**lti-Cell Decoder and Mutual Learning for ble Structure and Character Recognition**

Takaya Kawakatsu1

Preferred Networks, Inc., 1-6-1 Otemachi, Chiyoda, Tokyo, Japan.

kat.nii.ac.jp@gmail.com

https://researchmap.jp/t.kat

**bstract.** Extracting table contents from documents such as scientific pers and financial reports and converting them into a format that can processed by large language models is an important task in knowledge

ormation processing. End-to-end approaches, which recognize not only ble structure but also cell contents, achieved performance comparable state-of-the-art models using external character recognition systems, d have potential for further improvements. In addition, these models n now recognize long tables with hundreds of cells by introducing local tention. However, the models recognize table structure in one direction m the header to the footer, and cell content recognition is performed ependently for each cell, so there is no opportunity to retrieve useful

ormation from the neighbor cells. In this paper, we propose a multi-cell ntent decoder and bidirectional mutual learning mechanism to improve e end-to-end approach. The effectiveness is demonstrated on two large

tasets, and the experimental results show comparable performance to te-of-the-art models, even for long tables with large numbers of cells.

**ds:** Deep Learning, Table Recognition, Transformer, Mutual Learning

**roduction**

ion retrieval technology which provides high-quality knowledge to large models (LLMs) is attracting attention. Many researchers have worked rting scanned and imaged documents to machine-readable formats such code [37,42,15] and LaTeX code [3,12]. This initiative has direct and benefits. First, because past literature remains mostly in printed form, ssary to convert it into structured electronic documents. This is a direct econd, the realization of intelligence which recognizes hidden meanings

out of published documents that are intended to be read by humans is t from the perspective of human-machine interaction. This is an indirect

is paper, we will focus on table recognition, which includes two types of mely structure recognition and cell content recognition. A simple table

ontal and vertical borders, and each cell contains characters. Complex

ay contain cells which are merged vertically or horizontally, and/or cells

olve invisible borders. Persons can understand table structure from the t even without explicit boundaries, and this is a challenging problem. ent years, following the success of Transformer [31] models in language

al recognition tasks, many methods [21,36,18,19] based on Transformer n proposed for the table recognition task. Since we can utilize an external haracter recognition (OCR) system to parse cell contents, we can focus n table structure recognition. This task exploits cross attention between

tures and embedded representations of HTML tokens to predict HTML quentially. In previous studies [21,36,18,19,20], the token prediction was d in one direction from the header to the footer, from left to right. This he opportunity to focus on the table structure ahead.

d Takasu [19] reported that end-to-end learning of table structure and nt recognition tasks may improve overall table recognition performance.

on, tables may contain more than a few hundred cells, and the sequential n approaches may suffer from poor performance, which can be improved ucing local attention [18]. These previous studies performed cell content

on independently for each detected cell after structure recognition. This he opportunity of obtaining useful information from neighbor cells.

solution to the problems, we improve the end-to-end approach [18,19,20]

ose a method that refers to the recognition results of neighbor cells and g mechanism focusing on both previous and following cells. The former d by introducing a cell decoder that infers multiple cells and configuring hical decoder along with an HTML decoder for structural recognition. r is achieved by mutual learning [38] between a forward decoder which table structure from left to right and a backward decoder which reads structure from right to left. The effectiveness of the proposed method

strated using two large-scale tabular image datasets.

ain contributions of this paper are: 1) We propose a cell decoder which ltiple cells and obtains useful information from surrounding cells. 2) We a bidirectional mutual learning mechanism to force the proposed model tention to both previous and following cells. 3) Across all experimental ur proposed method achieved better performance than state-of-the-art

**lated Work**

l, the table recognition task is performed by two subtasks, namely table recognition and cell content recognition. Of course, the final output is

L [37,42,15] or LaTeX [3,12] document, so there is no need to distinguish the two subtasks. However, it is better to recognize tags (or commands) r visible characters using separate models. For cell content recognition

ting highly accurate OCR systems [17] are available, and thus previous 1,13,32,26,28,30] have mainly focused on table structure recognition.

structure recognition has been studied for a long time, and approaches hand-crafted features and heuristic rules [11,13,32] were proposed, but lication was limited to simple tables or tables with predefined patterns. development of deep learning, methods that automatically learn table

l patterns [39,27,26,28,30] have become mainstream. These studies can d into approaches based on object detection and segmentation [30,27],

oaches based on sequential token prediction [21,36].

e detection and segmentation approaches, Schreiber et al. [30] proposed d system using Faster R-CNN [29] and fully convolutional networks [16] table detection and table structure recognition. Raja et al. [28] proposed

ge model that estimates the relationships between cells after recognizing tions. Qiao et al. [27] won his first place in the ICDAR competition [37] ning text, cell, row and column recognition tasks using Mask R-CNN [5].

e sequential token prediction approaches, a simple image caption model ilized for cell detection because the order of cells is uniquely determined. [36] and Nassar et al. [21] proposed Transformer models with two types

ers for table structure recognition and cell localization. Peng et al. [25] performance comparable to a model using a deep convolutional encoder

nificantly reducing parameters by introducing a convolutional stem.

e 2020s, researchers are investigating end-to-end models that learn both ucture and cell content recognition tasks [3]. Zhong et al. [42] proposed that uses a ResNet [6] encoder and two LSTM [8] decoders to recognize le structure and cell contents, but its performance was inferior to models

ernal OCR.

d Takasu [19] proposed a multi-task model that detects table structure, ions, and cell contents. Their model uses a ResNet encoder with global ttention [2] and two Transformer decoders. The first decoder infers the

okens sequentially, and then the second decoder reads the cell contents e. This model achieved performance comparable to the models utilizing OCR. They also proposed weakly supervised learning to reduce the cost ing bounding box training data [20] and introduced local attention [1]

vely recognize tables with a large number of cells [18].

21, the scientific literature parsing competition [37] was held at ICDAR e competition consisted of document layout recognition task A and table

on task B. Task B required converting table images to HTML tags with nts. The PubTabNet [42] dataset and the final evaluation dataset were for this task. The training dataset consists of HTML tokens, cell texts,

bounding boxes. There were 30 submissions from 30 teams and most of 0 solutions exploited separate OCR models, additional annotation and

techniques.

ecSet [35] is a bilingual dataset containing rotated and distorted tables

otographs for three tasks, namely table detection, structure recognition, content recognition. Detection of such tables is outside the scope of this

**ckground**

o previous work [18,19,20], our proposed model uses a ResNet encoder TML decoder consisting of multiple attention blocks [31] to infer HTML presenting table structure. An additional decoder is exploited to infer nts. The encoder and two decoders are trained simultaneously using an d approach. In this section, we introduce some techniques used by the

method described in Section 4.

**coder**

studies [21,36,18] used a convolutional neural network (CNN) to extract tures and fed them to the decoder. CNN is useful for recognizing small rs while preserving locality such as character positions, reducing the size features, and improving the computational efficiency and performance

coder.

umber of convolutional layers contributes to recognition performance, y derivatives have been explored to increase it. ResNet [6], which consist number of residual blocks of multiple convolutional layers with simple ections, has been commonly used. In addition, ResNeXt [34] with group ion and DenseNet [9] with more complicated skip connections between

lutional layers were proposed.

f the weaknesses of CNN is its poor ability to recognize global context ng strongly on local features. As a solution, a global context attention lock [2] was proposed, defined by Eq. (1).

xij + W3 max(0, LayerNorm(W2 SoftMax(W1xij) xij)), (1)

i j

are the pixel indices, x, y are the input and output pixels, respectively. W3 are the weight matrices of three linear layers. LayerNorm means layer

ation. The softmax function is defined as follows.

SoftMax(zij) =

exp zij

exp zmn

m n

, (2)

, n are the pixel indices. The GCA block should be placed between some (or dense) blocks.

**coder**

mer [31] achieves superior performance in both language modeling and cognition tasks. Compared to recurrent neural networks including long m memory [8], a Transformer itself does not involve recursion, allowing

rocessing of sequential input and output data. It should be noted that , sequential inference is performed unless the prediction length is fixed.

, Transformer does not require recursion to recognize the context of the , avoiding vanishing gradients and providing better performance.

ey idea of Transformer is called scaled dot product attention. Let X be ce of length lx and dx channels, and Y be another input sequence. For tion, X, Y are the same sequence, and the Transformer pays attention parts of X in processing X. For cross attention, X and Y are in different and the Transformer pays attention to some parts from Y in processing mechanisms allow Transformer to learn the context of sequential data

relationship between visual and language domains.

ttention layer first generates a query Q, key K, and value V from X, Y d in Eq. (3).

qi = WQ xi,

kj = WK yj, vj = WV yj,

(3)

are the sequence indices, q, k, v, x, y are the elements of Q, K, V, X, Y , ely. WQ, WK, WV are the projection matrices.

utput Z of the attention layer is defined by Eq. (4).

QK⊤

Zi = WZ SoftMax(√

dk

)V, (4)

Z is the output projection matrix, and dk is the dimension of k. This chanism of the scaled dot product attention [31].

actice, the attention layer is divided into several groups, each of which ntion independently and combines the outputs at the end. This is called ad attention [31]. Through the above mechanism, Transformer can focus

c values of Y and incorporate them into the X series.

**cal Attention**

Transformer has superior ability to recognize long sequences compared ent neural networks, it is still known to perform poorly upon extremely ences. Local attention (LA) [1] is a technique designed to handle such

ences in Transformer.

ij be a mask to focus on the jth element from ith element of X. The f the local attention layer is defined by Eq. (5), involving causal masking

t leakage from subsequent elements.

QK⊤

√

Zi = WZ SoftMax(

dk

k matrix M is given by Eq. (6).

+ Mij)V. (5)

Mij =

0 0≤i−j≤w,−∞ otherwise,

(6)

j are the sequence indices, and w is the width of the sliding window.

**sitional Encoding**

mer [31] itself have poor ability to know the position of each element in nces, and the position information must be provided explicitly. Instead ing simple position values, two approaches have been proposed, namely l embedding [4] and positional encoding [31]. In general, the latter works

small training datasets.

utput p(n) of positional encoding for the index n is defined by Eq. (7).

 . 

 

sin 10000n 2dk 

d

 p(n) = 

, where k∈ 0,

. (7)



cos 10000n 2dk  .... 

2

st be added directly to the feature vector x(n) with d channels. sequence X has two-dimensional positions (i, j), 2D positional encoding by Zhao et al. [40] may be a better choice. It normalizes the horizontal ical coordinates to [0, 1], encodes each with Eq. (7), and then combines btain a single vector. The positional encoding of the i, jth pixel is given

).

p Hi



p2D(i, j) =p Wj 

, (8)

, W are the height and width for positional normalization. In this paper, ed this normalization.

**utual Learning**

e learning is commonly used to improve machine learning generalization nce and fitting accuracy by averaging or complementing the outputs of inference models. However, it is computationally more expensive than

dels due to the large number of parameters especially for deep learning es.

hieve similar effects using only a single model, knowledge distillation [7] n alternative solution. This is a technique which uses a large, complex twork, i.e., an ensemble model, as a teacher and a small, simple model

ent to obtain higher performance than simply training a student model ground-truth data.

al learning [38] may be another solution. Here, multiple student models ed simultaneously to teach each other, without training a teacher model ce. In particular, each student model performs supervised learning using ruth data and minimizing Kullback–Leibler (KL) divergence [14] so that ibutions of each other’s classification outputs match.

**etrics**

al. [42] introduced a tree edit distance based similarity (TEDS) metric rmance evaluation of both table structure and cell content recognition. verting the recognition results and the ground truth into tree structures

tags, the TEDS score is calculated according to Eq. (9).

TEDS(Ta, Tb) = 1−

EditDist(Ta, Tb) max(|Ta|, |Tb|)

, (9)

and Tb are the HTML trees, EditDist is the edit distance function, and number of nodes in T .

are two versions of TEDS, namely structural TEDS and total TEDS. er is calculated for HTML trees excluding cell contents and represents nition performance for table structures only. The latter is computed on HTML trees including cell contents and indicates the total recognition

nce.

dition, Zhong [42] classified the tables into two subsets, namely simple d complex tables. The former are tables without cells which are merged

or horizontally, and the latter are the other tables.

**posal**

osal consists of a ResNet encoder and two local-attention Transformer

. The two decoders infer table structure and cell contents, respectively. ional output layer estimates the cell bounding boxes.

wo main differences with previous studies [18,19] are 1) introduction of ell decoder, and 2) introduction of bidirectional mutual learning to the ecoder. In addition, 2D positional encoding is employed. We named the

method MuTabNet after mutual learning, multi-task learning, and the l decoder. Fig. 1 shows the network architecture.

**coder**

der consists of a CNN backbone and 2D positional encoding. The CNN image features of 65x65 pixels from an image of 520x520 pixels. For the adopted TableResNetExtra [36] with 26 convolutional layers and three cks. After 2D positional encoding, the image features are flattened into

nsional features with 512 channels for cross-attention at the decoders.

**ML Decoder**

L decoder consists of one embedding layer, three local attention blocks,

output layers. Each attention block accepts a table structure sequence a self-attention layer in the block. The attention block then incorporates tures into the table structure sequence through a cross-attention layer,

thead tr td /td td C h [SEP] P h

Flatten

Linear & Softmax Add & Layer Norm Feed Forward

Add & Layer Norm Global Attention

Image Feature

Add & Layer Norm

Local Attention

Linear & Softmax Add & Layer Norm Feed Forward

Add & Layer Norm Global Attention

Image Feature

Add & Layer Norm

Local Attention

Backbone x3 Blocks x1 Block

to 520x520

Phonetic Code

Alpha Bravo Charlie

ble Image

backbone.

Linear

LtoR LtoR LtoR LtoR LtoR [SOS] thead tr td /td

Embedding

[SOS] thead tr td /td

(b) HTML decoder.

Linear

td#1 td#1 td#1 td#2 td#2 [SOS] C h [SEP] P

Embedding

[SOS] C h [SEP] P

(c) Cell decoder.

Fig. 1: Proposed network architecture.

uts the sequence through a feed-forward layer. Several skip connections normalizations are inserted within the block. The output from the last block is converted into HTML tokens and cell bounding boxes by the

ut layers.

g training, the decoder predicts left or right shifted HTML tokens from HTML tokens. The shift direction is specified by an additional one-hot uring inference, the decoder predicts the following token and iteratively

the input sequence to obtain the complete HTML sequence.

dition to HTML tokens, the decoder accepts some special tokens, namely , and PAD. SOS is a token that triggers sequential inference and is inserted ginning of the tokens. EOS is a token that stops inference and is inserted d of the tokens. PAD is inserted after EOS to equalize the lengths of the

the mini batch.

ing previous studies [18,19], the HTML sequence was simply tokenized L tags, except for the <td> tag representing the start of a cell. A <td> enized as‘<td’,‘colspan="2"’,‘rowspan="3"’,‘>’if it contains colspan

an attributes. Otherwise, the tag is simply tokenized as‘<td>’.It should that FinTabNet [41] and PubTabNet [42] described in Section 5.1 are available with such tokenization applied. We then merged the <td> and

tely following </td> tokens into one token.

ermore, we assigned some special tokens to frequent sequence patterns. der cells in the dataset had bold text, we removed the <b> and completed t-processing. These methods follow previous studies [18,19].

**ll Decoder**

decoder consists of one embedding layer, one local attention block, and t layer. Following previous studies [18,19], the embedding layer accepts cters one by one. This is because cell contents typically consist of short

s or unknown words or numbers, making it difficult to utilize pretrained models.

asic structure of a cell decoder is similar to that of an HTML decoder, following differences. First, a special token SEP is inserted between cell to trigger movement to the next cells. Second, the cell decoder accepts ation of cell contents and their corresponding HTML features extracted

output of the HTML decoder. These improvements allow the proposal tially read the contents of multiple cells while referring to information

vious cells.

revious study [18] exploited local attention for the HTML decoder and tention for the cell decoder. This was because the cell decoder processed independently, and the cell contents were short in general. On the other

our proposed multi-cell decoder, the sequence of cell contents tends to Consequently, we employed local attention.

**directional Mutual Learning**

ose bidirectional mutual learning inspired by deep mutual learning [38] the HTML decoder. Here, two equivalent decoders are trained together t table structure in either a left-to-right (LtoR) or right-to-left (RtoL)

. To reduce model parameters, we implemented the mutual learning in ecoder by combining an additional one-hot vector that determines the with the embedded HTML tokens.

and←−x be the LtoR and RtoL sequences respectively, and let p(x) be

→−

nd-truth and q(x) the predicted probabilities. The loss L for the LtoR

is defined by Eq. (10).

→− 1 N −→ −→ 1 N ←−

q(←−xn)

L =−

N

p(xn) log q(xn) +

N

n=1

n=1 q(xn) log q(−→xn)

. (10)

Table 1: The statistics of the table image datasets.

Training Validation Evaluation

et 91,596 10,635 10,656 Net 500,777 9,115 9,064 Net250 114,111 2,161 -

**eriments**

ate the effectiveness of the multi-cell decoder and bidirectional mutual we conducted experiments on two public table datasets.

**tasets**

ed two large datasets, FinTabNet [41] and PubTabNet [42]. In addition, a subset named PubTabNet250 [18] for ablation studies. Table 1 shows stics for the datasets.

**et** is a large dataset of table images, including HTML labels and cell boxes, extracted from the annual reports of S&P 500 companies. The

ontains 112k tables and is divided into training set, validation set, and n set. It should be noted that the original FinTabNet confuses validation ation sets. Following previous studies [21,41], we treated the *validation*

ining 10,656 images as the evaluation set.

**Net** is a dataset built by collecting scientific articles from the PubMed pen access subset, containing 568k tables and corresponding structure content annotations and cell bounding boxes. PubTabNet provides the and validation sets, and the evaluation set was provided for the ICDAR

ion [37]. We classified the tables into simple tables and complex tables bed in Section 3.6.

**Net250** Ly and Takasu [18] extracted tables with 250 or more HTML

om PubTabNet and created a subset named PubTabNet250. They also ed subsets for tables containing at least 500, 600, and 700 tokens. These ere utilized originally [18] to demonstrate the effectiveness of the local mechanism. We also utilized these subsets to conduct ablation studies

n 5.4, approximately reducing training time from 179 hours to 45 hours el.

able 2: Table recognition results on FinTabNet evaluation set.

TEDS (%)

Structure Total [42] 90.60 -

[41] 87.14 - T) [41] 91.02 - rmer [21] 96.80 -

[10] 98.63 **98.21**

. [20] 98.72 95.32 Takasu [19] 98.79 - Takasu [18] 98.85 95.74

et **98.87** 97.69

**plementation**

osed model was implemented in PyTorch using mmcv [22], mmdet [23], cr [24] frameworks and trained on four NVIDIA V100 GPUs with batch total. We used Ranger [33] optimizer. The learning rate was initialized for the first 25 epochs, and decreased to 0.0001 and 0.00001 for the next

last two epochs, respectively.

tabular image was normalized and reduced to 520x520 pixels, padding ins with zeros if necessary. The cell bounding boxes were normalized to inimum value of 0 and a maximum value of 1.

L tokens and cell contents were converted to 512-dimensional embedded ations. The four attention blocks in the HTML and cell decoders have 8-head, 512-channel architecture, and the sliding window size for local was set to 300 by default, following previous work [18]. The maximum

or table structure sequences and cell content sequences were set to 800 , respectively, including special tokens. We employed greedy search for

l prediction.

sure a fair comparison with the previous studies, we did not utilize data ation or ensemble learning techniques. We also did not take advantage

topping.

**perimental Results**

ared the performance of the proposed model trained on FinTabNet and et with the claimed performance of existing models.

**et** We evaluated the experimental results of structure recognition and

ognition using the TEDS metric. Table 2 compares the TEDS scores in

able 3: Table recognition results on PubTabNet validation set.

TEDS (%)

Simple Complex Total [42] 91.20 85.40 88.30

ct-Net [28] - - 90.10 rmer [21] 95.40 90.10 93.60 [39] 94.80 92.50 93.70

&OCR [27] - - 94.60 p [36] - - 96.26 p&ME [36] - - 96.84

[10] - - 96.31

. [20] 97.89 95.02 96.48 Takasu [19] 97.92 95.36 96.67 Takasu [18] 98.07 95.42 96.77

et **98.16 95.53 96.87**

et between the proposal and previous models. The proposal outperforms

ous models with scores of 98.87% and 97.69%. The inference time using Us was 3.78 hours.

otal TEDS score of the proposal was lower than the score of VAST [10], uld be explained by the fact that VAST exploits external OCR for cell ecognition. In contrast, the structural TEDS score of VAST was lower

se of end-to-end approaches [20,19,18], including the proposal.

**Net** We evaluated the experimental results of table recognition on the n set using the TEDS metric. Table 3 compares the scores between the and previous methods. The proposal outperforms all previous methods es of 98.16%, 95.53% and 96.87% on simple tables, complex tables, and

, respectively. The inference time using the 4 GPUs was 3.23 hours.

lso evaluated our proposal on the evaluation set. Table 4 compares the the proposal with the top 10 solutions of the ICDAR competition [37]. scores achieved on both sets indicate high generalization performance

oposal. The inference time was 3.13 hours.

core of the proposal was higher than the score of VAST [10]. PubTabNet a large amount of training data, and the proposed model appears to be

ed for cell content recognition tasks.

uld be noted that VCGroup&ME [36] utilized additional annotation of boxes of text lines within cell contents and ensemble learning of three he proposed model outperforms all other non-end-to-end models which

able 4: Table recognition results on PubTabNet evaluation set.

TEDS (%)

Simple Complex Total [37] 97.18 92.40 94.84 [37] 96.95 93.43 95.23

AI [37] 97.35 93.79 95.61 [37] 97.30 93.93 95.65 [37] 97.39 93.87 95.66 [37] 97.38 94.79 96.11 [39] 97.60 94.89 96.27

p [36] 97.90 94.68 96.32 ab-OCR [37] 97.88 94.78 96.36

. [20] 97.51 94.37 95.97 Takasu [19] 97.60 94.68 96.17 Takasu [18] 97.77 94.58 96.21

et **98.01 94.98 96.53**

dditional annotation and ensemble learning even though our model did e such techniques.

**lation Studies**

ucted additional experiments for ablation studies using PubTabNet250

or training and PubTabNet subsets for evaluation.

**eness of Multi-Cell Decoder and Mutual Learning** We evaluated iveness of the proposed methods, namely multi-cell (MC) decoder and nal mutual learning (BML). We trained two models on the training set lated the validation scores as displayed in Table 5. We selected previous

ntal results [18] as baselines using exactly the same model architecture set except for MC and BML. LA in the table refers to local attention.

the previous study [18] focused on performance for long tables, we also d TEDS scores for tables containing at least 500, 600, and 700 structure he MC decoder outperforms the baselines at all table lengths, and BML

mproves table recognition performance.

ffect of BML was unclear in the structural TEDS scores but evident in TEDS scores. BML may still have improved the performance of implicit recognition and may have made an impact on cell content recognition,

quires precise content locations.

Table 5: Table recognition results with the proposed methods.

TEDS (%)

thods Structure Total

C BML 250+ 500+ 600+ 700+ 250+ 500+ 600+ 700+ - - - - - - 93.86 91.16 90.63 88.65 - - - - - - 94.28 92.99 91.29 89.61

- 96.60 96.71 96.75 96.67 95.02 94.59 93.73 93.14✓ 97.02 96.70 96.35 96.65 95.81 95.11 94.05 94.02

able 6: Table recognition results with respect to window sizes.

TEDS (%)

Structure Total

250+ 500+ 600+ 700+ 0+ 250+ 500+ 600+ 700+ 96.96 96.69 96.98 96.60 75.91 95.70 95.19 94.99 94.35 96.79 96.53 96.30 95.83 75.79 95.46 94.66 93.80 92.69 97.02 96.70 96.35 96.65 83.15 95.81 95.11 94.05 94.02 96.83 96.85 96.48 96.51 82.58 95.40 95.08 94.00 93.50 96.97 96.74 97.03 96.54 81.14 95.51 94.46 93.88 92.65

**Size of Cell Decoder** Ly and Takasu [18] has reported that a window 0 was optimal for the HTML decoder, whereas the cell decoder exploited tention. In this study, we determine the optimal window size for the MC

Table 6 shows the change in TEDS scores for the validation set as the ize varies from 100 to 500, while the window size for the HTML decoder

at 300.

eral, a window size of 300 achieved the highest score, with the exception containing more than 500 tokens, where a window size of 100 achieved st score. Tables with many cells tend to have fewer characters per cell,

orter window may be sufficient.

uld be noted that we used the PubTabNet250 dataset for training, and rmance for tables with fewer structure tokens was lower than the scores 3. We selected the window size of 300 as the optimal value for the entire

et dataset containing tables with fewer tokens from the perspective of ation performance.

**nclusion**

oved an end-to-end table recognition model based upon Transformer to erformance comparable to state-of-the-art models using external OCR

The proposed model consists of a ResNet encoder and two decoders for recognition and cell content recognition. After the first decoder infers

ture tokens, the second decoder reads the text within each cell. roposed a multi-cell decoder for cell content recognition to exploit useful ion from neighbor cells. Furthermore, we proposed bidirectional mutual to force the model to pay attention to both previous and following cells. ental results using two public datasets demonstrate the effectiveness of

osed methods.

ure work, we will further consider multitasking models that include the ecognizing the meaning of tables, which enables deep understanding of

ocuments, including table contents, and provides high-quality scientific e for LLMs and question-answering systems.

**nces**

gy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer

)

Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: Non-local networks meet squeeze- tion networks and beyond. In: IEEE/CVF International Conference on Com- Vision Workshop (ICCVW). pp. 1971–1980 (2019)

, Y., Rosenberg, D., Mann, G.: Challenges in end-to-end neural scientific table nition. In: International Conference on Document Analysis and Recognition AR). pp. 894–901 (2019)

ing, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional nce to sequence learning. In: International Conference on Machine Learning L). pp. 1243–1252 (2017)

., Gkioxari, G., Doll´ar, P., Girshick, R.: Mask R-CNN. In: IEEE International rence on Computer Vision (ICCV). pp. 2980–2988 (2017)

., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. EE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 78 (2016)

n, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network

)

reiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 80 (1997)

g, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolu- l networks. In: IEEE Conference on Computer Vision and Pattern Recognition

R). pp. 2261–2269 (2017)

g, Y., Lu, N., Chen, D., Li, Y., Xie, Z., Zhu, S., Gao, L., Peng, W.: Improving structure recognition with visual-alignment sequential coordinate modeling. EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

1134–11143 (2023)

ri, K.: Table structure recognition based on textblock arrangement and ruled

osition. In: International Conference on Document Analysis and Recognition AR). pp. 765–768 (1993)

l, P., Anand, M., Desai, H., Singh, M.: ICDAR 2021 competition on scien- able image recognition to LaTeX. In: International Conference on Document

sis and Recognition (ICDAR). pp. 754–766 (2021)

nger, T.G.: Table structure recognition based on robust block segmentation. hotonics West ’98. vol. 3305, pp. 22–32 (1998)

ack, S., Leibler, R.A.: On information and sufficiency. The Annals of Math- ical Statistics **22**(1) (1951)

., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: TableBank: Table benchmark age-based table detection and recognition. In: Language Resources and Eval-

n Conference (LREC). pp. 1918–1925 (2020)

, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic entation. In: IEEE Conference on Computer Vision and Pattern Recognition

R). pp. 3431–3440 (2015)

., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: MASTER: Multi- t non-local network for scene text recognition. Pattern Recognition **117**,

0 (2021)

.T., Takasu, A.: An end-to-end local attention based model for table recog-

. In: International Conference on Document Analysis and Recognition (IC- ). pp. 20–36 (2023)

.T., Takasu, A.: An end-to-end multi-task learning model for image-based recognition. In: International Joint Conference on Computer Vision, Imaging

omputer Graphics Theory and Applications (VISIGRAPP). pp. 626–634

)

.T., Takasu, A., Nguyen, P., Takeda, H.: Rethinking image-based table recog- using weakly supervised methods. In: International Conference on Pattern

gnition Applications and Methods (ICPRAM). pp. 872–880 (2023)

r, A., Livathinos, N., Lysak, M., Staar, P.: TableFormer: Table structure standing with transformers. In: IEEE/CVF Conference on Computer Vision attern Recognition (CVPR). pp. 4604–4613 (2022)

MMLab: MMCV, https://github.com/open-mmlab/mmcv

MMLab: MMDetection, https://github.com/open-mmlab/mmdetection MMLab: MMOCR, https://github.com/open-mmlab/mmocr

, A., Lee, S., Wang, X., Balasubramaniyan, R., Chau, D.H.: High-performance formers for table structure recognition need early convolutions. In: Table Rep- tation Learning Workshop (2023)

d, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: CascadeTabNet:

pproach for end to end table detection and structure recognition from image- documents. In: IEEE/CVF Conference on Computer Vision and Pattern

gnition Workshops (CVPRW). pp. 2439–2447 (2020)

L., Li, Z., Cheng, Z., Zhang, P., Pu, S., Niu, Y., Ren, W., Tan, W., Wu, F.:

A: Complicated table structure recognition with local and global pyramid alignment. In: International Conference on Document Analysis and Recog-

(ICDAR). pp. 99–114 (2021)

S., Mondal, A., Jawahar, C.V.: Table structure recognition using top-down

ottom-up cues. In: Computer Vision–ECCV. pp. 70–86 (2020)

S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object tion with region proposal networks. In: Advances in Neural Information Pro-

g Systems. vol. 28 (2015)

iber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: DeepDeSRT: Deep learning tection and structure recognition of tables in document images. In: Interna-

l Conference on Document Analysis and Recognition (ICDAR). pp. 1162–1167

)

ani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser,

olosukhin, I.: Attention is all you need. In: Advances in Neural Information ssing Systems. vol. 30, pp. 6000–6010 (2017)

, Y., Phillips, I.T., Haralick, R.M.: Table structure understanding and its rmance evaluation. Pattern Recognition **37**(7), 1479–1497 (2004)

t, L.: Ranger - a synergistic optimizer (2019), https://github.com/ 2020/Ranger-Deep-Learning-Optimizer

S., Girshick, R.B., Doll´ar, P., Tu, Z., He, K.: Aggregated residual transfor- ns for deep neural networks. In: IEEE Conference on Computer Vision and

rn Recognition (CVPR). pp. 5987–5995 (2017)

, F., Hu, L., Liu, X., Huang, S., Gu, Z.: A large-scale dataset for end-to-end recognition in the wild. Scientific Data **10**(1), 110 (2023)

., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., Xiao, R.: PingAn-VCGroup’s ion for ICDAR 2021 competition on scientific literature parsing task B: Table

nition to HTML (2021)

s, A.J., Zhong, P., Burdick, D.: ICDAR 2021 competition on scientific litera- arsing. In: International Conference on Document Analysis and Recognition

AR). pp. 605–617 (2021)

g, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In:

/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 320–4328 (2018)

g, Z., Zhang, J., Du, J., Wang, F.: Split, embed and merge: An accurate table ture recognizer. Pattern Recognition **126**, 108565 (2022)

, W., Gao, L., Yan, Z., Peng, S., Du, L., Zhang, Z.: Handwritten mathematical ssion recognition with bidirectionally trained transformer. In: International rence on Document Analysis and Recognition (ICDAR). pp. 570–584 (2021)

g, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor ): A framework for joint table identification and cell structure recognition visual context. In: IEEE Winter Conference on Applications of Computer

n (WACV). pp. 697–706 (2021)

g, X., ShafieiBavani, E., Yepes, A.J.: Image-based table recognition: Data, l, and evaluation. In: Computer Vision–ECCV. pp. 564–580 (2020)