2021 RIVF International Conference on Computing and Communication Technologies (RIVF) | 978-1-6654-0435-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/RIVF51545.2021.9642121

MC-OCR Challenge 2021: An end-to-end recognition framework for Vietnamese Receipts

Hung Le, Huy To, Hung An,
Khanh Ho, Khoa Nguyen
Faculty of Computer Science
University of Information Technology
VNU-HCMC, Vietnam
{18520797, 18520855, 17520531, 19520624, 18520929}@gm.uit.edu.vn

Thua Nguyen, Tien Do,
Thanh Duc Ngo, Duy-Dinh Le
Faculty of Computer Science
University of Information Technology
VNU-HCMC, Vietnam
{thuann,tiendv,thanhnd,duyld}@uit.edu.vn

***Abstract*—Recognizing text from receipts is a significant step in automating office processes for many fields such as finance and accounting. MC-OCR Challenge has formed this problem into two tasks (1) evaluating the quality, and (2) recognizing required fields of the captured receipt. Our proposed framework is based on three key components: preprocessing with receipt detection using Faster R-CNN, alignment based on the angle and direction of rotation; estimate the receipt image quality score in task 1 using EfficientNet-B4 which has been retrained using transfer learning; while PAN is for text detection and VietOCR [1] for text recognition. In the final round, our systems have achieved the best result in task 1 (0.1 RMSE) and a comparable result with other teams (0.3 CER) in task 2 which demonstrated the effectiveness of the proposed method.**

***Index Terms*—Deep learning, OCR, Receipt**

I. Introduction

Automated information extraction obtained from receipts plays a critical role in digital transformation in finance and accounting companies. Data extracted from receipts not only supports automatic payment but also helps obtaining and storing information efficiently. However, extracting information from mobile-captured receipts faces many challenges such as the quality of the photos taken with mobile devices and the complexity of the information field to be extracted from the receipt. To address them, MC-OCR Challenge [1] formed the problem into two tasks including (1) evaluating the quality of the captured receipt measured by the ratio of text lines associated with the “clear” label evaluated by human annotators, and (2) recognizing four predefined fields from the receipt.

To approach to this contest, we have built an end-to-end framework to score receipt quality and extract necessary information of receipt including the seller, timestamp, address, total (Figure 1). For input images, the system will crop the part which contains the receipt using Faster R-CNN [2]. At the same time, the system uses a classifier using EfficientNet-B7 [3] to identify rotated images or align skewed images. The preprocessing step is to ensure that data processed in the next step with the best quality possible. For Task 1, to evaluate the quality of the input image, we treat this task as a regression problem. From the preprocessed data, We use the transfer learning method to build a model based on the EfficientNet-B4 network architecture with various configurations to estimate the image quality level. To extract the information fields required by Task 2, the system uses PAN (Pixel Aggregation Network) [4] detection method which has been trained on the SROIE [5] dataset to locate the text-containing areas. Then, VietOCR is used to identify text content. And finally, we adopt the rule-based approach, in conjunction with some NLP techniques to identify content that is relevant to the information fields we need to extract.

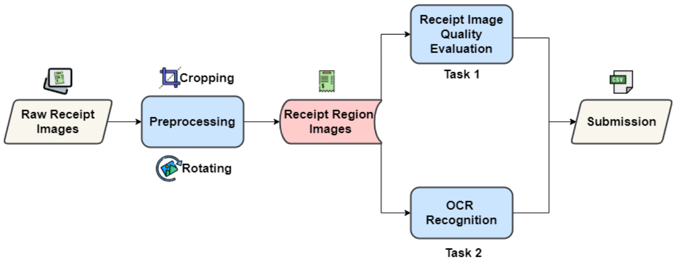

Fig. 1: The framework architecture.

The results on the final round showed that for Task 1 we ranked Top 1 with 0.1 RMSE (Root Mean Square Error). Besides, 0.3 with CER (Character Error Rate) measure is the result of task 2, it is not much different from the top 1 team with an accuracy of about 0.23. Experimental results demonstrate the effectiveness of the proposed methods. We also provide source code [2] and end-to-end live demo [3] for community.

II. Related Works

The system of extracting information from receipt is a complex system by combining many sub-problems. Based on the common pipeline of the system, we can divide related research into two main modules: Text Reading and Information Extraction

[1] VietOCR https://github.com/pbcquoc/vietocr

[2] Source Code https://github.com/tiendv/MCOCR2021

[3] Demo http://service.aiclub.cs.uit.edu.vn/receipt/

978-1-6654-0435-8/21/$31.00 ©2021 IEEE

A. Text Reading

The main goal of the text reading step is to locate the position of the text in the given image and recognition the text. Therefore, Text Detection and Text Recognition are two sub-tasks.

1) Text Detection: In recent years, along with the achievements of deep learning, approaches are commonly based on the CNN framework with two major categories: anchor-based methods and segmentation-based methods. For detail, Anchor-based methods [6] [7] [8] related to object detection methods, which could predict the existence of texts in the input image and return bounding boxes. Wang, Wenhai, et al [9] gave some brief descriptions of some recent typical methods [10] [11]. Wang, Wenhai, et al. [4] mentioned the tradeoff between heavy framework for high accuracy and simple structure to maintain a good balance between speed and accuracy among various methods such as TextBoxes [10], RRD [12], SSTD [6] and PAN [4]. Moreover, Tesseract OCR engine, Amazon Textract, Google vision API, are also already developed to support this task. However, they occasionally may not productively handle curve texts. Hence, another approach can detect the texts notwithstanding their orients and shapes called segmentation-based methods received more attention. As follows, according to He, Pan, et al. [6], the previous works generally concentrate on the bottom-up approach which uses hand-crafted features or sliding window methods to identify text regions in the image. They commonly include pixel-level binary text/non-text filter, thereby, multiple bottom-up steps are designed to put text-related pixels together into characters, then, characters to words, words to the text line. Nevertheless, mentioned bottom-up approaches are developed based on heuristic or hand-crafted features, and the application of low-level features here is not robust enough. Recently, thanks to the development of deep learning, segmentation-based methods are mainly inspired by fully convolutional networks (FCN) [13], which learn the pixel-level classification tasks to separate text regions apart from the background and achieved great progress [9] [11].

2) Text recognition: Recently, with the development of machine learning and deep learning, learning-based methods have become a new approach. Recurrent Neural Networks (RNNs) and CNN-based methods was combined with and introduced by [14] as CRNN framework, in addition, existing CTC decoder was replaced by the attention mechanism in this work. Thenceforward, many researchers have experimented and applied CRNN for their methods [15] [10]. After that, the TransformerOCR model has many advantages compared to the architecture of the CRNN model and the VietOCR library has taken advantage of that. Especially, this library is built for Vietnamese OCR problems which are complicated because of complex alphabetical structures and grammars.

B. Information Extraction

Information extraction is a crucial part of many fields, referring to the automatic extraction of structured information from unstructured or semi-structured documents. [16] [17] [18] deal with this task by position location information, considering the layout, and then operating on the word segmentation or reconstructed character of the printed file.

III. Methodology

In this competition, we have proposed an end-to-end framework as we visualized in Figure. 1. The preprocessing module is to generate the receipt region images. Then, we utilize those output to be the input of both task 1 and task 2 modules. That saves processing time than dealing with the preprocessing step in each module individually. After that, we combine the output of task 1 module with the output of task 2 module to create the final submission.

A. Data Preprocessing

Data pre-processing plays an important role in determining the accuracy of the whole system. Through observing the contest data, we found that there is a significant receipt amount that is not the main component in the image or the invoice is rotated or skewed. Therefore, receipt detection for cropping as well as receipt alignment are necessary step before the data is processed in the next stages.

1) Receipt Detection: To detect the receipt in the image we have prepared a data set for training some detectors. The dataset consists of 664 images, each image was labeled the position of 4 corners (top left, top right, bottom right, bottom left) as well as the receipt itself. (Figure 2)

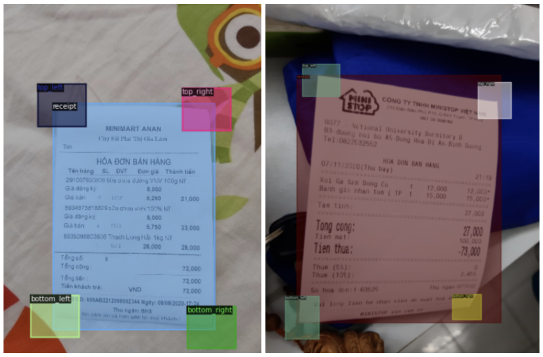

Fig. 2: An example of our receipt labeling.

We have evaluated the effectiveness of popular methods in the literature such as RCNN and YOLO (You Only Look Once) [19] with a wide variety of configurations. As the result produced by this subtask heavily affects the performance of our proposed system, we suggests choosing Faster RCNN using Resnet101 backbone with Feature Pyramid Network since this method offers high enough precision while also maintaining low execution time. (Table I)

Model	mAP@0.5-0.95	FPS
Faster-RCNN-R-101-FPN	0.779	6.6
Mask-RCNN-R-101-FPN	**0.787**	5.4
YOLOv3	0.739	**25.6**
YOLOv4	0.731	22.7

TABLE I: Results evaluated some receipt detector.

2) Receipt Alignment:

In practice, mobile-captured receipts may not be properly aligned along the image edges as they can fall into either one of two cases: **a)** slightly rotated or **b)** completely rotated with a large perpendicular angle i.e. 90, 180 or 270 degrees (Figure 3). Such instances could possibly harm the performance of our proposed system, especially when feeding cropped text lines from Text Detection module into Text Recognition and Information Extraction module where text orientation plays an important role. To cope with this issue, we propose dividing into two separate subtasks by solving **a)** and **b)** independently.

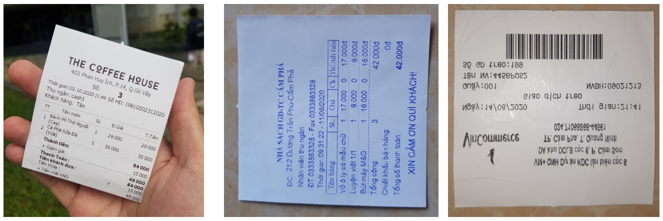

Fig. 3: Subtask **a)**: a slightly misaligned receipt (left). Subtask **b)**: a receipt rotated 90 degrees counterclockwise (middle) and a receipt rotated upside down (right).

On subtask a), we can empirically say that the orientation of the whole receipt depends on the orientation of each individual text line, a simple yet effective solution is to rotate receipt image based on the average rotation angle of predicted bounding boxes produced by PAN [4]. More specifically, suppose the coordinate of top-left and top-right vertices of said bounding boxes are respectively defined as (x_1, y_1) and (x_2, y_2), one can determine sine of the rotated angle by applying the cross-product formula below:

$$\sin(\alpha_i) = \frac{y_2 - y_1}{\|(x_2 - x_1, y_2 - y_1)\|_2}$$

Since calculating rotation angle of each bounding box using inverse trigonometric functions is computationally expensive, in order to keep the computational cost low, a simple workaround would be approximating them by averaging the sine quantity calculated above. Since the sine function is monotonic on $\left[-\frac{\pi}{2}; \frac{\pi}{2}\right]$, it can be approximated well when $\alpha_1, \alpha_2, \ldots, \alpha_n$ has low variance (which is a common occurrence). The average angle could then be estimated using cheap arithmetic operations:

$$\sin(\bar{\alpha}) = \sin\left(\frac{1}{n}\sum_{i=1}^{n}\alpha_i\right) \approx \frac{1}{n}\sum_{i=1}^{n}\sin(\alpha_i)$$

On subtask **b)**, the rotation angle is no longer be limited to $\left[-\frac{\pi}{2}; \frac{\pi}{2}\right]$, hence trionometric functions are no longer monotonic and thus become obsolete. Furthermore, our study reports that PAN does not perform very well on receipts rotated with an angle greater than 90 degrees (either clockwise or counterclockwise), therefore bounding boxes information will not be available for this subtask and an alternative solution has to be proposed.

An interesting observation from the dataset is that image rotation occasionally falls into one of four categories: 90 degrees clockwise rotation, 90 degrees counterclockwise rotation, 180 degrees rotation i.e. upside down, or no rotation at all. One may refer to the Image Classification task, where each class would be one of four rotation angles mentioned above. For this subtask, we are using EfficientNet network architectures to perform Image Classification task, as this type of architecture achieved state-of-the-art performance on ImageNet dataset while also maintaining lightweight computation cost. We ran an experiment to compare the effectiveness between different model sizes (Table II) and ultimately picked EfficientNet-B7 for its high effectiveness despite having such huge model size because preprocessing steps heavily affect the performance of our whole system. We also used Sharpness-Aware Minimization [20] to optimize network parameters, hoping to secure better network generalization and avoid overfitting.

Training data for this subtask is not naturally available, one may try to take detected receipts from previous section and proceed to rotate them to generate training data. In addition to rotating cropped images, we also suggest using basic image augmentation techniques to add non-linear noises such as modifying RGB image on HSV color space, gamma corrections, Gaussian blur, etc. By adopting the above approach, we achieved 97.14% accuracy on validation set.

Model	Accuracy
EfficientNet-B0	0.876
EfficientNet-B4	0.914
EfficientNet-B7	0.971

TABLE II: Comparison between different EfficientNet architectures on subtask b).

B. Task 1: Receipt Image Quality Evaluation

We have formed task 1 into a regression problem and proposed a framework for receipt image quality regression processing, which contains three mandatory steps. The first step is data preprocessing (III-A), after which the receipt will be properly aligned and noisy background will be cropped away. Next, we use cropped receipt regions as input data to train an EfficientNet regression model. Finally, the trained model from previous step is used to predict receipt

[4] One of the challenges we encountered is the output given by PAN is not always perfect and will sometimes returns noises. We suggest calculating the median of rotation angle instead, which is more robust to noises compared to averaging.

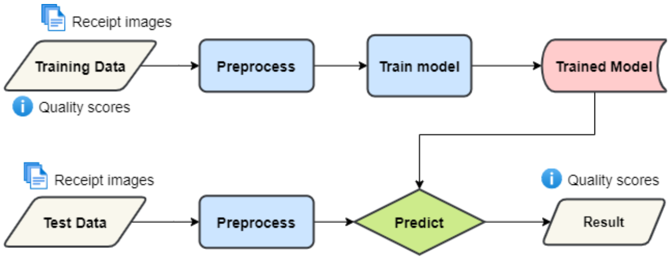

Fig. 4: An overview of our method for solving task 1.

quality scores and then export the results in the format that required by the organizers.

1) Training model:

The regression model we used for training is based on EfficientNet. EfficientNet is a powerful convolutional neural network and is brought into play a lot in common image classification transfer learning tasks. This network was proposed as a new scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. EfficientNet is a continuous family of models from B0 to B7 created by arbitrarily choosing scaling factor. The base model we used is a pretrained network previously trained on a large ImageNet dataset contains 1000 classes labels, we can take advantage of pretrained weights to extract useful image features without retraining from scratch.

As we mentioned above, EfficientNet is often used for classification. For the regression task like the receipt image quality evaluation task in this competion, we exclude the final fully connected (FC) layer that turns features on the penultimate layer into prediction of the labels of classes. Then, replacing the top layer with custom layers containing FC layers for receipt image quality regression allows using EfficientNet as a feature extractor in a transfer learning workflow. The features is fed into global average pooling (GAP) to generate a vector whose dimension is the depth of the feature, this vector is the input of the next FC Layer. The GAP layer outputs the mean of each feature map, this drops any remaining spatial information, which is fine because there was not much spatial information left at that point. The final FC layer 1×1 produces the quality of receipt ranged from 0 to 1.

Because the training data is not sufficient, freezing EfficientNet and training only custom top layers tends to underfit the training data, training both EfficientNet and custom top layers tends to overfit the training data. So, our approach is freezing some first layers of EfficientNet to make use of the low level features extracted by pretrained network on ImageNet datasets, then training the remaining layers and top layers. We use Adam optimizer with a learning rate of 0.0001 to minimize the Root Mean Square Error (RMSE) loss function. We trained 3 base models i.e B0, B4, B7, 1000 epochs for each model with a batch size of 16 and only save the best weight with the lowest validation loss.

2) Inference:

For this final step, we use the trained model that saved at previous step to predict receipt image quality scores from the private test data provided by the organizers in the final round. The private test data is preprocessed first to generate the receipt region images. Then, we pass those through our trained model to predict receipt image quality scores. Finally, we combine the above results with the results produced in task 2 to form the final CSV submission file. The RMSE results of each base model on the private test data are presented in the TABLE IV in IV-A.

C. Task 2: OCR Recognition

Extracting information of necessary fields from receipt is a complicated task because it depends on the results of many components. Figure 5 shows the main components of the module that extract four necessary fields of information. Input is an image that has been preprocessed, the system will use PAN - a text detector that we have trained to localize text fields on the image. We adopted VietOCR to recognize the text content of each line. And finally, we used a rule-based system combined with several NLP techniques to identify and extract content from four predefined fields on the receipt.

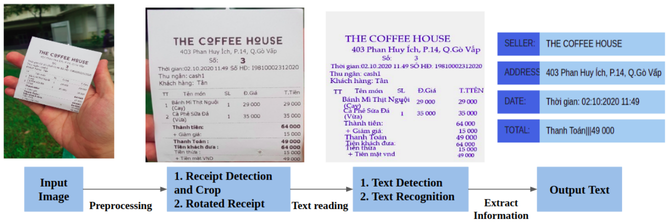

Fig. 5: Illustrate the main processing steps for task 2

1) Text Detection with PAN:

Text detection is a crucial component in our framework as it heavily influences the performance of identifying and extracting information from receipts. After evaluating several methods on SROIE-2019 - a dataset for the international invoice identification competition, we chose Pixel Aggregation Network (PAN) as it offers perfect balance between accuracy and execution time.

Pretrained models of PAN were originally trained on well-known scene text datasets such as CTW1500, SynthText, therefore it did not generalize very well to invoice data and has to be trained on SROIE19 dataset. The dataset was split into two parts: train-val consists of 600 English receipt images with their annotations and a test-set consists of 400 images. The backbone of PAN is set to ResNet18 by default, input size of image is 640x640, with segmetation head is

FPEM-FFM and 2 repeat FPEM. The initial learning rate is 10^{-3}. By using the configuration mentioned above, we achieved prominent results with 96.05% recall and 96.45% precision score on the validation set.

2) Text Recognition:

To recognize detected text lines from previous step, we used VietOCR - a well-known framework used in many researches in computer vision community. The framework provides two methods for Text Recognition which are AttentionOCR and TransformerOCR. AttentionOCR is an attention-based Seq2seq [21] architecture while Transformer OCR is a language model utilizing self-attention mechanism [22], both of which has paved the way for many researches in Computer Vision and Natural Language Processing. Since TransformerOCR is resource intensive and is not suitable for a lightweight and portable system, we adopted the AttentionOCR model pretrained by VietOCR's author. Despite some limitations, the model yields competent results and can be further improved by post-processing.

3) Information Extraction:

Our team has created three separated dictionaries containing keywords to identify the seller, total and address information fields. Words that frequently occur in receipts are included in our dictionaries, especially for the address dictionary where every commune, district and province in Vietnam is included.

Our study reports one notable feature of invoices dataset is that seller name and address are always located at the top of the receipt. Specifically, only 7 lines at the top will be considered to extract seller information by comparing their content to keywords included in the three dictionaries. For the timestamp field, we used regular expression to perform information extraction effectively.

IV. Analysis

The results of task 1 and task 2 in this paper are tested on one Quadro RTX 8000 GPU, one Intel(R) Xeon(R) CPU E5-2620 v3 2.40GHz CPU in a single thread and 65GB RAM.

A. Task 1

We have trained three EfficientNet base models i.e B0, B4 and B7. The GPU memory and time consumed to train each model are shown in the TABLE III below.

Base model	Training time (Hours)	GPU (GB VRAM)
B0	5	46
B4	16	
B7	26	

TABLE III: The training time and VRAM usage for task 1.

Because the number of parameters from B0 to B7 rises gradually, the training time also increases. As with B0, we only need 5 hours to complete the training, but with B7, it takes up to 26 hours to achieve the same task.

The next thing we do is using each model to predict the receipt image quality on the private test data provided by organizers. Finally, we compare the RMSE results of 3 models and show in TABLE IV below.

Base model	RMSE
B0	0.11513
B4	**0.10038**
B7	0.10652

TABLE IV: The RMSE results on the private test data.

Observing the TABLE IV, the base model B4 gives the best result with a RMSE of approximately 0.1 that ranked 1st on the task 1 final standing of this competition. The results of 2 remaining models are also good and quite close to the RMSE of B4 model. This proves the excellent performance of EfficientNet in the regression task.

B. Task 2

Since we extract information from receipts by solving a combination of different subtasks, solving these cases decreased CER score by 0.28 despite the pipeline having some limitations such as when localizing text the seller name, the model has not yet determined the name because the logo has a stylized font or a special symbol shape. During the text recognition process, there is still confusion between "," and "." due to the quality of the receipt e.g “Total 20,000” to “Total 20.200”. We therefore mitigate this confusion by averaging the number of ".", "," markers of the total cost in the receipt to determine whether "." or "," for total cost should be used.

Method	Loss reduce (CER)
Receipt detection + Baseline	0.05
Receipt detection + rotate receipt + baselines	0.19
Receipt detection + update rotate receipt + baseline	0.28

TABLE V: The results through our improvements to the baseline.

TABLE V shows the results of our baseline improvements by subsequently applying each component. First, when using only PAN + VietOCR, we noticed that baseline recognized words outside the background in some images, which affected the final extracted informations. So we added a receipt detection step to get rid of these cases. Next, we processed some receipt images that were rotated 90 or 180 degrees and also improved by 0.19 CER. Finally we took care of some tilted receipt cases and helped baseline improve by 0.28 CER.

Stage	Avg processing time / image (s)	GPU (MB VRAM)
Text Detection (PAN)	0.2	1099
Receipt Rotation	0.18	11778
Text Recognition (VietOCR)	2.2	1029
Receipt Detection	0.158	803
Extract Information	1.2	0
Full Pipeline	4.057	14409

TABLE VI: Computational processing times and resources required in each stage for task 2.

TABLE VI above indicate the average processing time per image and GPU memory required in each stage. In which, the text recognition using VietOCR takes the most time, about half of the entire processing time. Besides, our team have used EfficientNet-B7 that have extremely many parameters to determine which direction the receipt is rotated, so the GPU memory required for this stage is also large.

V. Conclusion

In this competition, our team has evaluated several methods and built an end-to-end framework to solve 2 tasks of the competition. We have used Faster-RCNN with the ResNet101 FPN backbone to perform a receipt detection. Then, receipts have been post-processed and applied EfficientNet-B4 to build a regression model to predict the receipt image quality and reached 0.1 (RMSE) on task 1. For task 2, we have to retrain the PAN to locate the text in the invoice and use the VietOCR to recognition the text in each previously detected bounding box, the result of task 2 is 0.3 (CER). We also provide an analysis of contributions and failed cases we encountered in the contest. At the current time, our team only use the rule base that we observed to extract the text field content. In the future, there are several approaches in the NLP field such as Named Entity Recognition that can be applied to automate this step.

Acknowledgement

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2021-11

References

[1] X.-S. Vu, Q.-A. Bui, N.-V. Nguyen, T.-T.-H. Nguyen, and T. Vu, "Mc-ocr challenge: Mobile-captured image document recognition for vietnamese receipts," *Proceedings of the 15th IEEE-RIVF International Conference on Computing and Communication Technologies*, 2021.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[3] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[4] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8440–8449.

[5] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "Icdar2019 competition on scanned receipt ocr and information extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1516–1520.

[6] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3047–3055.

[7] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1962–1969.

[8] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2550–2558.

[9] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9336–9345.

[10] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[11] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.

[12] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5909–5918.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[14] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[15] Y. Zhao, W. Xue, and Q. Li, "A multi-scale crnn model for chinese papery medical document recognition," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2018, pp. 1–5.

[16] T. I. Denk and C. Reisswig, "Bertgrid: Contextualized embedding for 2d document representation and understanding," *arXiv preprint arXiv:1909.04948*, 2019.

[17] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," *arXiv preprint arXiv:1809.08799*, 2018.

[18] R. B. Palm, F. Laws, and O. Winther, "Attend, copy, parse end-to-end information extraction from documents," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 329–336.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[20] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.