

# An End-to-End Multi-Task Learning Model for Image-based Table Recognition

Nam Tuan Ly<sup>a</sup> and Atsuhiko Takasu<sup>b</sup>  
*National Institute of Informatics (NII), Tokyo, Japan*

**Keywords:** Table Recognition, End-to-End, Multi-Task Learning, Self-Attention.

**Abstract:** Image-based table recognition is a challenging task due to the diversity of table styles and the complexity of table structures. Most of the previous methods focus on a non-end-to-end approach which divides the problem into two separate sub-problems: table structure recognition; and cell-content recognition and then attempts to solve each sub-problem independently using two separate systems. In this paper, we propose an end-to-end multi-task learning model for image-based table recognition. The proposed model consists of one shared encoder, one shared decoder, and three separate decoders which are used for learning three sub-tasks of table recognition: table structure recognition, cell detection, and cell-content recognition. The whole system can be easily trained and inferred in an end-to-end approach. In the experiments, we evaluate the performance of the proposed model on two large-scale datasets: FinTabNet and PubTabNet. The experiment results show that the proposed model outperforms the state-of-the-art methods in all benchmark datasets.


## 1 INTRODUCTION


The tabular format is one of the rich-information formats and is widely used in communication, research, and data analysis. Tables commonly appear in research papers, books, handwritten notes, invoices, financial documents, and many other places. Thus, table understanding becomes one of the essential techniques in document analysis systems and attracts the attention of numerous researchers.

Image-based table recognition is the key step of table understanding which refers to the representation of a table image in a machine-readable format, where its structure and the content within each cell are encoded according to a pre-defined standard (Zhong et al., 2020). The machine-readable format can be HTML code (Jimeno Yepes et al., 2021; Li et al., 2019; Zhong et al., 2020) or LaTeX code (Deng et al., 2019; Kayal et al., 2021). The choice of the machine-readable format is ultimately not very important, since one can be transformed into the other. Image-based table recognition is a challenging task due to the diversity of table styles and the complexity of table structures. In the past few decades, many table recognition methods have been proposed and can be

divided into two categories: end-to-end methods and non-end-to-end methods. Most of the previous works (Nassar et al., 2022; Qiao et al., 2021; Ye et al., 2021) focus on non-end-to-end approaches which divide the problem into two separate sub-problems: table structure recognition; and cell-content recognition, and then attempt to solve each sub-problem independently using two separate systems. On the other hand, the end-to-end approach attempts to solve the problem using a single model (specifically a Deep Neural Network) and achieves state-of-the-art results on many tasks such as machine translation (Vaswani et al., 2017), speech recognition (Bahdanau et al., 2016), and text recognition (Lu et al., 2021; Ly et al., 2021). However, to the best of our knowledge, there are few studies (Deng et al., 2019; Zhong et al., 2020) on the end-to-end approach to table recognition and their performance is mediocre compared to the non-end-to-end methods.

In this paper, we formulate the problem of table recognition as a multi-task learning problem, which requires the model to be jointly learned in three sub-tasks of table recognition. To address this problem, we propose a novel end-to-end multi-task learning model which consists of one shared encoder, one

<sup>a</sup>  <https://orcid.org/0000-0002-0856-3196>

<sup>b</sup>  <https://orcid.org/0000-0002-9061-7949>

shared decoder, and three separate decoders for three sub-tasks of table recognition: table structure recognition, cell detection, and cell-content recognition. The model takes an input table image and produces the table structure information, location of table cells, and contents of table cells, which can be easily transformed into the HTML code (or LaTeX code) representing the table. The shared components are repeatedly trained from the gradients received from three sub-tasks while each of three separate decoders is trained from the gradients of its task. The whole system can be easily trained and inferred in an end-to-end approach.

We have evaluated the performance of our model on the PubTabNet (Zhong et al., 2020) and FinTabNet (Zheng et al., 2021) datasets, demonstrating that our model outperforms state-of-the-art methods in both table structure recognition and table recognition. We further evaluated our model on the final evaluation set of Task-B in the ICDAR 2021 competition on scientific literature parsing (Jimeno Yepes et al., 2021) (ICDAR 2021 competition in short), demonstrating that our model achieves competitive results when compared to the top three solutions. The code will be publicly released to GitHub.

In summary, the main contributions of this paper are as follows:

- We present a novel end-to-end multi-task learning model for image-based table recognition. The proposed model can be easily trained and inferred in an end-to-end approach.
- Across all benchmark datasets, the proposed model outperforms the state-of-the-art methods.
- Although we used neither any additional training data nor ensemble techniques, our model achieves competitive results when compared to top three solutions in ICDAR2021 competition.

The rest of this paper is organized as follows. In Sec.2, we give a brief overview of the related works. In Sec. 3, we introduce the overview of the proposed model. In Sec. 4, we report the experimental details and results. Finally, we draw conclusions in Sec. 5.

## 2 RELATED WORK

Table understanding in unstructured documents can be defined in three steps: 1) table detection: detecting the bounding boxes of tables in documents (Casado-García et al., 2020; Huang et al., 2019); 2) table structure recognition: recognizing the structural information of tables (Itonori, 1993; Kieninger, 1998; Wang et al., 2004); 3) table recognition: recognizing

both the structural information and the content within each cell of tables (Deng et al., 2019; Ye et al., 2021; Zhong et al., 2020). We will briefly survey table structure recognition, and then table recognition.

**Table Structure Recognition** can be considered the first step of table recognition and has been studied by researchers around the world for the past few decades. Early works of table structure recognition are based on hand-crafted features and heuristic rules (Itonori, 1993; Kieninger, 1998; Wang et al., 2004). These methods are mostly applied to a simple structure or pre-defined data formats. In recent years, inspired by the success of deep learning in various tasks, especially object detection and semantic segmentation, many deep learning-based methods (Raja et al., 2020; Schreiber et al., 2017) have been presented to recognize table structures. S. Schreiber et al. (Schreiber et al., 2017) proposed a two-fold system named DeepDeSRT that applies Faster RCNN (Ren et al., 2015) and FCN(Long et al., 2015) for both table detection and row/column segmentation. Sachin et al. (Raja et al., 2020) presented a table structure recognizer named TabStruct-Net that combines cell detection and interaction modules to localize the cells and predicts their row and column associations with other detected cells. Structural constraints are incorporated as additional differential components to the loss function for cell detection. Recently, graph neural networks are also used for table structure recognition by encoding document images as graphs (Qasim et al., 2019).

**Table Recognition:** Most of the previous works of table recognition (Nassar et al., 2022; Qiao et al., 2021; Ye et al., 2021; Zhang et al., 2022) focus on non-end-to-end approaches which divide the problem into two separate sub-problems: table structure recognition; and cell-content recognition, and then attempt to solve each sub-problem independently using two separate systems. J. Ye et al. (Ye et al., 2021) proposed a Transformer-based model named TableMASTER for table structure recognition and combined it with a text line detector to detect text lines in each table cell. Finally, they employed a text line recognizer based on (Lu et al., 2021) to recognize each text line. Their system achieved second place in ICDAR2021 competition. A. Nassar et al. (Nassar et al., 2022) proposed a Transformer-based model named TableFormer for recognizing both table structure and the bounding box of each table cell and then using these bounding boxes to extract the cell contents from the PDF to build the whole table recognition system. Z. Zhang et al. (Zhang et al., 2022) proposed a table structure recognizer, Split, Embed, and Merge (SEM) for recognizing the table

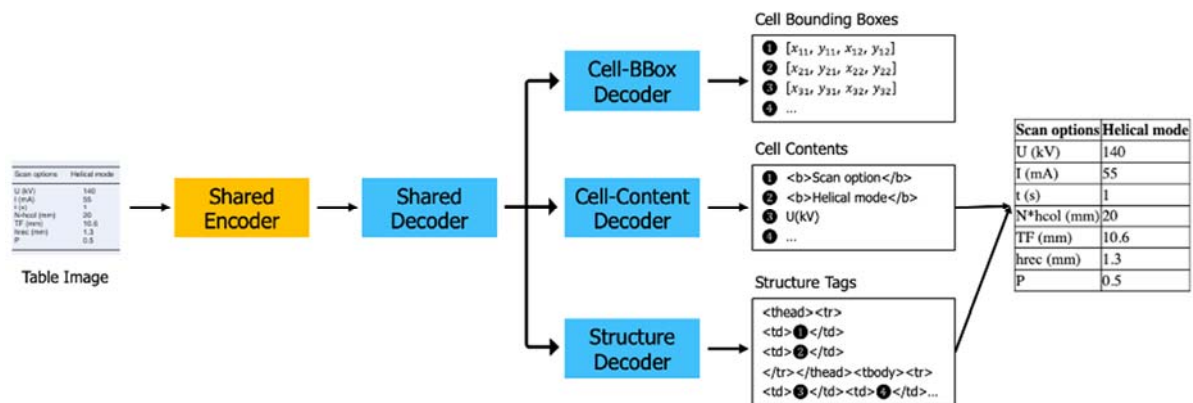


Figure 1: The overview of the proposed model.

structure. Then, they combined SEM with an attention-based text recognizer to build the table recognizer and achieved third place in the ICDAR2021 competition.

Recently, due to the rapid development of deep learning and the increase in the tabular data, some works (Deng et al., 2019; Zhong et al., 2020) try to focus on end-to-end approaches. However, their performance is still mediocre compared to the non-end-to-end methods. Y. Deng et al. (Deng et al., 2019) formulated table recognition as the image to latex problem and employed IM2TEX (Deng et al., 2016) model for table recognition. X. Zhong et al (Zhong et al., 2020) proposed an encoder-dual-decoder (EDD) model for recognizing both table structure and content of each cell. They also publicized a table recognition dataset PubTabNet to the community.

In 2021, IBM Research in conjunction with IEEE ICDAR held ICDAR2021 competition on scientific literature parsing (Jimeno Yepes et al., 2021) (ICDAR2021 competition in short). The competition consists of two tasks: Task A - Document layout recognition which identifies the position and category of document layout elements, including title, text, figure, table, and list; and Task B - Table recognition which converts table images into HTML code. For Task B, there are 30 submissions from 30 teams for the Final Evaluation Phase and most of the top 10 systems are non-end-to-end approaches and employ ensemble techniques to improve their performance.

### 3 METHODOLOGY

The proposed model consists of one shared encoder, one shared decoder, and three separate decoders for three sub-tasks of the table recognition problem as

shown in Fig. 1. The shared encoder encodes the input table image as a sequence of features. The sequence of features is passed to the shared decoder and then the structure decoder to predict a sequence of HTML tags that represent the structure of the table. When the structure decoder produces the HTML tag representing a new cell ('<td>' or '<td ...>'), the output of the shared decoder corresponding to that cell and the output of the shared encoder are passed into the cell-bbox decoder and the cell-content decoder to predict the bounding box coordinates and the text content of that cell. Finally, the text contents of cells are inserted into the HTML structure tags corresponding to their cells to produce the final HTML code of the input table image. Fig. 2 shows the detail of the five components in our model. We describe the detail of each component in the following sections.

#### 3.1 Shared Encoder

In this work, we use a CNN backbone network as the feature extractor followed by a positional encoding layer to build the shared encoder. The feature extractor extracts visual features from an input table image of the size  $\{h, w, c\}$  ( $c$  is the color channel), resulting in a feature grid  $F$  of the size  $\{h', w', k\}$  ( $k$  is the number of the feature maps). Then the feature grid  $F$  is unfolded into a sequence of features (column by column from left to right in each feature map) before being fed into the positional encoding layer to get the encoded sequence of features. The encoded sequence of features will be fed into the shared decoder and three separate decoders.

#### 3.2 Shared Decoder

The architecture of all decoders in our model is inspired by the original Transformer decoder

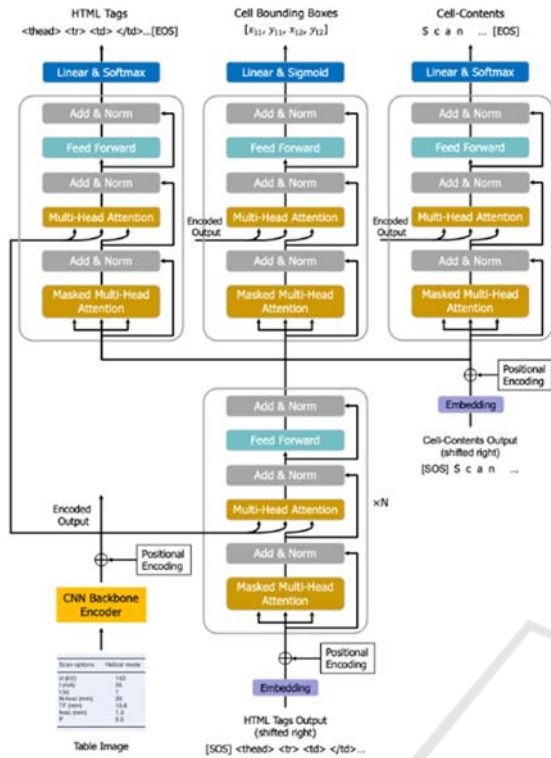


Figure 2: Network architecture of the proposed model.

(Vaswani et al., 2017) which is composed of a stack of  $N$  identical Transformer decoder layers where  $N$  can be a hyperparameter. Each identical Transformer decoder layer (identical layer in short) has three sub-layers: a multi-head self-attention mechanism; a masked multi-head self-attention mechanism; and a position-wise fully connected feed-forward network, and helps the decoders focus on appropriate places in the input table image.

At the top of the shared encoder, the shared decoder is composed of a stack of  $N=2$  identical layers. As shown in Fig.2, the output of the shared encoder is fed into the multi-head self-attention mechanism of each identical layer as the value and key vectors. During training, the right-shifted sequence of target HTML tags (structural tokens) of the table structure (after passing through the embedded layer and the positional encoding layer) is passed into the bottom of the shared decoder as the query vector. In the inference stage, the right-shifted sequence of target HTML tags is replaced by the right-shifted sequence of HTML tags outputted by the structure decoder. Finally, the outputs of the shared decoder will be fed into the three separate decoders to predict three sub-tasks of the table recognition problem.

### 3.3 Structure Decoder

At the top of the shared decoder, the structure decoder uses the outputs of the shared decoder and the outputs of the shared encoder to predict a sequence of HTML tags of the table structure. Inspired by the works in (Zhong et al., 2020), the HTML tags of the table structure are tokenized at the HTML tag level except for the tag of a cell. In our model, the form of '`<td></td>`' is treated as one token class. Note that this can largely reduce the length of the sequence. We also break down the tag of the spanning cells into '`<td>`', '`rowspan=`' or '`colspan=`', with the number of spanning cells, and '`>`'. Thus, the structural token of '`<td></td>`' or '`<td>`' represents a new table cell. As shown in Fig. 2, the structure decoder is composed of one identical layer followed by a linear layer and a softmax layer. The identical layer takes the outputs of the shared decoder as the query vector input and the outputs of the shared encoder as the key and value vector inputs. The output of the identical layer is fed into the linear layer, and then the softmax layer to generate the sequence of structural tokens.

### 3.4 Cell-BBox Decoder

When the structure decoder generates a structural token representing a new cell ('`<td></td>`' or '`<td>`'), the cell-bbox decoder is triggered and uses the output of the shared decoder corresponding to this cell to predict the bounding box coordinates of this cell.

As shown in Fig. 2, we use one identical layer followed by a linear layer and a sigmoid layer to build the cell-bbox decoder. The identical layer takes the output of the shared decoder and the output of the shared encoder as the input and learns to focus on appropriate places in the input image. The output of the identical layer is fed into the linear layer and then the sigmoid layer to predict the four coordinates of the cell bounding box.

### 3.5 Cell-Content Decoder

Similar to the cell-bbox decoder, the cell-content decoder selects the outputs of the shared decoder referring to the structural tokens representing a new cell ('`<td></td>`' or '`<td>`') and uses them to recognize the text contents of cells. The cell-content decoder in the proposed model can be considered a text recognizer and the text output are tokenized at the character level. In this work, we use one identical layer followed by a linear layer and a softmax layer to build the cell-content decoder as shown in Fig. 2. The output of the shared encoder are fed into the



identical layer as the input value and key vectors. During training, the right-shifted target of the cell content (the right-shifted output of the cell-content decoder in the testing phase) is passed through the embedded and the positional encoding layers and then added to the output of the shared decoder before being fed into the identical layer as the query vector. Finally, the output of the identical layer is fed into the linear layer and then the softmax layer to generate the cell content.

### 3.6 Network Training

In our model, the shared components are repeatedly trained from the gradients received from three sub-tasks while each of three separate decoders is trained from the gradients obtained from its task. The whole system can be trained end-to-end on pairs of table images and their annotations of the table structure, the text content, and its bounding box per non-empty table cell by stochastic gradient descent algorithms. The overall loss of our model is defined as the following:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{struc.}} + \lambda_2 \mathcal{L}_{\text{cont.}} + \lambda_3 \mathcal{L}_{\text{bbox}} \quad (1)$$

where  $\mathcal{L}_{\text{struc.}}$  and  $\mathcal{L}_{\text{cont.}}$  are the table structure recognition loss and the cell-content prediction loss, respectively that are implemented in Cross-Entropy loss,  $\mathcal{L}_{\text{bbox}}$  is the cell-bbox regression loss which is optimized by L1 loss.  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  are weight hyperparameters.

## 4 EXPERIMENTS

To evaluate the performance of the proposed model, we conducted experiments on two datasets: FinTabNet (Zheng et al., 2021) and PubTabNet (Zhong et al., 2020). The information of the datasets is given in Sec 4.1. The implementation details are described in Sect. 4.2; the experimental results are presented in Sect. 4.3; and the visualization results are shown in Sect. 4.4.

### 4.1 Datasets

In this paper, we conduct the experiments on the following two large-scale datasets that contain the annotations of both the structure of the table and the text content with the position of each non-empty table cell.

**PubTabNet** (Zhong et al., 2020) is a large-scale table image dataset that contains over 568k samples with their corresponding annotations of the table

structure presented in HTML format, the text content, and its bounding box per non-empty table cell. This dataset is created by collecting scientific articles from PubMed Central Open Access Subset (PMCOA). The dataset is used in the ICDAR2021 competition (Jimeno Yepes et al., 2021) and divided into 500,777 training samples and 9,115 validation samples in the development phase, and 9,064 final evaluation samples in the Final Evaluation Phase.

**FinTabNet** is another large-scale table image dataset published by X. Zheng et al. (Zheng et al., 2021). The dataset is composed of complex tables from the annual reports of the S&P 500 companies with detailed annotations of the table structure and table cell information like the PubTabNet dataset. This dataset consists of 112k table images which are divided into training, testing, and validation sets with a ratio of 81% : 9.5% : 9.5%.

### 4.2 Implementation Details

We use the ResNet-31 network (He et al., 2016) to build the CNN backbone in our model. To enable the CNN backbone network to model the global context from the input image, we add the Multi-Aspect Global Context Attention (GCAttention) proposed by Ning Lu et. al. (Lu et al., 2021) after each residual block of the ResNet-31 network. All images are resized to 480\*480 pixels and the feature map outputted from the CNN backbone has a dimension of 60\*60.

At the decoders, all identical layers have the same architecture with the input feature size of 512, the feed-forward network size of 2048, and 8 attention heads. The maximum length of a sequence of structural tokens in the structure decoder is 500 and the maximum length of a sequence of cell tokens in the cell-content decoder is 150. We empirically set all weight hyperparameters as  $\lambda_1 = \lambda_2 = \lambda_3 = 1$ .

Our model is implemented with Pytorch and the MMCV library (MMCV Contributors, 2018). The model is trained on two NVIDIA A100 80G with a batch size of 4 in each GPU. The initializing learning rate is 0.001 for the first 12 epochs. Afterward, we reduce the learning to 0.0001 and train for 8 more epochs or convergence.

### 4.3 Experimental Results

To evaluate the performance of the proposed model, we employ the Tree-Edit-Distance-Based Similarity (TEDS) metric as defined in (Zhong et al., 2020). We also denote TEDS-struc. as the TEDS score between

two tables when considering only the table structure information.

### 4.3.1 Table Structure Recognition

First, we conducted experiments to verify the effectiveness of the proposed model for recognizing the table structure. Table 1 compares the table structure recognition performance (TEDS-struct. scores) of the proposed model and the previous table structure recognition methods on PubTabNet and FinTabNet datasets.

As shown in Table 1, our model achieves superior performance on all benchmark datasets compared to the state-of-the-art models. Specifically, with TEDS-struct. of 98.79% on the FinTabNet dataset, our model improves TableFormer (Nassar et al., 2022) by 2% and other methods by more than 7.7%, and even GTE (FT) (Zheng et al., 2021) is pretrained in the PubTabNet dataset. On the PubTabNet dataset, our model achieved TEDS-struct. of 97.88% which again improves TableFormer and LGPMA (Qiao et al., 2021) by about 1.1% and other methods by more than 4.87%. Note that LGPMA requires additional

Table 1: Table structure recognition results on PubTabNet validation set (PTN) and FinTabNet (FTN).

Dataset	Model	TEDS-struct. (%)		
		Sim.	Com.	All
FTN	EDD (Zhong et al., 2020)	88.40	92.08	90.60
	GTE (Zheng et al., 2021)	-	-	87.14
	GTE <sup>(FT)</sup> (Zheng et al., 2021)	-	-	91.02
	TableFormer (Nassar et al., 2022)	97.50	96.00	96.80
	<b>Our Model</b>	<b>99.07</b>	<b>98.46</b>	<b>98.79</b>
PTN	EDD (Zhong et al., 2020)	91.10	88.70	89.90
	GTE (Zheng et al., 2021)	-	-	93.01
	LGPMA (Qiao et al., 2021)	-	-	96.70
	TableFormer (Nassar et al., 2022)	98.50	95.00	96.75
	<b>Our Model</b>	<b>99.05</b>	<b>96.66</b>	<b>97.88</b>

Sim. (Simple): Tables without multi-column or multi-row cells.

Com. (Complex): Tables with multi-column or multi-row cells.

(FT) Model was trained on PubTabNet and then finetuned.

annotation information for training. The proposed model also achieves state-of-the-art accuracies on complex tables (tables with multi-column or multi-row cells) of both datasets.

**Cell Detection:** Like any object detection model, the cell-bbox decoder produces the bounding boxes of the cells which can be evaluated by the PASCAL VOC mAP metric. Table 2 shows the mAP of the proposed model in comparison with the previous works in (Nassar et al., 2022) on the PubTabNet dataset. Our model achieves the state-of-the-art results and significantly improves TableFormer and EDD + BBox by more than 6.8%. Even without post-processing, the proposed model slightly outperforms TableFormer + PP which uses the information of the cell bounding boxes from the PDF document in post-processing.

Table 2: Cell detection results on PubTabNet validation set. PP: Post-processing.

Model	mAP (%)
EDD + BBox	79.20
TableFormer	82.10
EDD + BBox + PP	82.70
TableFormer + PP	86.80
<b>Our Model</b>	<b>88.93</b>

### 4.3.2 Results of Table Recognition

In this experiment, we evaluate the performance of the proposed model in the table recognition problem which recognizes both the structure of the table and the content within each cell. Table 3 shows table recognition results of the proposed model in comparison with the previous table recognition methods on the PubTabNet dataset. Our model outperforms all the previous methods without ensemble techniques. Specifically, our model achieved TEDS of 96.67% which improves VCGoup’s solution (Ye et al., 2021) by 0.41%, LGPMA + OCR (Qiao et al., 2021) by 2%, and others by more than 3%. Note that VCGoup’s solution, and SEM (Zhang et al., 2022) are the 2nd ranking, and 3rd ranking solutions in ICDAR2021 competition. LGPMA (Qiao et al., 2021) is the table structure recognizer component in the 1st ranking solution in ICDAR2021 competition. All other methods except EDD (Zhong et al., 2020) are non-end-to-end approach and the methods in (Qiao et al., 2021; Ye et al., 2021; Zhang et al., 2022) requires additional annotation information for training.

Table 3: Table recognition results on PubTabNet validation set.

Model	TEDS (%)		
	Sim.	Com.	All
EDD (Zhong et al., 2020)	91.20	85.40	88.30
TabStruct-Net (Raja et al., 2020)	-	-	90.10
GTE (Zheng et al., 2021)	-	-	93.00
TableFormer (Nassar et al., 2022)	95.40	90.10	93.60
SEM <sup>(3)</sup> (Zhang et al., 2022)	94.80	92.50	93.70
LGPMA + OCR <sup>(1)</sup> (Qiao et al., 2021)	-	-	94.60
VCGoup <sup>(2)</sup> (Ye et al., 2021)	-	-	96.26
<b>Our Model</b>	<b>97.92</b>	<b>95.36</b>	<b>96.67</b>
VCGoup + ME <sup>(2)</sup> (Ye et al., 2021)	-	-	<b>96.84</b>

(1)(2)(3) are 1st, 2nd, and 3rd ranking solutions in ICDAR2021 competition. ME: Model Ensemble.

Without ensemble technique as well as additional annotation information, however, our model achieves the competitive results when compared to VCGoup’s solution + ME in (Ye et al., 2021) which requires annotations of text-line bounding boxes of cell contents in table images and employs three model ensembles in the table structure recognition and three model ensembles in the text line recognition.

We also evaluate our model on the final evaluation set of the PubTabNet dataset which is used for the Final Evaluation Phase in ICDAR2021 competition. Table 4 compares TEDS scores by our model and the top 10 solutions in ICDAR2021 competition (Jimeno Yepes et al., 2021). Although we used neither any additional training data nor ensemble techniques, our model outperforms the 4th ranking solution named YG and achieves competitive results when compared to the top three solutions in the final evaluation set of Task-B in the ICDAR 2021 competition. Furthermore, the proposed model achieves a similar TEDS score on complex tables with the 2nd ranking solution named VCGroup. Note that the 1st ranking solution is a non-end-to-end approach which employs LGPMA (Qiao et al., 2021) to recognize the structure of the table and then uses attention-based text recognizer to provide the OCR information of the table cells. They also adopt multi-scale ensemble strategy to further improve the performance. The other 9 solutions also are non-end-to-end approaches and most of them use additional data for training as well as ensemble methods.

Table 4: Table recognition results on PubTabNet final evaluation set.

Team Name	TEDS (%)		
	Simp.	Comp.	All
Davar-Lab-OCR	97.88	94.78	<b>96.36</b>
VCGroup	<b>97.90</b>	94.68	96.32
XM	97.60	<b>94.89</b>	96.27
<b>Our Model</b>	97.60	94.68	96.17
YG	97.38	94.79	96.11
DBJ	97.39	93.87	95.66
TAL	97.30	93.93	95.65
PaodingAI	97.35	93.79	95.61
anyone	96.95	93.43	95.23
LTIAYN	97.18	92.40	94.84

#### 4.4 Visualization Results

In this section, we show some visualization results of the proposed model on the PubTabNet dataset. As shown in Fig. 3, the left image is the original image with detected bounding boxes of table cells, and the right image is the predicted HTML code of the table view on the web browser. As it is shown, our model is able to predict complex table structure as well as bounding boxes and contents for all table cells, even for the empty cells or cells that cross span multiple rows/columns.

## 5 CONCLUSIONS

In this paper, we formulate the problem of table recognition as a multi-task learning problem and propose a novel end-to-end multi-task learning model which consists of one shared encoder, one shared decoder, and three separate decoders for three sub-tasks of table recognition: table structure recognition, cell detection, and cell-content recognition. The shared components are repeatedly trained from the gradients received from three sub-tasks while each of three separate decoders is trained from the gradients of its task. Extensive experiments on two large-scale datasets demonstrate our model achieved state-of-the-art accuracies in both table structure recognition and table recognition. Although we used neither any additional training data nor ensemble techniques, our model outperforms the 4th ranking solution and achieves competitive results when compared to the top three solutions in ICDAR 2021 competition.

In the future, we will conduct experiments of the proposed model on the table image datasets of other languages. We also plan to incorporate language

Item	Total responders		North		Centre		South		p
	No	%	No	%	No	%	No	%	
Patient's opinion of doctor's explanation of the health problem									
Positive	1875	96.0	528	76.3	364	63.0	330	55.4	< 0.000
Negative			159	18.5	197	18.5	114	19.1	
Don't know			105	15.2	126	21.5	152	25.5	
Patient's view of the best information source on disease treatment									
Dermatologist	1870	95.7	434	63.2	333	56.7	292	49.0	< 0.000
General practitioner			162	23.6	60	10.2	137	23.0	
Patient association			30	4.3	132	22.5	64	11.1	
Other			61	8.9	62	10.4	101	16.9	
Patient's view as to whether patients require more information concerning therapy									
Yes	1884	96.4	607	88.1	575	96.5	482	80.5	< 0.000
No			82	11.9	21	3.5	117	19.5	
Patient's opinion of the best information source on psoriasis									
General practitioner	1842	94.3	148	23.0	113	19.0	121	20.0	< 0.000
Pharmacist			21	3.3	16	2.8	44	7.3	
Dermatologist			285	41.2	207	34.8	197	32.6	
Illustrated medication leaflet			37	5.1	37	6.2	19	3.2	
Health personnel			69	9.6	42	7.1	39	6.5	
Friends and family			13	2.0	15	2.6	16	2.6	
Health magazines			30	4.7	17	2.9	33	5.5	
Books			6	0.9	11	1.9	6	1.0	
Internet			25	3.9	53	9.0	42	7.0	
Newspapers			7	1.1	11	1.8	11	1.8	
Information campaigns			42	6.5	34	5.7	20	3.3	
Patient associations			19	3.0	95	16.0	34	5.6	
Other			11	1.7	4	0.6	22	3.6	
Knowledge of patients' rights									
Yes	1726	88.3	154	26.0	178	30.8	178	32.0	0.05
No			438	74.0	400	69.2	378	68.0	
Knowledge of homeopathic medication and herb-based products									
Yes	1855	94.9	124	19.1	123	20.6	187	30.7	< 0.000
No			526	80.9	473	79.4	422	69.3	
Knowledge of therapies such as acupuncture, the use of phototherapy									
Yes	1848	94.6	159	24.5	150	25.4	195	32.0	0.005
No			489	75.5	441	74.4	414	68.0	

Figure 3: Visualization results on PubTabNet.

models into the structure decoder as well as the cell-content decoder to improve the performance of the proposed model.

## ACKNOWLEDGEMENTS

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) Second Phase and “Big-data and AI-enabled Cyberspace Technologies” by New Energy and Industrial Technology Development Organization (NEDO).

## REFERENCES

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4945–4949. <https://doi.org/10.1109/ICASSP.2016.7472618>

Casado-García, A., Domínguez, C., Heras, J., Mata, E., & Pascual, V. (2020). The benefits of close-domain fine-tuning for table detection in document images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12116 LNCS, 199–215. [https://doi.org/10.1007/978-3-030-57058-3\\_15](https://doi.org/10.1007/978-3-030-57058-3_15)

Deng, Y., Kanervisto, A., Ling, J., & Rush, A. M. (2016). Image-to-Markup Generation with Coarse-to-Fine Attention. *34th International Conference on Machine*

*Learning, ICML 2017*, 3, 1631–1640. <https://doi.org/10.48550/arxiv.1609.04938>

Deng, Y., Rosenberg, D., & Mann, G. (2019). Challenges in end-to-end neural scientific table recognition. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 894–901. <https://doi.org/10.1109/ICDAR.2019.00148>

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>

Huang, Y., Yan, Q., Li, Y., Chen, Y., Wang, X., Gao, L., & Tang, Z. (2019). A YOLO-based table detection method. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 813–818. <https://doi.org/10.1109/ICDAR.2019.00135>

Itonori, K. (1993). Table structure recognition based on textblock arrangement and ruled line position. *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*, 765,766,767,768-765,766,767,768. <https://doi.org/10.1109/ICDAR.1993.395625>

Jimeno Yepes, A., Zhong, P., & Burdick, D. (2021). ICDAR 2021 Competition on Scientific Literature Parsing. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12824 LNCS, 605–617. [https://doi.org/10.1007/978-3-030-86337-1\\_40](https://doi.org/10.1007/978-3-030-86337-1_40)

Kayal, P., Anand, M., Desai, H., & Singh, M. (2021). ICDAR 2021 Competition on Scientific Table Image Recognition to LaTeX. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12824 LNCS, 754–766. [https://doi.org/10.1007/978-3-030-86337-1\\_50](https://doi.org/10.1007/978-3-030-86337-1_50)



- Kieninger, T. G. (1998). Table structure recognition based on robust block segmentation. *Https://Doi.Org/10.1117/12.304642*, 3305, 22–32. <https://doi.org/10.1117/12.304642>
- Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2019). *TableBank: A Benchmark Dataset for Table Detection and Recognition*. <https://doi.org/10.48550/arxiv.1903.01949>
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 07-12-June-2015*, 3431–3440. <https://doi.org/10.1109/CVPR.2015.7298965>
- Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., & Bai, X. (2021). MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117, 107980. <https://doi.org/10.1016/J.PATCOG.2021.107980>
- Ly, N. T., Nguyen, H. T., & Nakagawa, M. (2021). 2D Self-attention Convolutional Recurrent Network for Offline Handwritten Text Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12821 LNCS, 191–204. [https://doi.org/10.1007/978-3-030-86549-8\\_13/COVER](https://doi.org/10.1007/978-3-030-86549-8_13/COVER)
- MMCV Contributors. (2018). *{MMCV: OpenMMLab} Computer Vision Foundation*. <https://github.com/open-mmlab/mmcv>
- Nassar, A., Livathinos, N., Lysak, M., & Staar, P. (2022). *TableFormer: Table Structure Understanding with Transformers*. <https://doi.org/10.48550/arxiv.2203.01017>
- Qasim, S. R., Mahmood, H., & Shafait, F. (2019). Rethinking table recognition using graph neural networks. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 142–147. <https://doi.org/10.1109/ICDAR.2019.00031>
- Qiao, L., Li, Z., Cheng, Z., Zhang, P., Pu, S., Niu, Y., Ren, W., Tan, W., & Wu, F. (2021). LGPMA: Complicated Table Structure Recognition with Local and Global Pyramid Mask Alignment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12821 LNCS, 99–114. [https://doi.org/10.1007/978-3-030-86549-8\\_7/TABLES/4](https://doi.org/10.1007/978-3-030-86549-8_7/TABLES/4)
- Raja, S., Mondal, A., & Jawahar, C. v. (2020). Table Structure Recognition Using Top-Down and Bottom-Up Cues. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12373 LNCS, 70–86. [https://doi.org/10.1007/978-3-030-58604-1\\_5/FIGURES/8](https://doi.org/10.1007/978-3-030-58604-1_5/FIGURES/8)
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. <https://doi.org/10.48550/arxiv.1506.01497>
- Schreiber, S., Agne, S., Wolf, I., Dengel, A., & Ahmed, S. (2017). DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1, 1162–1167. <https://doi.org/10.1109/ICDAR.2017.192>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-December*, 5999–6009.
- Wang, Y., Phillips, I. T., & Haralick, R. M. (2004). Table structure understanding and its performance evaluation. *Pattern Recognition*, 37(7), 1479–1497. <https://doi.org/10.1016/J.PATCOG.2004.01.012>
- Ye, J., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., & Xiao, R. (2021). *PingAn-VCGroup's Solution for ICDAR 2021 Competition on Scientific Literature Parsing Task B: Table Recognition to HTML*. <https://doi.org/10.48550/arxiv.2105.01848>
- Zhang, Z., Zhang, J., Du, J., & Wang, F. (2022). Split, Embed and Merge: An accurate table structure recognizer. *Pattern Recognition*, 126, 108565. <https://doi.org/10.1016/J.PATCOG.2022.108565>
- Zheng, X., Burdick, D., Popa, L., Zhong, X., & Wang, N. X. R. (2021). Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 697–706. <https://doi.org/10.1109/WACV48630.2021.00074>
- Zhong, X., ShafieiBavani, E., & Jimeno Yepes, A. (2020). Image-Based Table Recognition: Data, Model, and Evaluation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12366 LNCS, 564–580. [https://doi.org/10.1007/978-3-030-58589-1\\_34/TABLES/3](https://doi.org/10.1007/978-3-030-58589-1_34/TABLES/3)