Vietnamese Scene Text Detection and Recognition using Deep Learning: An Empirical Study

Nhat Truong Pham[†]
Institute for Computational Science
Ton Duc Thang University
Ho Chi Minh City, Vietnam
phamnhattruong@tdtu.edu.vn

Van Dung Pham[†]
Deep Learning and Application
Ho Chi Minh City, Vietnam
dungpv.ai@gmail.com

Qui Nguyen-Van
Department of Telecommunications Engineering
Ho Chi Minh City University of Technology
Ho Chi Minh City, Vietnam
nvqui.sdh20@hcmut.edu.vn

Bao Hung Nguyen
SK-Global, JSC.
Ho Chi Minh City, Vietnam
baohung.ng@gmail.com

Duc Ngoc Minh Dang
Computing Fundamental Department
FPT University
Ho Chi Minh City, Vietnam
ducdnm2@fe.edu.vn

Sy Dzung Nguyen[*]
Institute for Computational Science
Ton Duc Thang University
Ho Chi Minh City, Vietnam
nguyensydung@tdtu.edu.vn

***Abstract*—Scene text detection and recognition are vital challenging tasks in computer vision, which are to detect and recognize sequences of texts in natural scenes. Recently, researchers have investigated a lot of state-of-the-art methods to improve the accuracy and efficiency of text detection and recognition. However, there has been little research on text detection and recognition in natural scenes in Vietnam. In this paper, a deep learning-based empirical investigation of Vietnamese scene text detection and recognition is presented. Firstly, four detection models including differentiable binarization network (DBN), pyramid mask text detector (PMTD), pixel aggregation network (PAN), and Fourier contour embedding network (FCEN), are employed to detect text regions from the images. Then, four text recognition models including convolutional recurrent neural network (CRNN), self-attention text recognition network (SATRN), no-recurrence sequence-to-sequence text recognizer (NRTR), and RobustScanner (RS) are also investigated to recognize the texts. Moreover, data augmentation methods are also applied to enrich data for improving the accuracy and enhancing the performance of scene text detection and recognition. To validate the effectiveness of scene text detection and recognition models, the VinText dataset is employed for evaluation. Empirical results show that PMTD and SATRN achieve the highest scores among the others for text detection and recognition, respectively. For knowledge-sharing, our implementation is publicly available at https://github.com/ThorPham/VN_scene_text_detection_recognition.**

***Index Terms*—Data Augmentation, Deep Learning, Optical Character Recognition, Scene Text Detection, Scene Text Recognition.**

I. Introduction

Scene text detection and recognition are to detect and recognize texts in natural scenes with a wide range of applications, such as self-driving cars, document parser and analysis, traffic monitoring, and image retrieval [1], [2]. Scene text detection and recognition face some vital problems, such as complicated backgrounds, weather conditions, and the quality of scene text images. In the past decade, researchers have investigated many techniques for scene text detection and recognition. For example, Zhou *et al.* [3] suggested EAST (efficient and accuracy scene text), a fast and accurate scene text detector based on the U-shaped design in the U-Net architecture. Tian *et al.* [4] also created the CTPN (connectionist text proposal network) component-level architecture by stacking recurrent neural networks (RNN) on top of convolutional neural networks (CNN) to anticipate and link segments of scene text with deep neural networks. He *et al.* [5] developed the DTRR (deep-text recurrent network) framework for scene text recognition, which uses CNN to extract features from scene objects and RNN for sequence learning. However, there are limited studies that investigated Vietnamese scene texts.

In this study, a deep learning-based empirical investigation of Vietnamese scene text detection and recognition is introduced as follows. Firstly, four detection models DBNet, PMTD, PAN, and FCEN are designed to detect text in natural scenes. Secondly, four recognition models are also employed to recognize the texts, including CRNN, SATRN, NRTN, and RS. The VinText [6] is used to obtain experimental results. Additionally, *Recall*, *Precision*, and H_{mean} metrics are used to evaluate text detection models while *Word Acc* (case-insensitive word accuracy) ignore symbol and *1-NED* (normalized edit distance) are used to evaluate text recognition ones. Furthermore, data augmentation methods are also investigated to generate and enrich data to improve the accuracy of scene text detection and recognition. Empirical results show that PMTD and SATRN reach the highest scores for Vietnamese scene text detection and recognition, respectively.

The rest of this paper is structured as follows. Some related studies are summarized in Section II. The methodology is introduced in Section III. Experimental results and discussion are presented and analyzed in Section IV. Finally, the conclusion of this study is summarized in Section V.

[*] *Corresponding author.*
[†] *These authors have contributed equally to this work.*

II. Related Work

In the past decade, deep learning and its applications have been investigated in many research fields, such as computer vision, natural language processing, speech processing, sound classification, and even COVID-19 detection [7]–[10]. Especially many deep learning approaches have been investigated for scene text detection and recognition. For instance, Naiemi *et al.* [11] proposed a novel pipeline of deep learning framework for scene text detection and recognition by improving the ReLU (Rectified Linear Unit) activation function, enhancing the inception layer, and employing a new local word directional pattern for feature extraction. This study could tackle some vital challenges in scene text detection and recognition, such as complex backgrounds, font size variations, and arbitrary-shaped texts. A blind deconvolution model [12] was proposed for scene text detection and recognition in the video by introducing a combined quality metric for estimating the degree of blur. By suppressing the blurred pixels, the proposed model could improve the edge intensity. In addition, many challenges in conjunction with *The International Conference on Document Analysis and Recognition* (ICDAR) have been organized to solve the tasks of scene text detection and recognition using deep learning approaches in the past decade, such as ICDAR2013, ICDAR2015, ICDAR2017, ICDAR2019, and ICDAR2021 competitions.

A robust and generalized deep learning model usually requires a lot of labeled or annotated datasets for training. However, it is time-consuming and expensive to collect and annotate a large dataset. Recent studies have tried to use adversarial learning like generative adversarial networks (GANs) to generate or synthesize more and more data to improve the accuracy of deep learning models. For instance, Zhan *et al.* [13] proposed a geometry-aware domain adaptation network for scene text detection and recognition that converts a source-domain image into multiple images with different views by modeling cross-domain shifts in both appearance and geometry spaces simultaneously. Besides, multimodal spatial learning and a novel disentangle cycle-consistency loss are applied to improve the stability and robustness of learning in both two spaces. Atienza [14] proposed an augmentation method named STRAug for scene text recognition by employing 36 image augmentation functions.

Recently, Nguyen *et al.* [6] contributed the VinText dataset, the biggest Vietnamese scene text collection, which was either acquired from the Internet or taken by employees. Within 2,000 completely annotated images, VinText has 56,084 text occurrences. The images were divided into three subsets: 1,200 images for the training subset, 300 images for the validation subset, and 500 images for the testing subset. As a result, this dataset leads researchers to put attention to Vietnamese scene text detection and recognition.

III. Methodology

A deep learning-based empirical study of Vietnamese scene text detection and recognition is provided and contrasted using multiple deep learning models, motivated by the aforementioned previous works. For text detection, four deep learning models are employed and compared. For text recognition, four other deep learning models are employed and compared. The short descriptions of the employed models are summarized as follows.

A. Scene Text Detection

1) Differentiable Binarization Network (DBN): DBN [15] is a text detection model based on a segmentation technique that uses a differentiable binarization (DB) module to execute binarization in a segmentation network. Rather than utilizing a threshold to distinguish foreground and background as in typical segmentation approaches, an adaptive functional threshold was developed alongside DBN during the training phase, which not only simplifies post-processing but also improves text detection performance. DBN is empirically used to train a text detection model in our work, employing the backbone of ResNet-50 [16].

2) Pixel Aggregation Network (PAN): PAN [17] is an efficient and accurate arbitrary-shaped text detector. It is also a scene text identification approach that uses segmentation and features a low-cost segmentation head and learnable post-processing. There are three key components in the architecture of PAN: (1) the feature pyramid enhancement (FPE) module is a cascadable U-shaped module that can introduce multi-level information to guide better segmentation; (2) the feature fusion module can combine the features provided by the FPE modules of different depths into a final feature for segmentation; and (3) the learnable post-processing is implemented by pixel aggregation that can precisely aggregate text pixels using the predicted similarity vectors.

3) Pyramid Mask Text Detector (PMTD): By providing a "soft" semantic segmentation between the text region and non-text region, PMTD [18] was built based on Mask R-CNN with ResNet-50 as the baseline backbone. It uses location-aware supervision to execute pixel-level regression, resulting in a more informative soft text mask for each text occurrence. PMTD reinterprets the acquired 2D soft mask into 3D space to construct text boxes. By utilizing the 3D coordinate, a plane clustering method predicts a more accurate text box and improves resilience against faulty bounding box predictions.

4) Fourier Contour Embedding Network (FCEN): FCEN [19] is another arbitrary-shaped text detector. To encode arbitrarily formed text contours as compact signatures, a text instance is first translated into the Fourier domain using the Fourier transform, and then an FCE (Fourier contour embedding) approach is used. Next, an FCEN with the backbone of ResNet-50 with deformable convolutional networks, feature pyramid networks, and the Fourier prediction header was built as an arbitrary-shaped text detector. In the Fourier domain, FCEN predicts text instances' Fourier signature vectors, then reconstructs text contour point sequences in the image spatial domain using inverse Fourier transform and non-maximum suppression.

B. Scene Text Recognition

1) Convolutional Recurrent Neural Network (CRNN): To identify sequence-like objects in images, CRNN [20] is a combination of deep CNN and RNN. Three key components of CRNN: (1) deep CNN layers automatically extract a feature sequence from each input image; (2) RNN layers predict each frame of the feature sequence output by the deep CNN layers; and (3) a transcription layer translates the per-frame predictions by the recurrent layers into a label sequence. In terms of learning image representations and providing a series of labels, CRNN shares the same characteristics as CNN and RNN. Furthermore, in both the training and testing stages, CRNN just requires normalized height normalization, with no requirement for sequence-like object lengths.

2) Self-Attention Text Recognition Network (SATRN): Based on the Transformer architecture, the SATRN [21] was designed for recognizing texts of any form. SATRN is made up of two primary parts: (1) A shallow CNN to extract local features and a SATRN encoder with an adaptive 2D positional encoding and a locality-aware feedforward layer to embed local patterns into 2D feature maps; (2) a Transformer decoder to recover sequence labels (characters) from the embedded 2D feature maps.

3) No-Recurrence sequence-to-sequence Text Recognizer (NRTR): Because convolutional-recurrent neural networks for scene text detection and recognition have limitations in terms of training speed and model complexity, NRTR [22] was created using a self-attention encoder-decoder architecture to extract features and recognize texts. A modality-transform block with numerous CNN layers is designed to change the input image to the correct sequence to enhance features in the encoder component. As a result, NRTR not only solves the training speed and model complexity limitations in CRNN-based systems but also simplifies the operation and improves the performance.

4) RobustScanner (RS): Because the attention-based encoder-decoders perform less accurately on contextless texts, RobustScanner [23] modified the decoder of the encoder-decoder structure with attention-based techniques. Based on empirical testing, this study found that the sequence labels (character-level) of the decoder process contain not just context but also position information. The encoder-decoder is employed by three basic components: (1) a CNN encoder based on an adopted 31-layer ResNet; (2) a hybrid branch containing two LSTM layers with a hidden state size of 128 and one attention module; and (3) a positional enhancement branch containing one position embedding layer, one position-aware module, and one attention module. The feature maps extracted from the CNN encoder are fed into hybrid and positional enhancement branches to obtain further representations. Finally, a dynamically-fusing module is also designed to dynamically fuse the obtained representations from two branches for predictions.

C. Data Augmentation

The deep learning approach usually requires a lot of labeled or annotated datasets. However, it is time-consuming and costly to collect and annotate a large dataset. Therefore, in this study, two augmentation methods are employed to enrich data for training a robust and generalized deep learning model as follows.

1) Learn to Augment (LearnAug): LearnAug [24] is a scene text and handwritten text recognition job that combines data augmentation and network optimization as follows: (1) an agent network is trained to predict the distribution of the moving state with the goal of creating a harder training sample from a set of custom fiducial points on the image; (2) an augmentation module generates augmented samples based on the random and predicted moving states, respectively; and (3) a recognition network predicts text strings on the augmented images. The recognition network uses the edit distance metric to assess the complexity of the pair of samples. The agent learns from the shifting state which increases the challenge and reveals the recognition network's flaws.

2) Synthetic Text Image GEneratoR (SynthTIGER): SynthTIGER [25] was proposed by analyzing and combining the most successful synthetic text image generator techniques into one. Text selection and text rendering are the two fundamental components of SynthTIGER. There are five processes in the text rendering module: text shape selection, text style selection, transformation, blending, and post-processing. Text selection approaches based on text length distribution and text character distribution are proposed. As a result, manipulating the text styles and text distributions in the synthetic dataset has an impact on the development of more generalized scene text recognition models.

IV. Experimental Results

A. Experimental Settings

1) Hyper-parameter: In this empirical study, Adam [26], SGD (stochastic gradient descent) [27], and Adadelta [28] optimizers are employed to train different deep learning models with hyper-parameter settings as presented in Table I.

TABLE I: Hyper-parameter settings for different scene text detection and recognition models.

Hyper-parameter settings				
Model	***Batch size***	***Optimizer***	***Learning rate***	***Epochs***
DBN	8	SGD	7e-3	150
PAN	8	Adam	1e-3	150
PTMD	8	SGD	8e-2	150
FCEN	8	SGD	1e-2	150
CRNN	256	Adadelta	1	5
SATRN	32	Adam	3e-4	5
NRTR	128	Adam	1e-3	5
RS	128	Adam	1e-3	5

Besides, in this empirical study, LearnAug is applied to augment around 50% of the dataset, while SynthTIGER is employed to generate one million images for scene text recognition.

2) Evaluation Metrics: Inference time (*IT*) and the number of parameters (*Params*) are considered as evaluation metrics for both scene text detection and recognition.

a) For scene text detection: In scene text detection, *Precision*, *Recall*, and H_{mean} are used as evaluation metrics:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$H_{mean} = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{3}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively.

b) For scene text recognition: In scene text recognition, case-insensitive word accuracy (*Word Acc*) ignore symbols and adopted normalized edit distance ($1 - NED$) are chosen as evaluation metrics. *NED* is defined as follows:

$$NED = \frac{1}{n} \sum_{i=1}^{n} L_D \frac{(s_i, \hat{s}_i)}{max(s_i, \hat{s}_i)} \tag{4}$$

where n is the maximum number of matching pairs of ground truth and detected regions, L_D is the Levenshtein distance, s_i and $\hat{s}_i$ are the ground truth in the regions and the corresponding predicted text line in the string, respectively.

B. Results

1) For Scene Text Detection: Fig. 1 depicts the detection results using different models, while the results are presented in details in Table. II in terms of Recall, Precision, H_{mean}, *IT* per image, and the number of parameters (*Params*), respectively. As shown in Table. II, PMTD achieves the highest scores of 81.7%, 91.9%, and 86.5% for *Recall*, *Precision*, and H_{mean}, respectively. However, PMTD has the highest *IT* and the highest *Params*, so it consumes a lot of computing resources. The comparison of H_{mean} and computing resources is visualized in Fig. 2.

TABLE II: The results of different models for text detection.

Text Detection					
Model	***Recall*** **(%) ↑**	***Precision*** **(%) ↑**	H_{mean} **(%) ↑**	***IT*** **(ms) ↓**	***Params*** ($\times 10^6$)
DBN	78.6	91.2	84.4	11.2	26.3
PAN	67.2	83.7	74.6	9.8	24.2
PMTD	81.7	91.9	86.5	34.6	43.7
FCEN	69.7	84.8	76.5	81.0	26.3

where ↑ means the higher the better and ↓ means the lower the better.

2) For Scene Text Recognition: Fig. 3 depicts the recognition results using different models, while the results are presented in details in Table. III in terms of *Word Acc*, *1-NED*, *IT* per image, and the number of parameters (*Params*), respectively. As shown in Table. III, SATRN achieves the highest scores of 86.3% and 91.6% for *Word Acc*, and *1-NED*, respectively while NRTR has the highest *IT* and the highest *Params* compared to the others. The comparison of total accuracy and computing resources is visualized in Fig. 4.

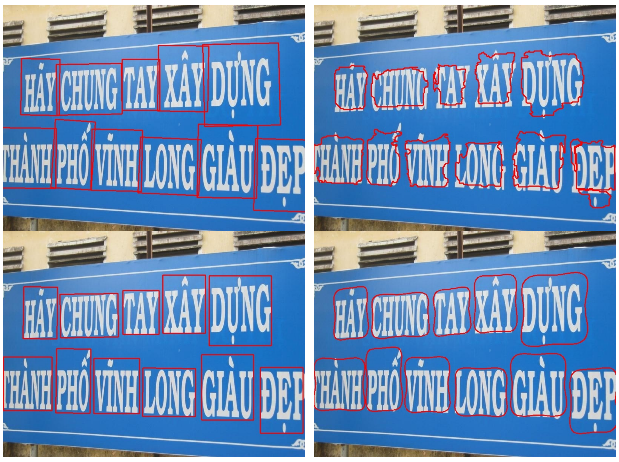


Fig. 1: Scene text detection results of DBN (top-left), PAN (top-right), PMTD (bottom-left), and FCEN (bottom-right), respectively.

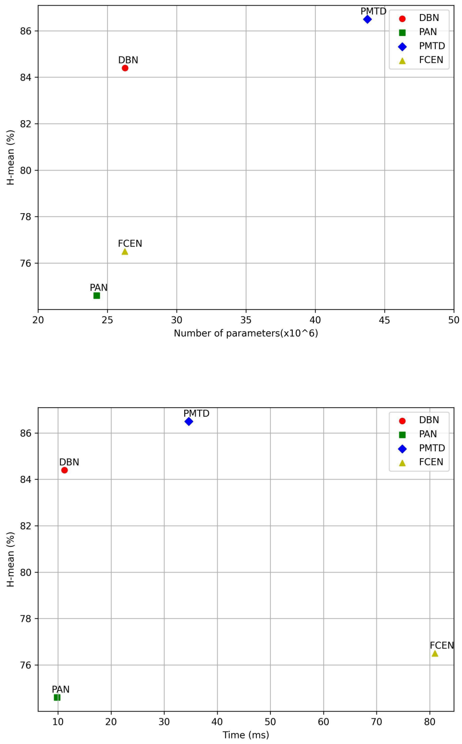


Fig. 2: Comparison of different detection models in terms of H_{mean} (%), inference time per image, and number of parameters.

Moreover, to improve the accuracy of scene text recognition models, data augmentation methods are applied in this study. Fig. 5 depicts the visualization of examples of synthesis and augmentation methods.

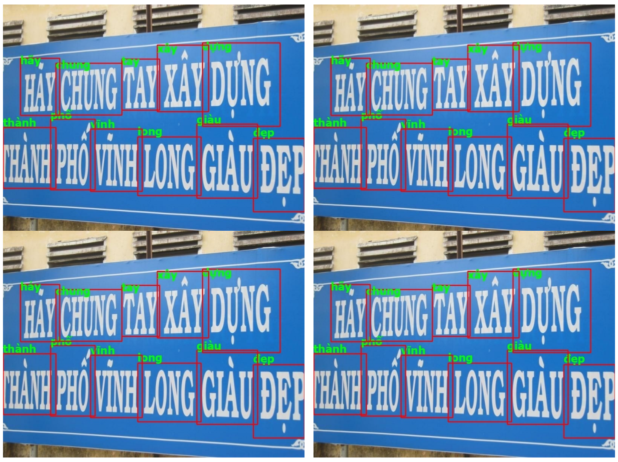


Fig. 3: Scene text recognition results of CRNN (top-left), SATRN (top-right), NRTR (bottom-left), and RS (bottom-right), respectively, using detected text regions obtained by PMTD.

TABLE III: The results of different models for text recognition.

Text Recognition				
Model	***Word Acc (%)*** ↑	***1-NED (%)*** ↑	***IT* (ms)** ↓	***Params*** ($\times 10^6$)
CRNN	79.1	87.6	3.0	8.4
SATRN	86.3	91.6	140.0	65.7
NRTR	84.2	90.6	220.0	66.7
RS	78.2	85.5	40.0	48.0

C. Discussion

In scene text detection and recognition, the higher accuracy the higher the computing resources. Hence, system developers have to trade-off between accuracy and performance to design a suitable model for the desired system.

In addition, data augmentation methods can augment and generate more and more data samples to improve the accuracy of scene text recognition.

Furthermore, the results in this empirical study could be used as baseline systems for further research.

V. Conclusion

In this study, an empirical deep learning-based study has been experimented with and compared for Vietnamese scene text detection and recognition. Four detection models and four recognition models have been investigated and analyzed. From which, PMTD and SATRN have achieved the highest scores on the VinText dataset for Vietnamese scene text detection and recognition, respectively. Moreover, data augmentation methods have been employed to enrich data, hence, improving the accuracy of text recognition models. Based on this empirical study, researchers would put more attention to Vietnamese scene text detection and recognition as well as contribute more and more Vietnamese scene datasets. Besides, our implementation is publicly available as an open-source so that readers can reproduce our experimental results and system developers can choose a suitable model for the desired system in applications, such as robot navigation, image retrieval, self-driving car, and traffic monitoring.

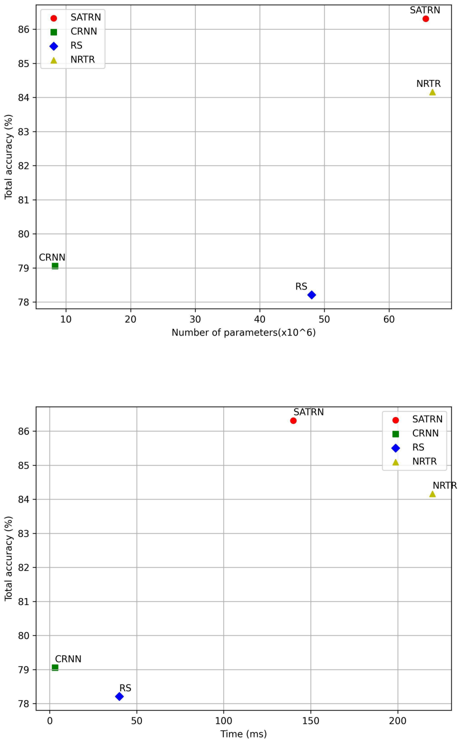
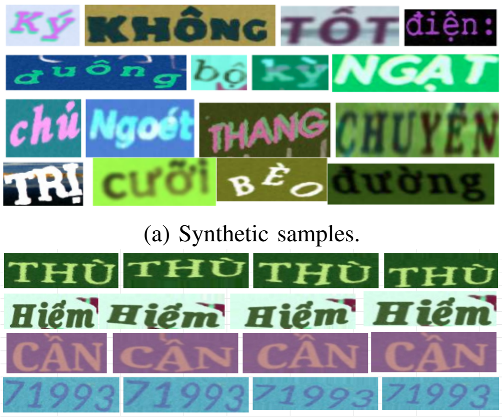


Fig. 4: Comparison of different recognition models in terms of total accuracy (%), inference time per image, and number of parameters.



(a) Synthetic samples.

(b) Augmented samples.

Fig. 5: Visualization of synthetic (a) and augmented (b) samples.

Acknowledgment

The authors would like to thank the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.01-2019.328.

References

[1] C. Zhang, Y. Tao, K. Du, W. Ding, B. Wang, J. Liu, and W. Wang, "Character-level street view text spotting based on deep multisegmentation network for smarter autonomous driving," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 297–308, 2021.

[2] C. Zhang, W. Ding, G. Peng, F. Fu, and W. Wang, "Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4727–4743, 2020.

[3] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2642–2651. [Online]. Available: https://doi.org/10.1109/CVPR.2017.283

[4] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9912. Springer, 2016, pp. 56–72. [Online]. Available: https://doi.org/10.1007/978-3-319-46484-8_4

[5] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 3501–3508. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12256

[6] N. Nguyen, T. Nguyen, V. Tran, M. Tran, T. D. Ngo, T. H. Nguyen, and M. Hoai, "Dictionary-guided scene text recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 7383–7392. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Nguyen_Dictionary-Guided_Scene_Text_Recognition_CVPR_2021_paper.html

[7] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, "A method upon deep learning for speech emotion recognition," *Journal of Advanced Engineering and Computation*, vol. 4, no. 4, pp. 273–285, 2020.

[8] L. H. Nguyen, N. T. Pham, V. H. Do, L. T. Nguyen, T. T. Nguyen, V. D. Do, H. Nguyen, and N. D. Nguyen, "Fruit-cov: An efficient vision-based framework for speedy detection and diagnosis of sars-cov-2 infections through recorded cough sounds," *arXiv preprint arXiv:2109.03219*, 2021.

[9] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *arXiv preprint arXiv:2109.09026*, 2021.

[10] T. Tran, N. T. Pham, and J. Lundgren, "A deep learning approach for detecting drill bit failures from a small sound dataset," *Scientific Reports*, vol. 12, no. 1, pp. 1–13, 2022.

[11] F. Naiemi, V. Ghods, and H. Khalesi, "A novel pipeline framework for multi oriented scene text image detection and recognition," *Expert Systems with Applications*, vol. 170, p. 114549, 2021.

[12] V. Khare, P. Shivakumara, P. Raveendran, and M. Blumenstein, "A blind deconvolution model for scene text detection and recognition in video," *Pattern Recognition*, vol. 54, pp. 128–148, 2016.

[13] F. Zhan, C. Xue, and S. Lu, "GA-DAN: geometry-aware domain adaptation network for scene text detection and recognition," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 9104–9114. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00920

[14] R. Atienza, "Data augmentation for scene text recognition," in *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*. IEEE, 2021, pp. 1561–1570. [Online]. Available: https://doi.org/10.1109/ICCVW54120.2021.00181

[15] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 11474–11481. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6812

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 8439–8448. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00853

[18] J. Liu, X. Liu, J. Sheng, D. Liang, X. Li, and Q. Liu, "Pyramid mask text detector," *arXiv preprint arXiv:1903.11800*, 2019.

[19] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 3123–3131.

[20] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2646371

[21] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2d self-attention," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 2326–2335. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w34/Lee_On_Recognizing_Texts_of_Arbitrary_Shapes_With_2D_Self-Attention_CVPRW_2020_paper.html

[22] F. Sheng, Z. Chen, and B. Xu, "NRTR: A no-recurrence sequence-to-sequence model for scene text recognition," in *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*. IEEE, 2019, pp. 781–786. [Online]. Available: https://doi.org/10.1109/ICDAR.2019.00130

[23] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "Robustscanner: Dynamically enhancing positional clues for robust text recognition," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12364. Springer, 2020, pp. 135–151. [Online]. Available: https://doi.org/10.1007/978-3-030-58529-7_9

[24] C. Luo, Y. Zhu, L. Jin, and Y. Wang, "Learn to augment: Joint data augmentation and network optimization for text recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 13743–13752. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Luo_Learn_to_Augment_Joint_Data_Augmentation_and_Network_Optimization_for_CVPR_2020_paper.html

[25] M. Yim, Y. Kim, H. Cho, and S. Park, "Synthtiger: Synthetic text image generator towards better text recognition models," in *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, J. Lladós, D. Lopresti, and S. Uchida, Eds., vol. 12824. Springer, 2021, pp. 109–124. [Online]. Available: https://doi.org/10.1007/978-3-030-86337-1_8

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[27] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 1139–1147. [Online]. Available: http://proceedings.mlr.press/v28/sutskever13.html

[28] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.