separate streams that interact through co-attentional transformer layers. Moreover, SelfDoc [19] establishes the contextualization over a block of content via cross-modal learning to manipulate visual features and textual features. Nevertheless, these previous co-attention based methods can only handle two modalities. By comparison, our proposed NCGM focuses on modality collaboration rather than simple fusion. Further, NCGM can not only process the interaction among more than two individual modalities, but also alternatively conduct intra-modality context extraction and inter-modality collaboration, which exploits more useful information provided by different modalities.

3. Methodology

3.1. Overall Architecture

The overview of the proposed Neural Collaborative Graph Machines (NCGM) is shown in Fig. 2. It mainly consists of collaborative blocks, which have two successive Multi-head Attention-based [42] modules, *i.e.*, Ego Context Extractor (ECE) and the Cross Context Synthesizer (CCS). First, three modalities of feature embeddings ($\mathbf{F}^{\sim} \in \{\mathbf{F}^{\mathrm{G}}, \mathbf{F}^{\mathrm{A}}, \mathbf{F}^{\mathrm{C}}\}$) in terms of table elements are extracted, *i.e., geometry*, *appearance* and *content embeddings*. In each collaborative block, the extracted feature embeddings are built as context graphs which are separately applied by the ECE to shape "intra-modality stream". Afterwards, the CCS selectively fuses individual contextual information from different modalities as inter-modality interactions maintained in "inter-modality stream". Note, we set $\mathbf{M}^{\sim}_{(0)} = \mathbf{F}^{\sim}$ as the initial input of CCS. The block is stacked L layers to implement the intra-inter modality collaboration in a hierarchical way. To predict the final table structure, the output collaborative graph embeddings from the l-th layer of inter-modality stream are sampled as pairs for cells, rows and columns classification.

3.2. Feature Extraction

In this component, a set of multi-modality features in terms of table elements are extracted from table image, including geometry embeddings $\mathbf{F}^{\mathrm{G}} \in \mathbb{R}^{N\times d}$, appearance embeddings $\mathbf{F}^{\mathrm{A}} \in \mathbb{R}^{N\times d}$ and content embeddings $\mathbf{F}^{\mathrm{C}} \in \mathbb{R}^{N\times d}$. N denotes the number of text segment bounding boxes. A more detailed description is given in supplementary material.

3.3. Collaborative Block

Ego Context Extractor. Now we elaborate on how to extract contextual interactions within each modality of table elements with the help of the Ego Context Extractor (ECE). Specifically, each extracted modality of features input to the ECE is constructed as individual directed graph $\mathbf{G}^{\sim} = \{\mathcal{V}, \mathcal{E}\} \in \left\{\mathbf{G}^{\mathrm{G}}, \mathbf{G}^{\mathrm{A}}, \mathbf{G}^{\mathrm{C}}\right\}$. In each decoupled modality of graph, corresponding embedding of each text segment bounding box is regarded as node $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\} \subseteq \mathcal{V}$ which is connected to each other by edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. In the similar spirit with works [30, 34], we adopt the following asymmetric edge function $h_{\Theta}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \| (\mathbf{x}_i - \mathbf{x}_j)$ to combine graph edge features to each node, which can be denoted as $\mathbf{H}_{\Theta}^{\sim} \in \mathbb{R}^{(N\cdot(N-1)/2)\times d}$. In the constructed graphs, each node can be either an anchor or one of context of others. In previous works using DGCNN [30, 34], only local context of each node is selected by k-Nearest Neighbors algorithm (KNN) to be aggregated into node feature. However, the local context is not versatile for representing relationships of all modalities. Besides, the DGCNN-based methods apply CNN to perform local context aggregation. For graph representation, CNN with strong inductive bias (*e.g.,* local behavior) may not be the optimal choice. To tackle the above problems, our proposed ECE instead aggregates information of fully-connected graph for all three modalities via Multi-head Attention (MHA) [42] module, which has been verified that it makes few assumptions about inputs and can learn to combine local behavior and global information based on input content [3].

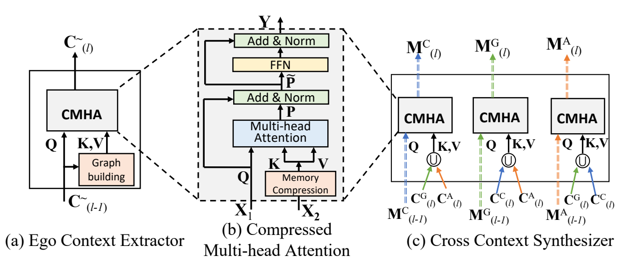


Figure 3. The proposed Ego Context Extractor and Cross Context Synthesizer modules in collaborative block. Best viewed in color.

More concretely, l-th ECE takes intra-modality features $\mathbf{C}^{\sim}_{(l\text{-}1)}$ as queries $\mathbf{Q}$ and the graph edge combined features $\mathbf{H}_{\Theta}^{\sim}$ as keys $\mathbf{K}$ and values $\mathbf{V}$ as illustrated in Fig. 3(a). Note, for the first layer, we input $\mathbf{F}^{\sim}$ as $\mathbf{C}^{\sim}_{(0)}$. However, the main limitation of using MHA is that the amount of input $\mathbf{K}$ and $\mathbf{V}$ can be very large ($N \cdot (N-1)/2$ in our case), which is infeasible to be trained. Given $\mathbf{Q} \in \mathbb{R}^{N\times d_q}, \mathbf{K} \in \mathbb{R}^{M\times d_k}, \mathbf{V} \in \mathbb{R}^{M\times d_v}$ and $M = N \cdot (N-1)/2$, the time complexity of the attention operation is $\mathcal{O}(NM)$ and the output is in $N \times d_v$ dimensionalities, of which the number is only relevant to that of $\mathbf{Q}$. Therefore, we can extend the MHA to a more memory-efficient Compressed MHA (CMHA) by introducing memory compression module which is utilized to reduce image pixel numbers in [43], as depicted in Fig. 3(b). In detail, the compression operation can be implemented as:

$$MC(\mathbf{H}) = Norm(Reshape(\mathbf{x}, \epsilon)\mathbf{W}^h), \tag{1}$$

where $Reshape(\mathbf{H}, \epsilon)$ denotes the operation of reshaping input $\mathbf{x} \in \mathbb{R}^{M\times d}$ to $\widetilde{\mathbf{x}} \in \mathbb{R}^{\epsilon M\times d/\epsilon}$, and $\epsilon \in [0, 1]$ is the compression ratio. Through this way, the complexity can be quadratically reduced from $\mathcal{O}(NM)$ to $\mathcal{O}(N\epsilon M)$. In default, we set $\epsilon = N/M$, where N is the number of queries $\mathbf{Q}$. And $Norm(\cdot)$ is the layer normalization. Additionally, we also equip the CMHA with residual connections in our method to make the query information flow unimpeded, which can be defined as:

$$\mathbf{Y} = Add\&Norm(FFN(\widetilde{\mathbf{P}}), \widetilde{\mathbf{P}}), \tag{2}$$

$$\widetilde{\mathbf{P}} = Add\&Norm(\mathbf{Q}, \mathbf{P}), \tag{3}$$

$$\mathbf{P} = MHA(\mathbf{Q}, MC(\mathbf{K}), MC(\mathbf{V})), \tag{4}$$

where "$FFN(\cdot)$" is the feed-forward layer and "$Add\&Norm(\cdot)$" denotes element-wise addition and layer normalization, which is similar to the work [42]. Conclusively, the contextual graph information is baked into graph node as $\mathbf{C}^{\sim} \in \left\{\mathbf{C}^{\mathrm{G}}, \mathbf{C}^{\mathrm{A}}, \mathbf{C}^{\mathrm{C}}\right\}$ within each modality through the CMHA in our ECE module.

Cross Context Synthesizer. Once heterogeneous context graph embeddings are obtained, our goals are to fuse them together in a collaborative way and to learn the collaborative patterns between different modalities. Also based on the CMHA, we design the Cross Context Synthesizer (CCS), as is shown in Fig. 3(c). In detail, the CCS has three parallel CMHA modules, and each of them takes one modality as queries while the other two are jointly regarded as keys and values. Take the first branch in Fig. 3(c) for example, the CMHA takes "content" modality of context graph embeddings as $\mathbf{Q}$, and the respective outputs of ECE for "geometry" and "appearance" are input as $\mathbf{K}$ and $\mathbf{V}$. In Fig. 3(c), "Ⓤ" denotes the union of two modality sets. For the similar purpose in ECE process, we also follow the similar rule to compress the number of "memory" to N which equals to that of $\mathbf{Q}$. Essentially, the query modality explores helpful information from another two modalities.

3.4. Table Structure Prediction

At the l-th layer of collaborative block, the outputs of CCS are to further fused as collaborative graph embeddings, which are denoted as $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N\} \in \mathbb{R}^{N\times d_e}$. Based on the embedings $\mathbf{E}$, our method constructs the i-th and j-th samples as pairs and concatenate them along channel axis as vectors $\mathbf{U} = \{\mathbf{u}_{1,1}, \mathbf{u}_{1,2}, ..., \mathbf{u}_{i,j}, ..., \mathbf{u}_{N,N}\} \in \mathbb{R}^{N^2\times 2d_e}$. Then three groups of FC layers are separately applied for predicting binary-class relations of $\mathbf{U}$, *i.e.,* whether the pair of i-th and j-th sample is belong to the same row, column or cell, as illustrated in Fig. 2. Each FC group consists of three FC layers with 256 dimensions and a 2-dimension FC with softmax layer.

3.5. Training Strategy

We train our proposed NCGM in an end-to-end way. The whole loss function is defined as $\mathcal{L} = \mathcal{L}_{cell} + \mathcal{L}_{col} + \mathcal{L}_{row}$, where $\mathcal{L}_{cell}$, $\mathcal{L}_{col}$ or $\mathcal{L}_{row}$ represents cell, column and row relationship losses. For each of them, we adopt the multi-task loss $\mathcal{L}_{\sim} = \lambda_1 \mathcal{L}_{class} + \lambda_2 \mathcal{L}_{con}$ to satisfy both the contrastive objective and to predict belonging classes of the output embedding pairs. $\mathcal{L}_{con}$ and $\mathcal{L}_{class}$ are contrastive loss and binary classification loss functions respectively. A more detailed description is given in supplementary material.

4. Experiments

4.1. Datasets and Evaluation Protocol

Datasets. We perform extensive experiments on various benchmark datasets. Among, ICDAR-2013 [8], ICDAR-2019 [6], WTW [22], UNLV [36], SciTSR [2] and SciTSR-COMP [2] are employed for physical structure recognition, while TableBank [18] and PubTabNet [46] are adopted for evaluating logical structure recognition performance. It should be noted that there is no training set in ICDAR-2013 and UNLV datasets, so we extend the two datasets to the partial versions, which is similar to TabStruct-Net [34]. A more detailed description about public benchmarks is given in supplementary material.

To further investigate the capacity of our proposed method under more challenging scenarios, we expand "SciTSR-COMP" dataset to "SciTSR-COMP-A" by applying two kinds of distortion algorithms. A more detailed description is given in supplementary material.

Evaluation settings. Several existing works are only applicable to table images alone, while others utilize additional information including text segment/cell bounding boxes or text contents. To compare in a unified protocol, we follow two different experimental setups in [34]: (a) Setup-A where only table image is taken as input without additional information and (b) Setup-B where table image along with additional features such as cell/text segment bounding boxes and text contents. For a fair comparison, we also incorporate the result boxes of detection in FLAG-Net [20] and the OCR results of Tesseract [37] as inputs in Setup-A.

Evaluation protocol. We employ precision, recall and F1-score [7] as protocol to evaluate the performance of our model for recognizing table physical structure including vertical and horizontal relations. For the recognition of table logical structure, BLEU score [28] used in [18] and Tree-Edit-Distance-based Similarity (TEDS) proposed in [46] are exploited.

4.2. Implementation Details

The framework is built on Pytorch [29]. We scale the input table images to a fixed size 512×512 to introduce scale invariance. In default, the layer number of collaborative blocks is set to 3 and the hidden size d is set to 64. Further, we set $h = 8, d_m = 64, d_k = d_v = 8$ for both Ego Context Extractor (ECE) and Cross Context Synthesizer (CCS) of each collaborative block. During training, the learning rate is initialized as $1\mathrm{e}{-4}$ and divided by 10 when the loss stops decreasing. For the training loss, we empirically set all weight parameters $\lambda_1 = \lambda_2 = 1$. For all experiments, the models are pre-trained on SciTSR for 10 epochs, and then fine-tuned on different benchmarks for 50 epochs, which is conducted on the platform with one Nvidia Tesla V100 GPU and 32 GB memory.

4.3. Comparison with State-of-the-arts

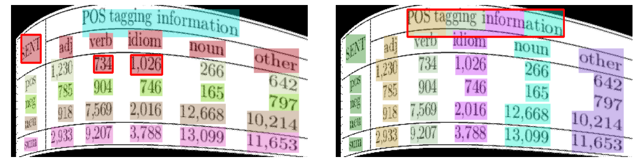
Results of physical structure recognition. As is shown in Tab. 1, our NCGM outperforms most of previous methods on different datasets for physical structure recognition. Compared with the strong baseline FLAG-Net [20], NCGM increases average F1-score on all datasets by round 2% under both Setup-A settings and Setup-B settings. When processing table images with complex distortions ("SciTSR-COMP-A"), it is worth mentioning that our NCGM can achieve about 11% and 12% higher F1-scores under Setup-A and Setup-B than the second-best FLAG-Net [20] without using distorted images as training data. If taking distorted data as training set, the performance of NCGM still can surpass it round 7% and 9% under both settings respectively. We also visualize row and column physical relationships of distorted table in Fig. 4. Note, the different color blocks in it merely visualize the belonging relationship rather than dividing the entire box. Taking right column of Fig. 4 for example, "POS tagging information" is one whole text segment bounding box. In logical, one can observe that "POS tagging information" box spans across five columns of word bounding boxes below it in column dimension. Therefore, the five columns attribute their respective colors to the "POS tagging information" box. By comparison, our method correctly recognizes both relationships while the FLAG-Net performs unsatisfactorily under distorted table scenes.

Results of logical structure recognition. In order to evaluate our model on logical structure recognition task benchmarks, *i.e.,* TableBank and PubTabNet, we perform lightweight post-processing (see supplementary material) on the NCGM's output results of row/column relationships to convert them to the HTML representation. Tab. 2 presents that our method achieves significant improvement compared with other methods for logical structure recognition task.

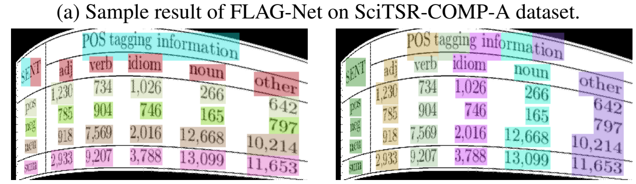
Computational complexity. A more detailed description is given in supplementary material.

Method	Train Dataset	Setup-A			Setup-B		
		P	R	F1	P	R	F1
ICDAR-2013-P							
DGCNN [30]	Sci. + IC13-P	-	-	-	98.6	99.0	98.8
TabStr. [34]	Sci. + IC13-P	93.0	90.8	91.9	99.1	99.3	99.2
GTE [45]	Pub. + IC13-P	94.4	92.7	93.5	-	-	-
LGPMA [31]	Sci. + IC13-P	96.7	99.1	97.9	-	-	-
C-CTRNet [22]	WTW + IC19	95.5	88.3	91.7	-	-	-
FLAG-Net [20]	Sci. + IC13-P	97.9	**99.3**	98.6	99.2	99.5	99.3
NCGM	Sci. + IC13-P	**98.4**	**99.3**	**98.8**	**99.3**	**99.9**	**99.6**
ICDAR-2019							
DGCNN [30]	Sci. + IC19	80.3	77.8	79.0	-	-	-
TabStr. [34]	Sci. + IC19	82.2	78.7	80.4	97.5	95.8	96.6
C-CTRNet [22]	WTW	-	-	80.8	-	-	-
FLAG-Net [20]	Sci. + IC19	**85.2**	83.8	84.5	96.1	96.3	96.2
NCGM	Sci. + IC19	84.6	**86.1**	**85.3**	**98.9**	**98.8**	**98.8**
WTW							
C-CTRNet [22]	WTW	93.3	91.5	92.4	-	-	-
FLAG-Net [20]	WTW	91.6	89.5	90.5	93.2	91.7	92.4
NCGM	WTW	**93.7**	**94.6**	**94.1**	**95.8**	**96.4**	**96.1**
UNLV-P							
DGCNN [30]	Sci. + UNLV-P	-	-	-	92.1	89.8	90.9
TabStr. [34]	Sci. + UNLV-P	84.9	82.8	83.9	99.2	99.4	99.3
FLAG-Net [20]	Sci. + UNLV-P	**89.2**	87.3	88.2	98.9	99.5	99.2
NCGM	Sci. + UNLV-P	88.9	**88.2**	**88.5**	**99.8**	**99.8**	**99.8**
SciTSR							
DGCNN [30]	Sci.	-	-	-	97.0	98.1	97.6
TabStr. [34]	Sci.	92.7	91.3	92.0	98.9	99.3	99.1
LGPMA [31]	Sci.	98.2	99.3	98.8	-	-	-
FLAG-Net [20]	Sci.	**99.7**	99.3	99.5	**99.8**	99.5	99.6
NCGM	Sci.	**99.7**	**99.6**	**99.6**	99.7	**99.8**	**99.7**
SciTSR-COMP							
DGCNN [30]	Sci.	-	-	-	96.3	97.4	96.9
TabStr. [34]	Sci.	90.9	88.2	89.5	98.1	98.7	98.4
LGPMA [31]	Sci.	97.3	98.7	98.0	-	-	-
FLAG-Net [20]	Sci.	98.4	98.6	98.5	98.6	99.0	98.8
NCGM	Sci.	**98.7**	**98.9**	**98.8**	**98.8**	**99.3**	**99.0**
SciTSR-COMP-A							
FLAG-Net [20]	Sci.	70.7	66.2	68.4	83.3	81.0	82.1
FLAG-Net [20]	Sci. + Sci.-C-A	82.5	83.0	82.7	88.8	87.5	88.1
NCGM	Sci.	**79.6**	**78.9**	**79.2**	**93.3**	**94.8**	**94.0**
NCGM	Sci. + Sci.-C-A	**88.4**	**90.7**	**89.5**	**97.2**	**97.5**	**97.3**

Table 1. Comparison results of physical structure recognition on ICDAR-2013-P, ICDAR-2019, WTW, UNLV-P, SciTSR, SciTSR-COMP and SciTSR-COMP-A dataset. "-P" means partial dataset and "-A" represents augmented dataset by distortion. "P", "R" and "F1" stand for "Precision", "Recall" and "F1-score" respectively. "TabStr." and "C-CTRNet" denote "TabStruct-Net" and "Cycle-CenterNet" individually.



(a) Sample result of FLAG-Net on SciTSR-COMP-A dataset.



(b) Sample result of NCGM on SciTSR-COMP-A dataset.

Figure 4. Visualization of physical relationships of distorted table between FLAG-Net and NCGM. The first and second column indicate the predictions of rows and columns respectively. The boxes belonging to the same relationships are filled in the same colors. The boundaries of the text segment boxes with misrecognized relationships are marked in red lines. Our NCGM shows better tolerance for the challenging scenarios compared with FLAG-Net.

Method	Train Dataset	Setup-A
		BLEU
TableBank		
Image-to-Text [18]	TableBank	73.8
TabStruct-Net [34]	SciTSR	91.6
FLAG-Net [20]	SciTSR	93.9
NCGM	SciTSR	**94.6**

Method	Train Dataset	Setup-A
		TEDS
PubTabNet		
EDD [46]	PubTabNet	88.3
TabStruct-Net [34]	SciTSR	90.1
GTE [45]	PubTabNet	93.0
LGPMA [31]	PubTabNet	94.6
FLAG-Net [20]	SciTSR	95.1
NCGM	SciTSR	**95.4**

Table 2. Comparison results of logical structure recognition on TableBank and PubTabNet datasets.

4.4. Ablation Study

In this subsection, we perform several analytic experiments under Setup-B settings on SciTSR-COMP benchmark to investigate the contributions of intra-modality and inter-modality interactions in our proposed NCGM.

Effect of intra-modality interactions. For intra-modality interactions, Tab. 3 compares the effectiveness of various extractors, including DGCNN [30] and Transformer [42], with ECE in our method. "Mixed" means all modality features are **early-fused** by concatenation as the input and "Individual" denotes each modality is input into context extractor separately. Tab. 3 shows ECE can achieve the best performance when taking either mixed features or individual features as input while "Transformer" performs the worst. For "DGCNN", it only aggregates information from top K similar nodes of each node instead of all ones. Compared with "DGCNN", although "Transformer" can deploy the global information of nodes, it ignores the directed edge effects between nodes. Encouragingly, our CMHA-based ECE can not only consider the directed relationships between nodes, but also extract the context information from all nodes. Additionally, we can also observe that individual features can yield better results than the mixed ones, which proves that decoupling the individual modality from each other is indeed a more preferable way to solve the Hetero-TSR problem.

Fusion Method	Input		Intra.			Inter.		Setup-B		
	Mix.	Ind.	DG.	Tr.	ECE	Con.	CCS	P	R	F1
Early Fusion	✓	✗	✓	✗	✗	✗	✗	96.3	97.4	96.8
	✓	✗	✗	✓	✗	✗	✗	95.1	95.6	95.3
	✓	✗	✗	✗	✓	✗	✗	97.8	98.3	98.0
Late Fusion	✗	✓	✓	✗	✗	✓	✗	96.9	98.2	97.5
	✗	✓	✗	✓	✗	✓	✗	94.9	96.1	95.5
	✗	✓	✗	✗	✓	✓	✗	98.4	98.2	98.3
NCGM	✗	✓	✗	✗	✓	✗	✓	**98.8**	**99.3**	**99.0**

Table 3. Ablation studies of NCGM on SciTSR-COMP dataset. "Intra." and "Inter." stand for intra-modality interactions and inter-modality interactions respectively. "Mix." and "Ind." are short for "Mixed" and "Individual". "DG." and "Tr." denote "DGCNN" and "Transformer". "Con." represents "Concatenation".

Effect of inter-modality interactions. We compare the proposed CCS with the "Concatenation" operation of multi-modal features in Tab. 3. It can be observed that CCS improves the accuracy of predicting adjacency relationship compared with directly **late-fused** multiple model features via concatenation. This confirms the benefits of CCS that enables one modality to positively collaborate with the others, and can capture the complex implicit modality relationships. Moreover, it also proves that the CCS module combined with ECE can further boost the performance.

4.5. Further Analysis on Collaborative Block

What does ECE learn from the intra-modality? As suggested by recent works [24, 25, 33] on interpreting attention mechanism, separate attention heads may learn to look for various relationships between inputs and introducing more sparsity and diversity for attention may improve performance and interpretability. To explore the intra-modality interactions learned by ECE in collaborative block, we in Fig. 6 visualize the multi-head attention maps from last blocks of ECE. For comparison, we also visualize the multi-head self-attention maps from the last blocks of "Transformer-Mixed" [42] and KNN (K = 5) selection heatmaps of all layers in DGCNN [30], where a lighter color indicates a closer relationship. The KNN results of DGCNN show that the feature aggregation of one node only pays attention to the top K similar features of other nodes instead of all the nodes, and relies on the choice of K. The attention maps of Transformer-Mixed present equilibrium status, which lacks sparsity and diversity. Comparatively, our "ECE-Mixed" taking mixed features presents more diversified attention maps in eight heads, which indicates ECE can more effectively capture context information. Moreover, the attention maps generated by "ECE-Individual" show different intriguing focus patterns for different features. Specifically, ECE prefers to extract interactions for appearance and geometry features in global scope while content features bring more local focus patterns.

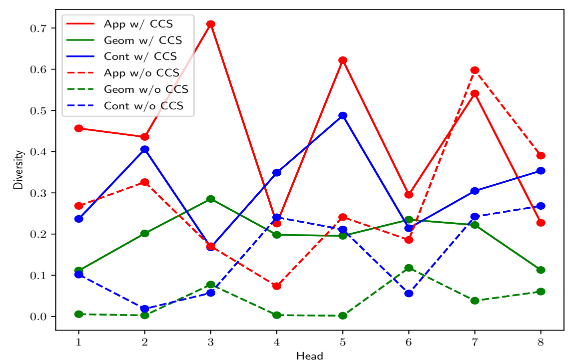


Figure 5. Diversities of attention maps for different modalities with or without CCS. Solid lines (∼ w/ CCS) represent the diversity distributions of attention in CCS when one modality features are regarded as queries and others as keys/values. Dashed lines (∼ w/o CCS) present diversity of attention weights in ECE for each modality.

How do different modalities collaborate with each other? To investigate the working pattern of CCS, we adopt Jensen-Shannon Divergence [4] (see supplementary materials) to measure the average diversity of attention map in CCS when the model also takes input table image shown in Fig. 6. As shown in Fig. 5, solid lines (∼ w/ CCS) represent the diversity distributions when one modality features are regarded as queries and others as keys/values. After removing CCS, diversity of attention weights in ECE for each modality is also presented by dashed lines (∼ w/o CCS). For those with CCS, the higher value indicates the query modality is in a closer collaboration with the other modalities. Particularly, appearance modality has the strongest collaborative relationship with others while geometric one requires the least collaboration. By comparison, the diversities of attention weights in ECE also follow a similar trend, but with lower values on average.

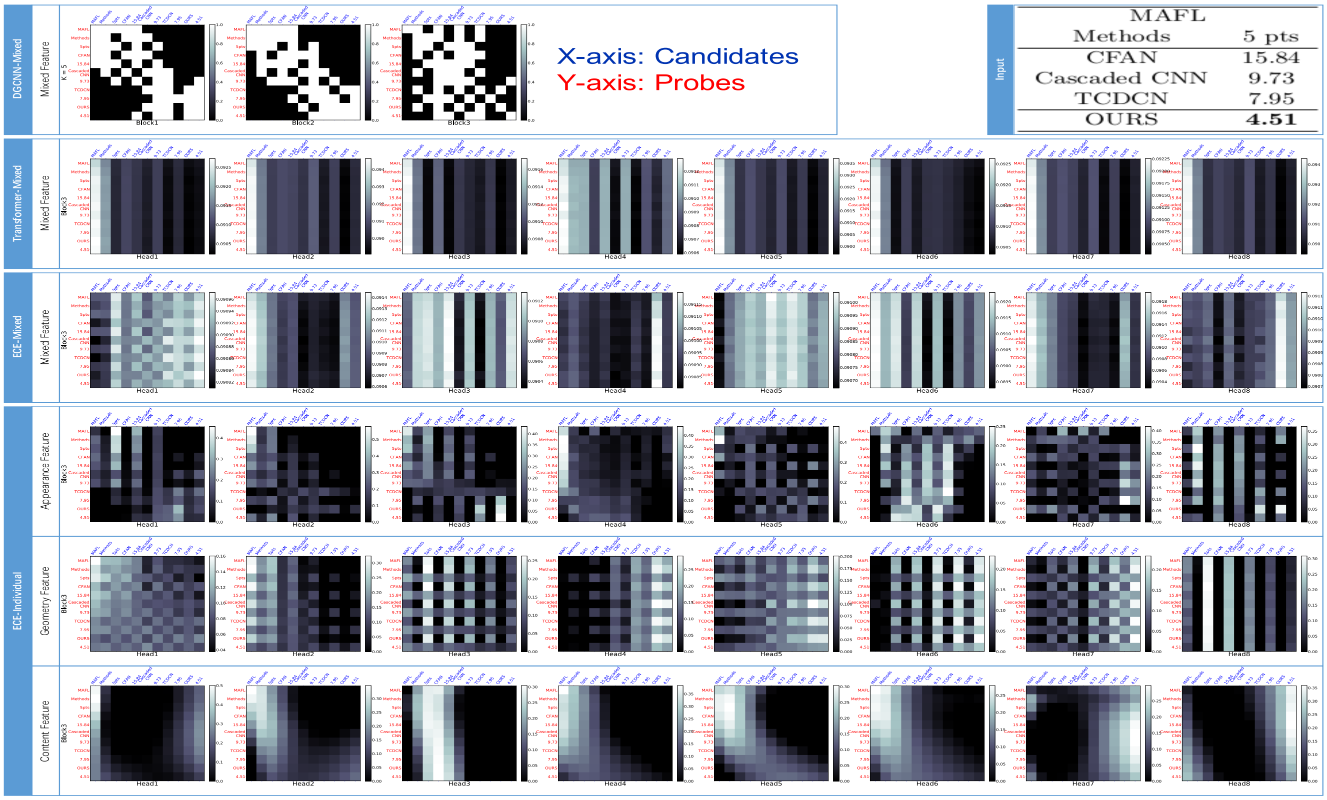


Figure 6. Visualization of the heat-maps generated by DGCNN and multi-head attention maps from the Transformer and ECE. Y-axis (red) and X-axis (blue) are "probes" and "candidates" respectively. For ECE, probes are graph node features and candidates are edge combined features. For Transformer and DGCNN, probes and candidates are both non-graph features. The heat-maps of DGCNN show a local hard selection way in terms of context. And Transformer yields attention maps lacking sparsity and diversity. In contrast, ECE-Mixed presents more diversified attention maps and ECE-Individual extracts interactions in global or local pattern conditioned on different features. Best viewed in color.

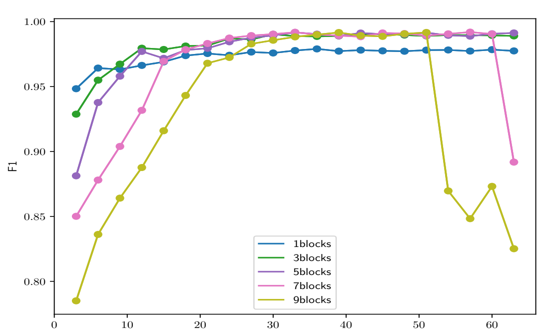


Figure 7. The relationship between block number of NCGM and F1-score on SciTSR-COMP dataset.

The more collaborative blocks, the better performance? To further explore the effect of the collaborative block number on the NCGM performance, we conduct a set of experiments setting block numbers from 1 to 9, respectively. It can be seen from Fig. 7 that it is a trade-off problem. Small block number can render faster convergence to the model. As the number increases, the performance keeps improving until block number increases to 5, but the convergence speed of the network keeps slowing down. In particular, we observe that the F1-score decreases sharply when NCGM with more than 7 blocks is trained over round 50 epochs, which indicates more blocks are easier to cause model training collapse problem. Based on the above observation, we set it to 3 as default number.

5. Conclusion and Limitation

We present a novel graph-based method for heterogeneous table structure recognition through learning intra-inter modality collaboration. Extensive experiments on public benchmarks demonstrate its superiority over state-of-the-art methods, especially under challenge scenarios. However, there exist two limitations could be improved in future. The first one is the inevitable problem of computational complexity increase brought by introducing multiple modalities and decoupled processing. The second one lies in the fact that NCGM with deeper blocks is easier to suffer from the training collapse problem. We may introduce more inductive bias into the attention model to tackle it.

References

[1] John R Anderson. *Cognitive psychology and its implications*. Macmillan, 2005. 2

[2] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 1, 2, 3, 5

[3] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *Eighth International Conference on Learning Representations-ICLR 2020*, number CONF, 2020. 4

[4] Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019. 7

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[6] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1510–1515. IEEE, 2019. 5

[7] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. A methodology for evaluating algorithms for table understanding in pdf documents. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 45–48, 2012. 5

[8] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1449–1453. IEEE, 2013. 5

[9] E Green and M Krishnamoorthy. Recognition of tables using table grammars. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 261–278, 1995. 2

[10] Yuki Hirayama. A method for table structure analysis using dp matching. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 583–586. IEEE, 1995. 2

[11] Katsuhiko Itonori. Table structure recognition based on textblock arrangement and ruled line position. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 765–768. IEEE, 1993. 2

[12] Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–483, 2016. 1

[13] Saqib Ali Khan, Syed Muhammad Daniyal Khalid, Muhammad Ali Shahzad, and Faisal Shafait. Table structure extraction with bi-directional gated recurrent unit networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1366–1371. IEEE, 2019. 1, 2

[14] Thomas Kieninger and Andreas Dengel. The t-recs table recognition and analysis system. In *International Workshop on Document Analysis Systems*, pages 255–270. Springer, 1998. 2

[15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 3

[17] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2016. 1

[18] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1918–1925, 2020. 1, 2, 5, 6

[19] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 3

[20] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, Bo Ren, and Rongrong Ji. Show, read and reason: Table structure recognition with flexible context aggregator. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1084–1092, 2021. 1, 2, 3, 5, 6

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[22] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 944–952, 2021. 1, 2, 5, 6

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 3

[24] Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–376, 2018. 7

[25] David Mareček and Rudolf Rosa. Extracting syntactic trees from transformer encoder self-attentions. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 347–349, 2018. 7

[26] Bruno A Olshausen, Charles H Anderson, and David C Van Essen. A neurobiological model of visual attention and

invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993. 2

[27] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 128–133. IEEE, 2019. 1, 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 5

[30] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. Rethinking table recognition using graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 142–147. IEEE, 2019. 1, 2, 3, 4, 6, 7

[31] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. Lgpma: Complicated table structure recognition with local and global pyramid mask alignment. *arXiv preprint arXiv:2105.06224*, 2021. 1, 2, 6

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 3

[33] Alessandro Raganato, Jörg Tiedemann, et al. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics, 2018. 7

[34] Sachin Raja, Ajoy Mondal, and CV Jawahar. Table structure recognition using top-down and bottom-up cues. In *European Conference on Computer Vision*, pages 70–86. Springer, 2020. 1, 2, 3, 4, 5, 6

[35] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1162–1167. IEEE, 2017. 1, 2

[36] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 113–120, 2010. 5

[37] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007. 5

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[39] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 3

[40] Chris Tensmeyer, Vlad I Morariu, Brian Price, Scott Cohen, and Tony Martinez. Deep splitting and merging for table structure decomposition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 114–121. IEEE, 2019. 1, 2

[41] Scott Tupaj, Zhongwen Shi, C Hwa Chang, and Hassan Alam. Extracting tabular information from text files. *EECS Department, Tufts University, Medford, USA*, 1996. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017. 3, 4, 6, 7

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 4

[44] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. 3

[45] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 697–706, 2021. 1, 2, 6

[46] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. *arXiv preprint arXiv:1911.10683*, 2019. 1, 2, 5, 6