LHFAN: Scene Text Recognition Method Based on Multi-level Feature Fusion and Enhancement of Semantic Knowledge

Ruturaj Mahadshetti, Guee-Sang Lee*, Hyung-Jeong Yang, and Soo-Hyung Kim

Department of Artificial Intelligence Convergence
Chonnam National University, Gwangju 61186, South Korea
ruturajm9770@gmail.com, gslee@jnu.ac.kr, hjyang@jnu.ac.kr, and shkim@jnu.ac.kr

Abstract. In various computer vision tasks, STR has been a sensual research subject and performs a considerable utility. Modern deep-learning algorithms for scene text recognition (STR) have contrived significant advancements over the last few years. However, they aren't optimal for recognizing the text from degraded images or even when images are clear. Many STR methods employ transformers to utilize linguistic information for arduous recognition tasks rather than purely visible classification, but the recognition outcome is not superior for degraded images and confusing text fonts. Early scene text recognition approaches may even yield a substantial percentage of erroneous outputs when an image acquire in natural conditions. In this study, we proposed a novel approach called LHFAN to address the above issues by utilizing low-level features with high-level features, which improves the ability of visual features and semantic features. LHFAN employs upsampling method with ResNet and blends different scale features to enrich the feature capability of text content. We demonstrate with experimental results that our proposed method outperforms on standard text recognition datasets and also achieves state-of-the-art performance when working with blurred and perspective images.

Keywords: Scene Text Recognition · Deep Learning · Transformer · Convolutional Neural Network

1 Introduction

Scene text recognition, also known as optical character recognition (OCR) in natural scenes, is the process of automatically identifying and extracting written text from images and videos. Scene Text Recognition (STR) has emerged as a popular and intriguing study area in both academic and commercial circles as a subfield of computer vision. This technology has a wide range of practical applications, such as in self-driving cars, mobile document scanning, and augmented reality. However, the problem of scene text recognition is challenging due to the variability in text appearances, such as different fonts, colors,

and orientations, as well as the presence of noise and occlusions. Numerous text recognition techniques have recently been offered as solutions to this challenging issue, such as pictorial feature extraction methods [13, 25], attention-based mechanisms [32, 38, 39], resolution enhancement [3, 22, 26], and rectification mechanisms as pre-processing tasks [61, 1, 13].

Fig. 1. Failure results from the previous framework. (1) LevOCR [9], (29) MGP-STRF, and ground truth (gt).

In recent years, Scene text recognition approaches have made considerable progress and achieved significant performance because of their ability to learn and extract features from large amounts of data. One popular deep learning approach for scene text recognition is the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNN's can be utilized to extract features from an image, while RNNs can be used to process these features and recognize the text. Another approach is to use attention-based mechanisms to focus on the most important parts of the image, like the text itself, during the recognition process. Recently, End-to-end trainable models such as the Connectionist Temporal Classification (CTC) and Attention-based methods have been proposed, which directly take an image as input and output the recognized text. These models have been shown to achieve high levels of accuracy and can be trained on large datasets to generalize well to unseen data. In addition to CNN's and RNNs, other deep learning techniques, such as Generative Adversarial Net-

works (GANs) and the Generative Pre-trained Transformer (GPT) employed to improve scene text recognition.

Recent scene text recognition techniques [9, 29] have demonstrated high performance, but their effectiveness diminishes when applied to Pixelated or partially obscured images. Many methods [9, 29, 19] take advantage of the Vision-Language Transformer and fuse the semantic feature knowledge with visual features instead of considering it separately. Da et al. [9] introduce the LevOCR framework using Levenshtein Transformer, and the refinement process uses two basic character-level operations (deletion, insertion) to enrich the accuracy of the text recognition task. These operations are learned using imitation learning, which allows for the parallel decoding of text and dynamic change in the length of text. Wang et al. [29] proposed a framework utilizing a Multi-Granularity Prediction technique to take advantage of linguistic knowledge. Wu et al. [19] employ Masked Autoencoders for a masking mechanism that allows the model to ignore the background pixels and only focus on the text pixels. [19] use the iterative correction process to improve performance. However, previous methods have a significant drawback in recognizing text from images that have partial obstructions, variations in text appearances that are seen in their curved shapes, various orientations, size variations, and fancy font styles. Fig. 1 shows the failure results of earlier work. Low-level features for scene text recognition are basic image features that are extracted from the input image and used as inputs to the text recognition system. Low-level features are used as inputs to more sophisticated text recognition algorithms, such as optical character recognition (OCR) and convolutional neural networks (CNNs). High-level features in scene text recognition capture the meaning and context of the text in an image through techniques like text semantic segmentation, confidence scores, and text-image relationships. These high-level features enhance the robustness and precision of text recognition systems, particularly in complex real-world scenarios where the text may be obscured, distorted, or written in multiple languages.

Prior approaches have acquired promising results on several benchmarks, indicating their potential effectiveness in solving a particular problem or achieving a specific task. However, recent frameworks fail to generate robust linguistic knowledge and visual features. To overcome these issues, we propose a novel approach utilizing different level features that are helpful to enhance the shallow semantic knowledge for recognition. LHFAN unite ResNet and the convolutional pyramid method to enrich extracted visual features and linguistic information about text instance. The Convolutional pyramid process upsamples features at different levels. After that, LHFAN blends low-level features with high-level features from several layers. Low-level features play a crucial role in the process of scene text recognition by providing a foundation for identifying and extracting text from an image. Combined low-level features and deep features information improve the accuracy of the text recognition process by providing additional information about the layout and structure of the text in an image. This work primarily offers the following contributions:

1. We explore an implicit method to unite low-level features and deep features to improve the robustness of extracted features and linguistic information and prove the importance of low-level features for Scene Text Recognition.
2. The proposed framework is designed to be robust and able to effectively reduce the number of false positives and increase recognition accuracy.
3. The LHFAN algorithm demonstrates exceptional results and surpasses current methodologies in the field.

2 Related work

The classical method for scene text recognition (STR) involves using a convolutional neural network (CNN) to extract visual features, an recurrent neural network (RNN) for ordering labeling, and Connectionist Temporal Classification (CTC) [10] to calculate the loss [13, 20]. Recently, a method known as GTC [4] has been developed to improve the CTC-based method by incorporating a graph convolutional network (GCN) [31] to learn local correlations of features in irregular text images. Some works have also aimed to correct these irregular text images, such as RPI [24] uses the quadratic Bezier curves, and ScRN adds symmetry constraints in addition to Thin-Plate-Spline transformation. Recent studies [9, 27, 19] have proposed innovative solutions for handling difficult situations, such as occlusion and noise, by utilizing models that incorporate semantic information.

Language-aware methods. In some approaches [5, 32], semantics are extracted using external language models. For example, SRN [32] uses the predicted text from the visual model to construct a global semantic reasoning module. In ABINet [5], the method is developed as a bidirectional cloze network, which uses an iterative correction for the language model and makes better use of bidirectional linguistic information. Another approach [14, 33, 12] involves impliedly learning semantics without using language models. Local and global mixing blocks are used in SVTR [7] to recognize both characters and their long-term dependency. Visual-semantic interaction is achieved with VST [33] by extracting semantic information from a visual feature map. ConCLR [27] framework first creates characters with varied contexts using basic image concatenation operations and then optimizes the contrastive loss on the obtained embeddings. ConCLR mitigates the side-effect of overfitting to specific contexts by gathering together clusters of identical characters within different contexts in the embedding space. Recently, various models that adopt the language model approach into account to improve performance by taking into account spatial context [], incorporating real-world images in a semi-supervised manner, or using re-ranking techniques to generate more valuable output.

Modern **Transformer-based methods** have demonstrated their effectiveness in scene text recognition (STR). For example, PIMNet [30] uses a bidirectional Transformer-based parallel decoder to iteratively gather contextual information. Also, ViTSTR [35] utilizes a Vision Transformer (ViT) encoder for recognition tasks without a decoder, and it is utilized with pre-trained parameters from DeiT. Also, ViTSTR employs a Vision Transformer (ViT) encoder for

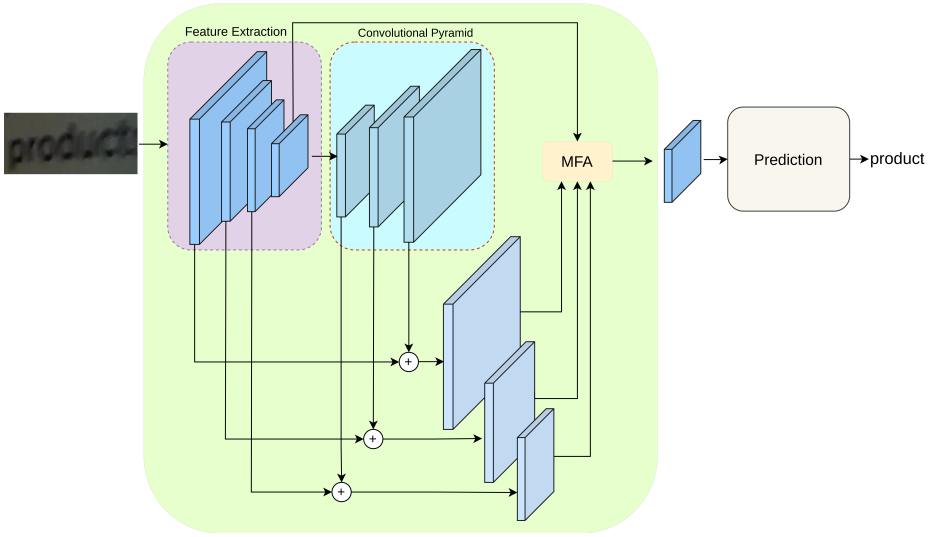

Fig. 2. The illustration of the LHFAN method, LHFAN consists of four parts: Feature Extraction, Convolution Pyramid, Multi-level Feature Aggregation (MFA), and Prediction.

recognition tasks without a decoder, and it is utilized with pre-trained parameters from DeiT. Wang et al. [29] proposed a framework using a Multi-Granularity Prediction technique to take advantage of linguistic knowledge. Wu et al. [19] employ Masked Autoencoders for a masking mechanism that allows the model to ignore the background pixels and only focus on the text pixels. In our proposed model, we employ a transformer for the prediction process.

3 Methodology

The detailed architecture of LHFAN describe in Fig. 2. Our proposed LHFAN consists of four parts: feature extraction, Convolution Pyramid, Multi-feature aggregation (MFA), and Prediction. We employ ResNet-50 as a backbone to extract features and utilize a feature aggregation approach to merge deep features and low-level features. The text images share as input to ResNet and extract features. The input image size is $\mathrm{R}^{H\times W\times 3}$, and the output size is $\mathrm{R}^{H\times W\times D}$. Here H is height, W is weight, and D is channel size. After extracting features (P_0, P_1, P_2, P_3), Convolutional Pyramid adopted a feature upsampling approach, which enhances the feature's representation ability (P_0*, P_1*, P_2*). After upsampling, Extracted features and output of the Convolutional Pyramid merge at a different level.

Multi-level Feature Aggregation Module

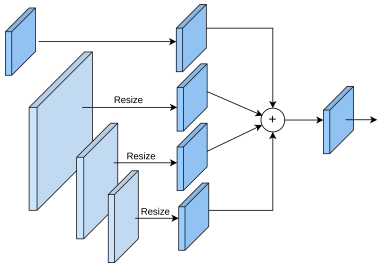

Fig. 3. Architecture of multi-level feature aggregation

The primary aim of Multi-level feature aggregation (MFA) is to unite low-level features with high-level features to enrich the visual and linguistic features from a different levels. The illustration of Multi-level feature aggregation is shown in Fig. 3. P_3 and combined features after upsampling share as input to the MFA module, which captures robust features of text content. The combined features resize to the same size as the size of P_3 using Conv2d. After resizing all layers, MFA combines these features with P_3 and shares them as input to the prediction module.

Parallel Prediction

Unlike previous approaches that separate the process of capturing visual and linguistic information into two distinct stages, we introduce a framework into a single unified architecture. LHFAN utilized a self-attention approach to the prediction task. The proposed method utilizes N transformer units, which have been shown to effectively handle long-term dependencies in recent computer vision tasks. To take into account the position of pixels, the method employs position encoding. This approach utilizes transformer units for sequence modeling instead of only for language modeling, which eliminates the effect of the length of words. The model makes predictions for multiple characters simultaneously, using the previously processed information.

4 Experimental Results

4.1 Dataset

We utilize **SynthText** and **SynthText90K** datasets for training and six datasets to evaluate the framework (three regular (IC13, SVT, IIIT5k) and three irregular datasets(IC15, SVTP, CUTE80)). SynthText [15] dataset consists of 8 million synthetic word samples. SynthText90K [8] dataset include 9 million images covering 90k English word instances. The proposed method is evaluated using the **ICDAR 2013** (IC13), **ICDAR 2015** (IC15), **IIIT 5K-Words** (IIIT5k), **Street View Text** (SVT), **Street View Text-Perspective** (SVTP), and **CUTE80** (CUTE) datasets.

4.2 Implementation Details

Our model is built using ResNet as its backbone, and we specifically set the stride default for the initial stage and increase it for other stages. The weights are initialized using the default initialization method. In our experiments, we used an image size of 128x32 and employed a data augmentation process that included random rotation, color jittering, and perspective distortion. We conducted the experiments on an NVIDIA GTX 3090 GPU with a batch size of 192. The Adam optimizer is used to train the network completely, with a learning rate of 1e-4, in an end-to-end fashion and uses the cross-entropy loss to calculate the loss. The recognition system is designed to cover 37 characters, including letters a-z, numbers 0-9, and an end-of-sequence symbol.

Table 1. Scene text recognition accuracy compared with other STR methods on six standard benchmarks.

Method	**IC13**	**SVT**	**IIIT5K**	**IC15**	**SVTP**	**CUTE**
Chu et al. [1]	95.50	95.50	96.60	84.40	89.90	90.30
Loginov rt al. [2]	96.80	94.70	93.5	80.20	89.90	-
ABINet[5]	97.30	93.50	96.20	86.00	89.30	89.20
Cheng et al. [7]	93.30	85.90	87.40	70.60	-	-
PIMNet[30]	95.4	94.70	96.70	85.90	88.20	92.70
S-GTR [16]	95.80	94.10	96.80	87.90	84.60	92.30
CornerTransformer[21]	96.40	94.60	95.90	86.30	91.50	92.00
VisionLAN [12]	95.80	95.70	91.70	83.70	86.00	88.50
SVTR-L[14]	97.20	91.70	96.30	86.60	88.40	95.10
LevOCR[9]	96.85	92.89	96.63	86.42	88.06	91.67
Zhang et al.[27]	**97.70**	94.30	96.50	85.40	89.30	91.30
MGP-STRF[29]	97.32	94.74	96.40	87.24	91.01	90.28
MVLT[19]	97.30	94.70	96.80	87.20	90.90	91.30
Ours*	97.00	**96.01**	**96.81**	**88.40**	**92.28**	**95.20**

4.3 The effectiveness of LHFAN

We evaluate the LHFAN approach to previous top-performing methods on six commonly utilized datasets, as displayed in Table 1. Generally, approaches that consider the intricacies of language tend to perform better compared to those that do not. The LHFAN approach outperforms both language-based and language-free methods by effectively utilizing both low-level features and linguistic information from different levels to enrich the recognition performance on all six benchmarks. As shown in Fig. 4, The proposed method effectively recognizes text from some challenging images, such as degraded, low-quality, and irregular images. The LHFAN approach particularly excels on regular datasets, achieving

Fig. 4. Recognition performance of the proposed method on some challenging examples. The predictions of LHFAN represent below of images.

a 0.31% and 0.01% improvement on the SVT and IIIT5K datasets. On irregular datasets, the LHFAN also shows significant improvement, with increases of 0.5%, 0.78%, and 0.1% on IC15, SVTP, and CUTE, respectively.

5 Conclusion

In this paper, we introduce a novel scene text recognition approach on multi-level feature fusion and enhancement of linguistic information, which can improve the recognition accuracy of distorted and low-quality images. The recognition accuracy is improved on indistinct and confusing text images by effectively increasing the spatial information and adding different levels of semantic information and visible features. Future work on the suggested method should focus on handling complex scenarios of erroneous text appearances.

6 Acknowledgements

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A3B05049058 & NRF-2020R1A4A1019191).

References

1. Chu, X., Wang, Y., Shen, C., Chen, J., Chu, W. (2022). Training Protocol Matters: Towards Accurate Scene Text Recognition via Training Protocol Searching. arXiv preprint arXiv:2203.06696.
2. Loginov, Vladimir. "Why You Should Try the Real Data for the Scene Text Recognition." arXiv preprint arXiv:2107.13938 (2021).

3. Chen, J., Li, B., Xue, X. (2021). Scene text telescope: Text-focused scene image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12026-12035).
4. Wenyang Hu, Xiaocong Cai, Jun Hou, Shuai Yi, and Zhiping Lin. GTC: guided training of CTC towards efficient and accurate scene text recognition. 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, pages 11005–11012. AAAI Press, 2020.
5. Fang, Shancheng, et al. "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
6. Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. Expert Systems with Applications, 41(18):8027–8048, 2014.
7. Chen, X., Jin, L., Zhu, Y., Luo, C., Wang, T. (2021). Text recognition in the wild: A survey. ACM Computing Surveys (CSUR), 54(2), 1-35.
8. Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. NIPS, 2014.
9. Da, Cheng, Peng Wang, and Cong Yao. "Levenshtein OCR." European Conference on Computer Vision. Springer, Cham, 2022.
10. Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. In William W. Cohen and Andrew W. Moore, editors, Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, volume 148 of ACM International Conference Proceeding Series, pages 369–376. ACM, 2006.
11. Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In BMVC, 2012.
12. Wang, Yuxin, et al. "From two to one: A new scene text recognizer with visual language modeling network." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
13. Baoguang Shi, Xiang Bai, and Cong 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., 39(11):2298–2304, 2017.
14. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., ... Jiang, Y. G. (2022). SVTR: Scene Text Recognition with a Single Visual Model. arXiv preprint arXiv:2205.00159.
15. Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localization in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2315–2324, 2016.
16. He, Y., Chen, C., Zhang, J., Liu, J., He, F., Wang, C., Du, B. (2022, June). Visual semantics allow for textual reasoning better in scene text recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 1, pp. 888-896).
17. Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In 2013 12th International Conference on Document Analysis and Recognition, pages 1484–1493. IEEE, 2013.

18. Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In 2011 International Conference on Computer Vision, pages 1457–1464. IEEE, 2011.
19. Wu, Jie, et al. "Masked Vision-Language Transformers for Scene Text Recognition." arXiv preprint arXiv:2211.04785 (2022).
20. Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang. Reading scene text in deep convolutional sequences. In Dale Schuurmans and Michael P. Wellman, editors, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, pages 3501–3508. AAAI Press, 2016.
21. Xie, Xudong, et al. "Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition." European Conference on Computer Vision. Springer, Cham, 2022.
22. Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image superresolution. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2472–2481, 2018.
23. Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pages 1156–1160. IEEE, 2015.
24. Xu, C., Wang, Y., Bai, F., Guan, J. and Zhou, S., 2022. Robustly Recognizing Irregular Scene Text by Rectifying Principle Irregularities. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3061-3068).
25. Jiang, H., Xu, Y., Cheng, Z., Pu, S., Niu, Y., Ren, W., ... Tan, W. (2021, September). Reciprocal Feature Learning via Explicit and Implicit Tasks in Scene Text Recognition. In International Conference on Document Analysis and Recognition (pp. 287-303). Springer, Cham.
26. Zuzana B ´ ılkova and Michal Hradi ´ s. Perceptual license plate super- resolution with CTC loss. ˇ J. Electron. Imaging, 2020(6):52–1, 2020.
27. Zhang, X., Zhu, B., Yao, X., Sun, Q., Li, R., Yu, B. (2022). Context-based Contrastive Learning for Scene Text Recognition. AAAI.
28. Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In Proceedings of the IEEE International Conference on Computer Vision, pages 569–576, 2013.
29. Wang, P., Da, C., Yao, C. (2022). Multi-granularity Prediction for Scene Text Recognition. In European Conference on Computer Vision (pp. 339-355). Springer, Cham.Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, 16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, 5-10, 2021, Proceedings, Part I, volume 12821 of Lecture Notes in Computer Science, pages 319–334. Springer, 2021.
30. Qiao, Z., Zhou, Y., Wei, J., Wang, W., Zhang, Y., Jiang, N., ... Wang, W. (2021, October). PIMNet: a parallel, iterative, and mimicking network for scene text recognition. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 2046-2055).
31. Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

32. Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and ErruiDing. Towards accurate scene text recognition with semantic reasoning networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 12110–12119. Computer Vision Foundation / IEEE, 2020.
33. Xin Tang, Yongquan Lai, Ying Liu, Yuanyuan Fu, and Rui Fang. Visual-semantic transformer for scene text recognition. arXiv preprint arXiv:2112.00948, 2021.
34. Yue He, Chen Chen, Jing Zhang, Juhua Liu, Fengxiang He, Chaoyue Wang, and Bo Du. Visual semantics allow for textual reasoning better in scene text recognition. In ThirtySixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 888–896. AAAI Press, 2022.
35. Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, 16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, 5-10, 2021, Proceedings, Part I, volume 12821 of Lecture Notes in Computer Science, pages 319–334. Springer, 2021.