Scene Table Structure Recognition with Segmentation and Key Point Collaboration

Zhuoming Li[1]*, Fan Peng[1]*, Yang Xue[1](✉), Ni Hao[2], and Lianwen Jin[1]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
{202121014036,202020112442,yxue,eelwjin}@mail.scut.edu.cn
[2] Dept of Mathematics University College London h.ni@ucl.ac.uk

Abstract. This paper proposes a Segmentation and Key point Collaboration Network (SKCN) for structure recognition of complex tables with geometric deformations. First, we combine the cell regions of the segmentation branch and the corner locations of the key point regression branch in the SKCN to obtain more reliable detection bounding box candidates. Then, we propose a Centroid Filtering-based Non-Maximum Suppression algorithm (CF-NMS) to deal with the problem of overlapping detected bounding boxes. After obtaining the bounding boxes of all cells, we propose a post-processing method to predict the logical relationships of cells to finally recover the structure of the table. In addition, we design a module for online generation of tabular data by applying color, shading and geometric transformation to enrich the sample diversity of the existing natural scene table datasets. Experimental results show that our method achieves state-of-the-art performance on two public benchmarks, TAL_OCR_TABLE and WTW.

Keywords: table structure recognition · segmentation and key point collaboration · centroid filtering NMS · online generation of tabular data

1 Introduction

Table is widely used as an effective representation of structured data in various types of documents in daily life. With the rise digitalization, table recognition has become an important research topic in the field of document understanding. How to correctly recognize the structure of a table is an important step in table recognition, whose main task is to identify the internal structure of a table. It aims to locate all the physical position of cells in the table and obtain information about the rows and columns in order to better understand the table as a whole. However, it's a challenging task for natural scene tables which can be complex in structure, vary in style and content, and may cause geometric distortions or even

*Authors contributed equally as first author

bending during the image acquisition process. With the explosive growth in the number of documents, applying table detection and table recognition techniques to reconstruct tables from document images has become one of the important techniques in current document understanding systems that can facilitate many downstream tasks and has significant research value.

Early table recognition studies mainly focused on hand-crafted features and heuristic rules [7–9]. Most of them were applied to simple table structures or specific data formats, such as PDF. Recently, research scholars have proposed more general models for structure recognition, such as LGPMA [22] and FlagNet [14]. The advantage is only one model is needed for all types of wired and wireless tables in document and natural scenes. However, these models are generally complex and the feature that can be utilized is the intersection of features extracted from different types of tables, thus ignoring the unique feature of each type of table. In real life, table recognition tasks are usually applied to fixed scenes, which require more targeted table recognition models. For example, for distorted wired table recognition in natural scenes, in order to obtain accurate cell boundaries, it is necessary to take full advantage of the most obvious visual features of the cells, i.e., the box lines and four corner points of each cell; whereas generic models, in order to be applicable to both wired and wireless table recognition, often do not take full advantage of these most salient visual features. Therefore, it also makes sense to design a specialized table recognition model to take full advantage of the salient features of each type of table.

Cycle-CenterNet [17] proposed a detection-based table structure recognition method that works well for wired table recognition in seven sub-scenarios. It first locates the four corner points of each cell and further infers the overall logical structure of the table from the coordinates of the cell. However, it only utilizes the corner point features of the wired tables and ignores the box line features of the tables. For table recognition of complex natural scenes with challenges such as geometric distortion, overlay, occlusion and blurring, it is inadequate to completely describe the overall position information of a cell by only four corner points. Better results can be achieved if a scheme can be proposed to extract both corner point features and box line features of wired tables.

Based on this, we propose a Segmentation and Key point Collaboration Network (SKCN) that combines the cell region of the segmentation branch and the corner locations of the key point regression branch to obtain a more reliable detection bounding box for better recognition performance. On the one hand, these two branches can assist each other during training. On the other hand, their respective results can interact and fuse to obtain refined detection results. In order to effectively filter redundant detected bounding boxes, we propose a centroid filtering algorithm based on the standard NMS algorithm, which achieves accurate cell detection results. Base on the refined cell boxes, we design a post-processing scheme to predict the logical relationship of the cells to recover the structure of the table.

The main contributions of this paper are as follows:

1. We combine the cell region of the segmentation branch and the corner locations of the key point regression branch to obtain a refined detected bounding box.

2. We propose a Centroid Filtering-based Non-Maximum Suppression algorithm (CF-NMS). To address the challenge of overlapping bounding boxes in table recognition tasks, we use CF-NMS to filter out prediction results with high IOU values that overlap with the target cell, thus improving the model detection performance.

3. We propose a module to generate tabular data online by applying color, shading and geometric transformation to enrich the sample diversity of existing natural scene table datasets.

2 Related Work

Early methods for table structure recognition [7–9, 24, 26] were mainly based on well-designed handcrafted features and heuristic rules. Most of these methods were applied to specific data formats, such as PDF files. However, in these traditional methods, there are strong assumptions about the layout of the tables, which limits their generality. With the rapid development of deep neural networks, image-based table structure recognition methods have shown great potential and outperform traditional methods by a large margin. We roughly divide these methods into four categories: image-to-token generation method, graph-based method, segmentation-based method, and object detection-based method.

2.1 Image-to-token generation method

This method treats table structure recognition as an image-to-token generation problem, typically using an encoder-decoder structure that directly converts the source table image into target token to adequately describe tabular data structure and its cell content. Existing approaches have tried several attempts to convert table images into symbols or HTML sequences [3, 11, 30, 33]. However, these methods usually rely on a large amount of data to train for convergence. In some cases, especially with large and complex tables, this approach may lead to performance degradation. Due to the limited length of the sequences ,these methods usually adopt certain trade-off strategies for large tables and have difficulty in tuning parameter and network design with their weep explanatory.

2.2 Graph-based method

The graph-based approach [21, 23] treats the bounding boxes of cell regions or text regions as nodes in a graph and uses graph neural networks to predict the logical relationship of each sampled node pair. GraphTSR [1] introduces the attention module to predict whether the sampled node pair belong to same row or same column. FLAG-Net [14] combines Transformer with graph-based context

aggregator in an adaptive way to exploit the advantages of both. NCGM [13] leverages graphs and modality interaction to enhance the multi-modal representation of text embeddings. However, these methods rely on bounding boxes of cell regions or text regions used as additional input, which are not available directly from the table images, thus bringing extra network cost.

2.3 Segmentation-based method

The segmentation-based approach first obtains the segmentation results from the table image and then parses the segmentation results to reconstruct the table structure. There are two broad types of this approach. One is to first obtain the segmentation of the rows and columns, and then use the segmentation results to grid out the cell boundaries. DeepdeSRT [25] and TableNet [19] semantically segment rows and columns, and intersect the segmentation results of rows and columns to obtain cell segmentation. To deal with spanning cells, SPLERGE [28] uses the split model to segment cell boundaries and then uses the merge model to further merge adjacent cells to obtain spanning cell boxes. SEM [31] follows the idea of multimodality and introduces textual feature to fuse with visual feature for each cell. The other is to recover cell boundaries to obtain cell boxes directly. CascadeTabNet [20] classifies tables into bordered and borderless tables, then predicts cell segmentation for borderless tables and extracts cells from bordered tables using traditional algorithms. LGPMA [22] combines local and global feature to accurately reconstruct cell boundaries by using soft pyramidal masks. However, these methods cannot handle distorted tables because they rely on table-axis alignment.

2.4 Object detection-based method

The method based on object detection first obtains the basic cells of a table from a table image by directly detecting the bounding box of a cell or text. Heuristic rules are then used to predict the logical relationships between detected cells to further reconstruct the logical structure of the table. [23, 27, 32] propose to detect the bounding boxes of table cells directly. After obtaining the bounding boxes of cells, [23, 32]designed some rules for clustering cells into rows and columns. However, the methods mentioned above assume that the table is well aligned and the target bounding boxes are rectangular, which are not suitable for natural scene tables. Cycle-CenterNet [17] introduces a cyclic pairing module to predict quadrilateral bounding boxes. Our method also uses quadrilateral bounding boxes for detection, which are more adaptable to the complexity of natural scene tables and achieve better performance in experiments. However, quadrilateral bounding boxes are still difficult to accurately describe curved cells and also may bring the potential of degrading detection performance. Sequential-free box discretization (SBD) [16] parameterizes bounding boxes as key edges and predicts the coordinates of four key points of the box from which the box is subsequently recovered. It can output more qualitative and accurate results

in natural scene table recognition. Therefore, we use SBD to predict four corner points in our method. The model is built based on the box discretization network [16], which use SBD as an additional branch to Mask-RCNN [6].

3 Methodology

Our approach consists of two main components: the Segmentation and Key point Collaboration Network (SKCN) and the Centroid Filtering-based Non-Maximum Suppression module(CF-NMS). The former is to obtain cell regions from both the segmentation branch and the key point regression branch to generate refined bounding boxes. The latter deals with the problem of overlapping cell boxes under the natural scene table. After obtaining bounding boxes, we use a post-processing algorithm to cluster cells into rows and columns and then parse the table structure. The details of our approach are described separately in the following sections.

3.1 SKCN

As shown in Fig. 1, the input image is first transformed into the output of four branches, i.e., box classification, box regression, box segmentation and point regression. The box regression branch outputs the minimum area bounding rectangle about the cell. The box classification branch predicts the category of the cell, such as images, text, formulas and other categories. Among them, the box segmentation branch and the point regression branch play an important role. The box segmentation branch focuses on the box-line characteristics of the wired table to get the segmentation result of the cell region wrapped by boundaries, which better adapts to the arbitrary deformation of the cell. The point regression branch mainly locates the four key points by using the key-point characteristics of the cell. The advantage of our model is to fully capture the feature of table elements to achieve a more accurate detection.

Since box segmentation and point regression serve for the same task of cell detection, previous studies usually selected only one of the two in this case. However, we believe that each of these two branches has its own characteristics. The box segmentation branch outputs pixel-level instance segmentation of cells, so the predicted box will be closer to ground truth. However, when it comes to complex tables with geometrical distortions or incomplete linear characteristic, it is difficult to separate out closely adjacent cell instances, which can easily lead to missed detection. The point regression branch only needs to return the four key points of the target. We first predict the eight boundary key edges of the cell in the process of locating the key points, and then combine them into four key points, which makes it easier to learn. The SBD branch tends to predict cells more completely, but the drawback is also obvious. If a predicted error occurs at one of the four key points, the detected bounding box becomes imprecise. Based on this, we propose to fuse results by a proper process to achieve better performance. We give these two branches different priorities in different

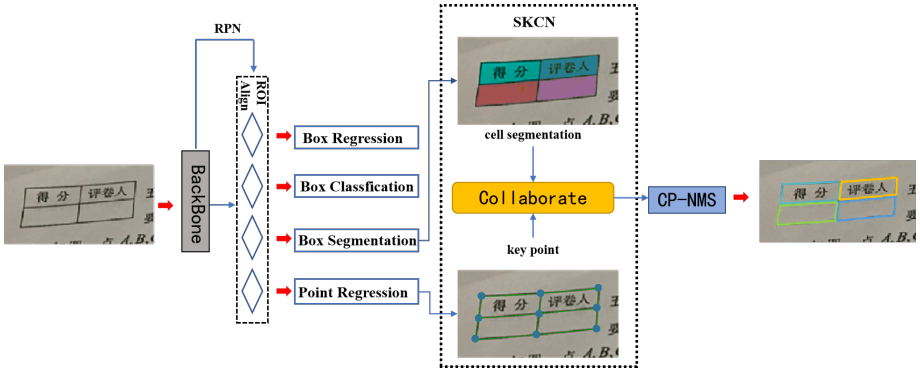


Fig. 1. The architecture of SKCN

confidence ranges. Firstly, we add a small constant of 0.03 to the segmentation results with confidence higher than 0.9. In this way, the high confidence segmentation results are preferred by the Non-Maximum Suppression. Then we select the results of SBD with confidence higher than 0.2 and mix the results of both branches together into the CF-NMS module to obtain the final results.

3.2 Key point prediction

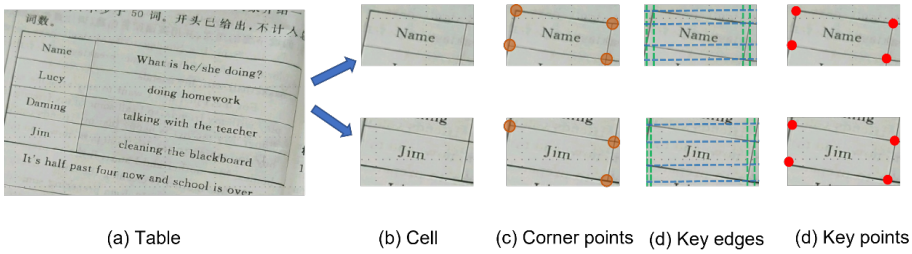


Fig. 2. We visualize the results of the corner point detection method and our key point detection method. For the cell "Name", both methods predict correctly. However, for the cell "Jim", the corner point detection method predicts incorrectly because the cell misses a lower-left corner point. However, our key point detection method avoids this error by obtaining the critical point from the critical edge with the help of SBD's edge detection.

The most intuitive way of key point regression is to directly predict the corner points of the target to localize it like CornerNet [10], and then Liu et al. [16] proposed a method called SBD to solve the LC (Learning Confusion) problem [15], which first predicts the eight boundary values of the target and

then combines them to obtain the four key points of the target. We refer to it to design our key point regression branch. Compared to direct corner point prediction, our method has certain advantages. As shown in Fig. 2, the corner characteristics of the cells may be incomplete for defective tables and wireless tables. In this case, it is difficult for the direct corner point prediction method to accurately predict the four corner points, which brings a large error. But for SBD, the eight bounding key edges of the cell are obtained first. Key points can be predicted with the help of border information, text information, not just relying on the four corners. Our key point prediction method locates four key points of a cell with the help of the location of eight boundaries, which is more adaptive in natural scene table recognition.

3.3 Centroid Filtering Non Maximum Suppression (CF-NMS)

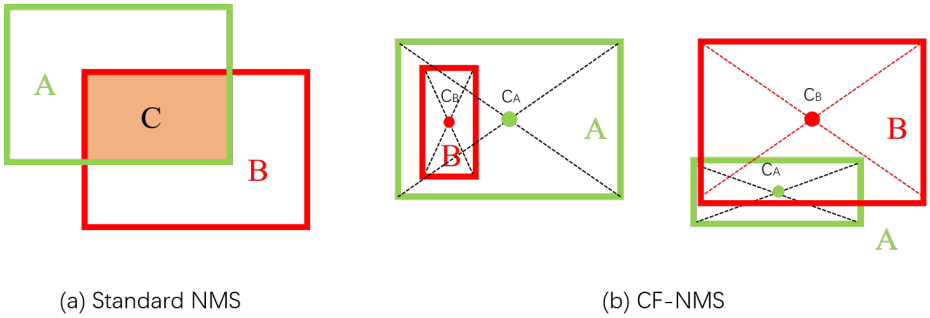


Fig. 3. Illustration of NMS and our CF-NMS. (a) is the standard NMS and (b) is our proposed CF-NMS.

In the process of fusion of two branches of SKCN, there's bound to be many overlapped and redundant bounding boxes. How to effectively filter the wrong detection boxes is the key to ensure the model performance. The process of the standard NMS algorithm (Fig.3(a)) consists of: (1) set the confidence threshold for the target box, (2) arrange the list of candidate boxes in descending order of their confidence, (3) select the box A with the highest confidence and add it to the output list,while removing it from the list of candidate boxes, (4) calculate the IOU value of all boxes in the list of candidate boxes with A, and remove the candidate boxes with IOU values greater than the threshold value, (5) repeat the above process until the list of candidate boxes is empty, and return the output list. The effectiveness of the standard NMS algorithm depends on the setting of the IoU threshold. A relatively high threshold can result in a large number of false positives, while a lower threshold can result in missing highly overlapped correct results.

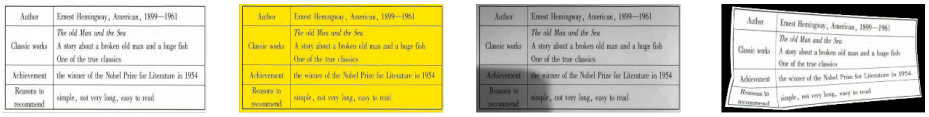
The standard NMS algorithm is not fully applicable to the filtering of candidate boxes in natural scene tables where many complexities exist. On the one

hand, the threshold of NMS cannot be set too low, because the bounding boxes of adjacent cells in the distorted skew table often have a large part of overlapping areas. In order to ensure the integrity of cell detection, we have to set the NMS threshold higher, which results in many redundant detection boxes not being filtered out. On the other hand, in some large tables, the GT boxes of cells are relatively dense and the area of each box is small. Therefore, it can also be regarded as a dense detection task of small targets. In this case, it is easy to predict some bounding boxes that are wrapped internally or around the perimeter of the correct box. Such errors are unavoidable due to the relatively high NMS threshold. Therefore, in order to solve the aforementioned problems of standard NMS in cell detection tasks, a new NMS algorithm is urgently needed to solve the existing problems of redundant bounding box detection.

Therefore, we propose a new non-maximum suppression algorithm based on centroid filtering (CF-NMS) to avoid threshold setting. CF-NMS filters overlapping bounding boxes by centroids, as shown in Fig. 3(b). Assuming that box A is the correct bounding box to be picked out, if the center of box B is inside box A, or conversely the center of box A is inside box B, then box B is judged to be redundant. This can effectively eliminate the case of nested detection boxes and make the detection results more accurate.

3.4 Tabular data augmentation

To address the lack of tabular datasets, we propose a tabular data enhancement module (TabSynth) to expand the number and diversity of tables and improve the performance of the model online. We propose three types of enhancement methods. The first is color variation, whose change is achieved by changing the HSV value of the table image. The second is shading transformation, which changes the lighting conditions of the table by combining the collected shadow photos with the table image to get a new table image with shading. And the third is geometric transformation that changes the degree of tilt and distortion of the table. The steps to achieve it are similar to the document image composition process described in DocUNet [18]. Our enhancement is not a random enhancement, but a targeted solution to two problems. Firstly, the distribution of various types of tables in the existing datasets is not uniform. Some types of tables with larger percentages are better trained and therefore more likely to yield better results than others, while some types of tables with smaller percentages are not sufficiently trained, which often leads to poorer performance. Therefore, we address the problem of data distribution by using TabSynth to enrich the sample diversity of the existing natural scene table datasets to enable each type of table to be adequately trained. Secondly, the existing datasets also have some tables with extreme aspect ratios, distortions and skews, and the structure recognition of these tables is also very challenging with existing methods. Therefore, our augmentation module can increase the number of these difficult samples in a targeted manner, so that the structure recognition model can fully learn the characteristics of the difficult samples and achieve better performance.



(a) Original Image (b) Background colour (c) Light intensity (c) Geometric distortion

Fig. 4. Example of some augmented tabular data.

3.5 Table structure recovery

After obtaining refined detection cells, we further designed an adaptive adjacency matching algorithm to reconstruct the table structure. First, the four corner points of all cells in the table are arranged in the order of top-left, top-right, bottom-right and bottom-left. Then, we propose a center line matching strategy to perform row/column matching on these cells. For example, in row matching we first use the center point of the right boundary to match the rights, and adjust the coordinates of the right boundary used for matching according to the size of the matched cells, and then pass them to the right side one by one. Left row matching is similarly. Since cells in natural scenes are not usually aligned, and the idea of dealing with cross-row and cross-column cases is to use small cells to match large cells, we use an adaptive boundary matching strategy, which means that the current cell boundary used for matching will be adjusted according to the matched cells. For paired cell boxes in right matching, $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$ and $\{(x'_1, y'_1), (x'_2, y'_2), (x'_3, y'_3), (x'_4, y'_4)\}$, if $\frac{y_2+y_3}{2} >= y'_1$ and $\frac{y_2+y_3}{2} <= y'_4$, the paired boxes are predicted to belong to the same row. Then the coordinates of the right border used for matching are adjusted by $y_2 = min(y_2, y'_2)$ and $y_3 = max(y_3, y'_3)$.

4 Experiments

In this section, we conducted experiments on two publicly available natural scene table datasets to evaluate the performance of our proposed table structure recognition method. To verify the effectiveness of the SKCN and the CF-NMS for the table structure recognition task, we conducted ablation experiments. The following are the relevant details of the experiments.

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on two publicly available natural scene table datasets, WTW and TAL_OCR_TABLE.

WTW [17] is a challenging and complex dataset for table structure recognition in the wild with 10970 training images and 3611 testing images, a sum of 14581 images. WTW divides the data into 7 cases by their own characteristics and unique challenges: simple, inclined, extreme aspect ratio, occluded and

blurred, overlaid, multi-color and gird, and curved. The dataset annotation contains table ids, table coordinates, cell coordinates and row/column information about cells. We cropped out table regions from the original images and used the tilt angle of the table regions obtained by Hough Transform to rotation correction for training and testing. We followed [17] using the cell adjacency relationship (IoU = 0.6) [4] as the evaluation metric for this dataset. There are two versions of the evaluation metric for cell adjacent relationship, ICDAR2013 [5] and ICDAR2019 [4]. Because some tabular datasets do not have textual annotations, such as WTW, the previous version cannot be used in this case. We used the more general version of ICDAR2019 without exact text-matching.

TAL_OCR_TABLE (TAL) [2] is a natural scene table dataset provided in the PRCV2021 TAL table recognition competition, which focuses on wired tables for educational scenarios. The dataset contains 18,000 images, 16,000 of which have provided annotations for training and 2,000 for testing. The annotation of the dataset includes the physical location of the cells and the HTML code of the table. The physical location of the table is annotated by the four vertices of the quadrilateral. We also cropped out the table regions from the original images for training and testing.

Evaluation Metrics. There are two common evaluation metrics used in table recognition tasks, TEDS and cell adjacent relationship.

Tree Edit Distance based Similarity (TEDS) [33] represents the logical structure of a table with a tree structure and examines the table structure recognition results at the global tree-structure level. It uses the tree edit distance to evaluate the accuracy of table structure recognition, with higher values being better. The TEDS results contain the extra results of text recognition, and taking OCR errors into account may lead to unfair comparisons, since previous work used different OCR models. Therefore, the TEDS metrics in this paper only calculate the results for the logical structure of the table, without considering the OCR recognition results.

Cell adjacent relationship [4] is used to evaluate the effectiveness of structure recognition by the accuracy of the physical location and the row/column coordinates of each sampled adjacent cell pair. The adjacency relationship of each cell is generated with its horizontal and vertical adjacent cells. Then precision, recall and F1 scores are calculated to compare the predicted relationships with the ground truth.

4.2 Implementation details

All experiments were implemented in PyTorch with 4x2080Ti GPUs. In Table 1, we compared the experimental results of different backbones. Since the difference between them is not very significant, we regard ResNet-50 as the backbone of network by default in the subsequent experiments. From the comparison of the results of different cell detection strategies, the accuracy of the box segmentation branch is higher than that of the key point regression branch, while the recall is

Table 1. Results with different backbones and cell detection strategies on TAL dataset. Here, * denotes using our online tabular data generation module TabSynth. SEG refers to the box segmentation. KEY refers to the key point regression.

Training data	Backbone	Strategy	Prec.(%)	Rec. (%)	F1.(%)
TAL	ResNet-50	SEG	99.5	98.89	99.2
		KEY	99.12	99.41	99.27
		SKCN	99.54	99.4	99.47
TAL	ResNet-101	SEG	99.6	98.89	99.24
		KEY	99.18	99.43	99.31
		SKCN	99.6	99.39	99.49
TAL*	ResNet-50	SEG	99.87	99.36	99.61
		KEY	99.78	99.84	99.81
		SKCN	**99.88**	**99.82**	**99.86**

lower. We further find that the SKCN after the collaborative processing of these two strategies can improve the prediction results synthetically. After adding our TabSynth module for training, the detection performance is also further improved, and the effect of this improvement is greater than replacing ResNet-50 with ResNet-101. This module allows us to adequately explore the potential of the model and achieve better results with a smaller cost for the model.

4.3 Comparisons with prior arts

We have compared our proposed method with several state-of-the-art methods on the public datasets TAL and WTW. our method achieves a state-of-the-art performance of 99.35% in terms of TEDS, as shown in Table 2. The experiments for SPLERGE [28] and CascadeTabNet [20] were reproduced based on the authors' original design. Since they are designed for scanned tables, they could not perform well in natural scenarios. To validate the effectiveness of our method on boundary warping or bending tables in natural scenarios, we conducted experiments on the WTW dataset. The results in Table 3 show that our method outperforms existing methods in terms of F1 scores for cell adjacent relationship, improving by 1.2% over Cycle-CenterNet, designed specifically for natural scenes, and by 0.2% over TSRFormer [12], which is able to robustly identify the structure of distorted tables with and without borders.

To better verify the robustness of our approach to complex situations, we analyzed the F1 scores of different types of tables on WTW, as shown in Table 3. Although our performance is slightly lower than Cycle-CenterNet on three ordinary table subsets, our method shows significant improvements on complex scenarios. In particular, for the subset "overlaid", we achieve a 14% improvement with the mainly contribution of CF-NMS. The experiments on these subsets fully demonstrate the superiority of our method and the ability to deal with complex scenarios in table structure recognition.

Table 2. Comparison of TEDS on TAL dataset

Method	TEDS(%)
SPLERGE [28]	53.14
CascadeTabNet [20]	66.71
Table-Master [30]	94.30
SCAN [29]	98.45
TAL_First_Place [2]	99.20
Ours	**99.35**

Table 3. Comparison of cell adjacent relationship on WTW dataset

Method	Curved	Overlaid	Simple	Occluded and blurred	Extreme aspect ratio	Inclined	Multi color and grid	All		
	F1.(%)							Prec.(%)	Rec.(%)	F1.(%)
Cycle-CenterNet [17]	76.1	84.1	**99.3**	77.4	91.9	**97.7**	**93.7**	93.3	91.5	92.4
FLAG-Net [14]	-	-	-	-	-	-	-	91.6	89.5	90.5
TSRFormer [12]	-	-	-	-	-	-	-	93.7	**93.2**	93.4
Ours	**83.4**	**98.1**	98.9	**82.6**	**96.3**	97.2	92.7	**94.2**	93.1	**93.6**

Table 4. Ablation experiments on TAL dataset

Training data	TabSynth	SEG	KEY	CF-NMS	TEDS(%)
TAL		✓			93.2
	✓	✓			97.7
	✓		✓		98.5
	✓	✓	✓		98.7
	✓	✓	✓	✓	**99.4**

4.4 Ablation studies

We conducted a series of experiments on the TAL and WTW datasets to verify the effectiveness of the proposed modules, and the experimental results are shown in Table 4. For TAL dataset, after adding our tabular data generation module TabSynth for training, the training data of the model can simulate complex scenarios with random distortions, random light and multi-color background to overcome the difficulty of the lack of natural scene dataset, thus making the model more robust and achieving a 4.5% improvement. We thus use TabSynth to assist in training by default. The results show that the TEDS metric of the KEY branch is higher than that of the SEG branch, and the interactive results of the two branches outperform the results of the two branches individually, providing support for our method and demonstrating that our method is more suitable for challenging table structure recognition tasks in natural scenes. What's more, our CF-NMS module, designed for the dense detection in table scenes, contributes a 0.7% improvement over the standard NMS, whose threshold is set to 0.5 in our experiment.

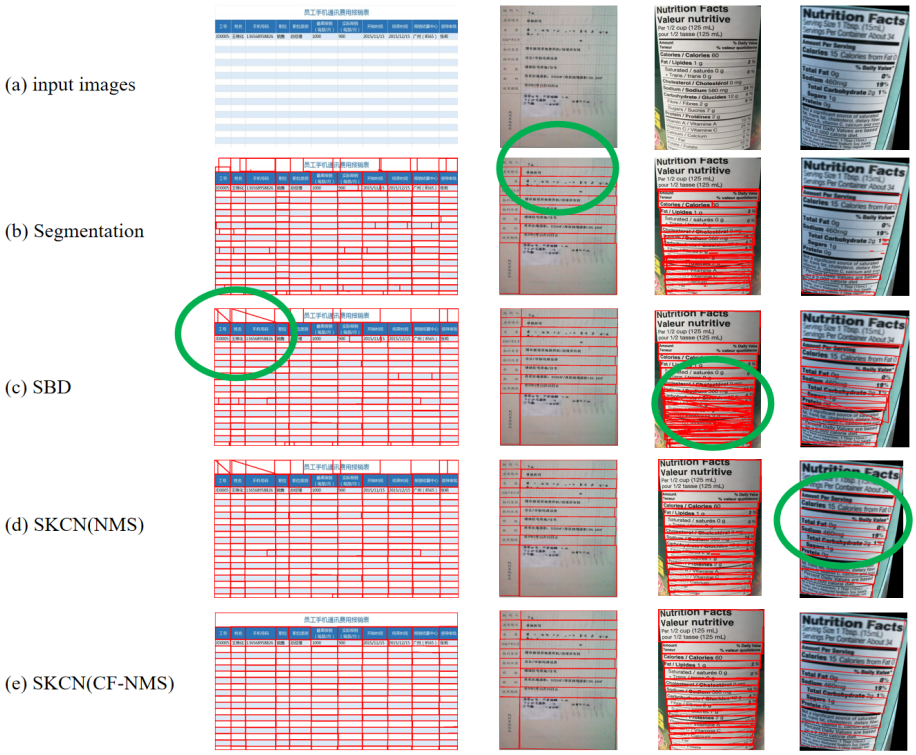


Fig. 5. Qualitative results of our approach.

Table 5. Ablation experiments on WTW dataset for cell detection(IOU=0.6)

Training data	SEG	KEY	CF-NMS	Prec.(%)	Rec.(%)	F1.(%)
WTW	✓			87.0	91.67	89.27
		✓		89.44	94.01	91.67
	✓	✓		93.4	93.59	93.5
	✓	✓	✓	**96.87**	**93.84**	**95.28**

For the WTW dataset, our approach also achieves considerable improvements in F1 scores. Replacing the standard NMS with our CF-NMS can effectively handle the challenges of dense cell detection scenarios and can yield improvements in precision and recall, improving the F1 score by nearly 1.8%. This also shows that the CF-NMS is designed to be very friendly for cell detection tasks. Fig. 5 gives a demonstration of the qualitative results of our method, and it can be seen that for large dense table detection tasks, the collaboration of the SEG and KEY branches outperforms both in terms of refinement results. Our proposed SKCN and CF-NMS modules can even be applied to other dense target detection tasks in the future.

5 Conclusion

In this paper, we consider that existing networks do not fully exploit the features of tables and propose a segmentation and key point collaboration network (SKCN) for table structure recognition in the wild. Unlike previous detection methods, the two branches of our model can complement each other during training process and the results of each branch can be fused to obtain a refined result. To better cope with complex table scenarios, we further propose CF-NMS and a tabular data generation module. Experimental results show that our method achieves state-of-the-art performance on two public benchmarks, including TAL and WTW.

6 Acknowledgments

This research is supported in part by GD-NSF (No. 2021A1515011870), NSFC (Grant no. 61771199), Zhuhai Industry Core and Key Technology Research Project (No. 2220004002350), and the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant 2020GH17).

References

1. Chi, Z., Huang, H., Xu, H.D., Yu, H., Yin, W., Mao, X.L.: Complicated table structure recognition. arXiv preprint arXiv:1908.04729 (2019)
2. Contributors, T.: TAL_OCR_TABLE:a scene table structure recognition benchmark. https://ai.100tal.com/dataset (2021)

3. Deng, Y., Rosenberg, D., Mann, G.: Challenges in end-to-end neural scientific table recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 894–901. IEEE (2019)
4. Gao, L., Huang, Y., Déjean, H., Meunier, J.L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: Icdar 2019 competition on table detection and recognition (ctdar). In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1510–1515. IEEE (2019)
5. Göbel, M., Hassan, T., Oro, E., Orsi, G.: A methodology for evaluating algorithms for table understanding in pdf documents. In: Proceedings of the 2012 ACM symposium on Document engineering. pp. 45–48 (2012)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
7. Itonori, K.: Table structure recognition based on textblock arrangement and ruled line position. In: Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93). pp. 765–768. IEEE (1993)
8. Kieninger, T., Dengel, A.: The t-recs table recognition and analysis system. In: International Workshop on Document Analysis Systems. pp. 255–270. Springer (1998)
9. Laurentini, A., Viada, P.: Identifying and understanding tabular material in compound documents. In: International Conference on Pattern Recognition. pp. 405–405. IEEE COMPUTER SOCIETY PRESS (1992)
10. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
11. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: Tablebank: Table benchmark for image-based table detection and recognition. In: Proceedings of The 12th language resources and evaluation conference. pp. 1918–1925 (2020)
12. Lin, W., Sun, Z., Ma, C., Li, M., Wang, J., Sun, L., Huo, Q.: Tsrformer: Table structure recognition with transformers. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6473–6482 (2022)
13. Liu, H., Li, X., Liu, B., Jiang, D., Liu, Y., Ren, B.: Neural collaborative graph machines for table structure recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4533–4542 (2022)
14. Liu, H., Li, X., Liu, B., Jiang, D., Liu, Y., Ren, B., Ji, R.: Show, read and reason: Table structure recognition with flexible context aggregator. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1084–1092 (2021)
15. Liu, Y., Jin, L.: Deep matching prior network: Toward tighter multi-oriented text detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1962–1969 (2017)
16. Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., Wang, Z.: Omnidirectional scene text detection with sequential-free box discretization. arXiv preprint arXiv:1906.02371 (2019)
17. Long, R., Wang, W., Xue, N., Gao, F., Yang, Z., Wang, Y., Xia, G.S.: Parsing table structures in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 944–952 (2021)
18. Ma, K., Shu, Z., Bai, X., Wang, J., Samaras, D.: Docunet: Document image unwarping via a stacked u-net. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4709 (2018)
19. Paliwal, S.S., Vishwanath, D., Rahul, R., Sharma, M., Vig, L.: 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). pp. 128–133. IEEE (2019)

20. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 572–573 (2020)
21. Qasim, S.R., Mahmood, H., Shafait, F.: Rethinking table recognition using graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 142–147. IEEE (2019)
22. Qiao, L., Li, Z., Cheng, Z., Zhang, P., Pu, S., Niu, Y., Ren, W., Tan, W., Wu, F.: Lgpma: Complicated table structure recognition with local and global pyramid mask alignment. In: International Conference on Document Analysis and Recognition. pp. 99–114. Springer (2021)
23. Raja, S., Mondal, A., Jawahar, C.: Table structure recognition using top-down and bottom-up cues. In: European Conference on Computer Vision. pp. 70–86. Springer (2020)
24. Rastan, R., Paik, H.Y., Shepherd, J.: Texus: A unified framework for extracting and understanding tables in pdf documents. Information Processing & Management **56**(3), 895–918 (2019)
25. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1162–1167. IEEE (2017)
26. Shigarov, A., Mikhailov, A., Altaev, A.: Configurable table structure recognition in untagged pdf documents. In: Proceedings of the 2016 ACM symposium on document engineering. pp. 119–122 (2016)
27. Siddiqui, S.A., Fateh, I.A., Rizvi, S.T.R., Dengel, A., Ahmed, S.: Deeptabstr: Deep learning based table structure recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1403–1409. IEEE (2019)
28. Tensmeyer, C., Morariu, V.I., Price, B., Cohen, S., Martinez, T.: Deep splitting and merging for table structure decomposition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 114–121. IEEE (2019)
29. Wang, H., Xue, Y., Zhang, J., Jin, L.: Scene table structure recognition with segmentation collaboration and alignment. Pattern Recognition Letters (2022)
30. Ye, J., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., Xiao, R.: Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: Table recognition to html. arXiv preprint arXiv:2105.01848 (2021)
31. Zhang, Z., Zhang, J., Du, J., Wang, F.: Split, embed and merge: An accurate table structure recognizer. Pattern Recognition **126**, 108565 (2022)
32. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: 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. pp. 697–706 (2021)
33. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European Conference on Computer Vision. pp. 564–580. Springer (2020)