Hindawi
International Journal of Antennas and Propagation
Volume 2022, Article ID 2996862, 10 pages
https://doi.org/10.1155/2022/2996862

Research Article

HRNet Encoder and Dual-Branch Decoder Framework-Based Scene Text Recognition Model

Meiling Li, Xiumei Li, Junmei Sun, and Yujin Dong

Hangzhou Normal University, School of Information Science and Technology, Hangzhou 311121, China

Correspondence should be addressed to Xiumei Li; xiumei_li@hotmail.com

Received 11 February 2022; Accepted 28 May 2022; Published 15 June 2022

Academic Editor: Giuseppe Castaldi

Scene text recognition (STR) is designed to automatically recognize the text content in natural scenes. Different from regular document text, text in natural scenes has the characteristics of irregular shapes, complex background, and distorted and blurred contents, which makes STR challenging. To solve the problems of STR for distorted, blurred, and low-resolution texts in natural scenes, this paper proposes a HRNet encoder and dual-branch decoder framework-based STR model. The model mainly consists of an encoder module and a dual-branch decoder module composed of a super-resolution branch and a recognition branch in parallel. In the encoder module, the HRNet is adopted to realize the cross-parallel aggregation representation with multiple resolutions during feature extraction and then outputs four kinds of feature maps with different resolutions. Moreover, the supervised attention module is used to strengthen the learning of the important feature information. In the decoder module, the dual-branch structure is adopted, in which the super-resolution branch takes the feature maps with the highest resolution obtained in the encoder module as input and restores images by upsampling through transposed convolution. The four kinds of feature maps with different resolutions are fused through independent transposed convolution layers for multiscale fusion in the recognition branch and then inputted into the attention-based decoder for text recognition. To improve the accuracy of text recognition, the feature extraction effect of the encoder module is together supervised by the super-resolution branch loss and the recognition branch loss. In addition, the super-resolution branch is only used for training and is abandoned during testing to reduce the complexity of the model. The proposed model is trained on Synth90K and SynthText datasets and tested on seven natural scene datasets. Compared with classical models such as ASTER, TextSR, and SCGAN, the recognition accuracy of the proposed model is improved and better recognition results can be achieved on irregular and blurred datasets such as IC15, SVTP, and CUTE80.

1. Introduction

Natural scene text refers to the text content in natural situations, such as billboards and road signs. Due to the high diversity of text in orientation, shape, and blurring, scene text recognition (STR), which is designed to automatically recognize the text content in natural scene images, is challenging [1]. With the development of deep learning, the deep learning-based STR can obtain good text recognition results and has become a research highlight in the field of document analysis and recognition [2]. Moreover, the deep learning-based STR is an essential research technology, which can be employed in many computer vision applications, such as image retrieval, autonomous driving, and handwriting recognition [3–6].

Early STR models are usually based on temporal feature classification, such as the convolutional recurrent neural network (CRNN) [7]. CRNN uses convolutional neural networks to extract visual features and uses recurrent neural networks to learn the bidirectional dependence of feature sequences and predict the probability of character sequences. Then, the predicted probabilities of character sequences are transcribed into text character sequences according to the predefined transformation mode in the transcription layer. However, the setting of the transcription layer in the CRNN requires that the feature sequences of image and text are aligned with each other, which is not beneficial to predict the text sequences with spatial dependence. The model based on the encoder-decoder framework [8] can avoid the alignment problem by training to predict the corresponding relationship between any two sequences. Generally, the visual features of an image are extracted using an encoder and then are converted into a fixed-length intermediate semantic feature sequence by means of a recurrent neural network. Then, the intermediate semantic feature sequence is decoded into a text character sequence through a decoder. The models based on the encoder-decoder framework have achieved higher performance than the earlier models based on temporal feature classification and provide an effective baseline model for further research [9].

However, scene text images are often disturbed by complex background and text distortion, which often cause the information loss of the visual features extracted by the encoder and then lead to the decoder's inaccurate recognition of the target sequences in the noisy decoding time steps. To alleviate the above problems, the ASTER model [10] based on the encoder-decoder framework is proposed and the thin-plate-spline (TPS) [11] is introduced to improve the text distortion, so that the encoder can extract more sufficient visual features from the rectified images. Based on the sequence model, the visual features are converted into the textual features and finally the attention mechanism is introduced to decode the textual features. When confronted with blurred images and low resolution, the models based on the encoder-decoder framework also suffer from low-recognition accuracy, which prompts researchers to introduce auxiliary networks to improve the resolution of scene text images and learn more accurate text information. Inspired by the success of multitask learning, the super-resolution network SRGAN [12] is used as a preprocessing method in text super-resolution (TextSR) [13] to restore the low-resolution image with the corresponding super-resolution image and then input it into the STR model to improve the recognition effect. The super-resolution network RCAN [14] is used as an auxiliary network in PlugNet [15] to update the parameters of the encoder, to achieve better recognition results. Similarly, in the text super-resolution network (TSRN) [16], a sequential-residual block is proposed to extract the sequential information of the scene text images and accomplish the super-resolution task. However, the super-resolution networks adopted by the above three models have complex structure and a large number of parameters, which increases the complexity of the model.

Recently, a parallel high-resolution network (HRNet) is proposed [17]. Instead of restoring high-resolution representations from low-resolution representations, HRNet maintains high-resolution representations at any time and performs multiscale fusion across parallel convolutions to enhance high-resolution representations, therefore greatly improving the detection and segmentation difficulties caused by image blurring and low resolution. Due to its advantages, HRNet is introduced into the STR task to effectively alleviate the text recognition difficulties caused by the scene text images with blurry and low resolution.

In this paper, a HRNet encoder and dual-branch decoder framework-based STR model is proposed to recognize the distorted and blurred text with low resolution. This model innovatively introduces a HRNet encoder to extract visual features and adopts dual-branch decoder structure composed of a super-resolution branch and a recognition branch following the encoder. The feature maps with highest resolution are inputted into the super-resolution branch for upsampling and image recovery. The feature maps with multiple resolutions are fused at multiscale in the recognition branch to accomplish the transformation of feature sequences and obtain the recognized text. The loss of the super-resolution branch and the loss of the recognition branch are together propagated back to enhance the feature extraction effect of the encoder module, therefore improving the performance of text recognition. The main contributions of this paper are as follows:

(1) The HRNet is innovatively used for feature extraction in STR and also performs as a super-resolution network. Moreover, the HRNet encoder provides effective feature maps for the super-resolution branch, therefore decreasing the model complexity caused by the introduction of an auxiliary super-resolution network, such as TextSR. Experiments on several natural scene datasets verify the effectiveness of the proposed model.

(2) In the encoder module, four kinds of feature maps with different resolutions are generated at the end of the HRNet. By using the supervised attention module (SAM) on the feature maps with the highest resolution, the important features are enhanced and the features with a small amount of information are suppressed. In the decoder module, the feature maps enhanced by SAM are upsampled through transposed convolution (Trans Conv2D) in the super-resolution branch to restore the super-resolution images. The other three feature maps with lower resolution are upsampled through independent transposed convolution layers (Independent Trans Conv2D Layers) in the recognition branch. The feature maps with the same size as the feature maps with the highest resolution are generated, and multiscale fusion is implemented to enhance the representation of the feature maps with multiple resolutions.

(3) The parallel dual-branch structure is adopted. In the training stage, the super-resolution branch and the recognition branch are adopted together to strengthen the feature extraction effect of the encoder module and constantly update the effective parameters in the model, so that the recognition branch can recognize the text on the more-effective feature maps. In the testing stage, the model is simplified, the super-resolution branch is abandoned, and only the recognition branch is used to

obtain the recognition results, which is helpful to reduce the model complexity.

2. Related Work

The ASTER model [10] based on the encoder-decoder framework introduces a rectification network TPS to alleviate the recognition difficulties caused by arbitrary arrangement and text distortion. The residual network (ResNet) is used in the encoder to encode the rectified images and obtain the visual feature sequences. Based on bidirectional long-short-term memory (Bi-LSTM), the visual feature sequences are converted into textual feature sequences and the text content is obtained by the decoder with attention mechanism. However, when the scene text images are severely distorted and the rectification is insufficient, the visual features extracted from the encoder are not sufficient. Based on the ASTER model, the ESIR model via iterative rectifications is proposed [18], which iteratively removes text distortion as driven by better recognition performance. However, for the images with few distortions, the rectified images obtained through the rectification network will be over rectified, the resolution of images will be reduced, and the rectified images will even lose some edge information. Meanwhile, to reduce the recognition errors caused by insufficient rectification, the SAR model is proposed [19], which abandons the rectification network and adopts a two-dimensional attention module to process the two-dimensional visual feature maps from the encoder. Then, the text characters are located and recognized by considering the regional information of each position in the feature maps.

In the encoder-decoder framework, the attention mechanisms are usually used in the decoder of the models. Most of the attention-based methods usually suffer from serious alignment problems due to its recurrent alignment operation, where the alignment relies on historical decoding results. Cheng et al. [20] put forward the concept of attention drift. The attention mechanism is easily affected by some problems, such as image blurring and complex background, and cannot get accurate alignment between feature maps and the targets of input images. The DAN model is proposed to alleviate the problem of attention drift [21]. The attention maps are obtained through the convolution alignment module based on visual features from the encoder. Moreover, a decoupled decoder is used to make the final prediction by jointly using the visual feature maps and attention maps. In addition, the RobustScanner model is proposed to alleviate the problem by introducing two branches [22]. Specifically, the position enhancement branch is specially designed to improve the ability of position encoding in the decoder. The hybrid branch is the traditional decoder with attention mechanism. The outputs of the two branches are combined through the dynamic fusion module and connected to an elementwise gate mechanism in the channel dimension. By the selection of features, the RobustScanner can adaptively adjust the importance of contextual information and positional information to obtain better performance of text recognition. However, the above models are complex in terms of the model structure.

Since the transformer [23] has achieved remarkable achievements in the tasks of natural language processing, researchers are beginning to explore its application in the field of STR. The 2DOCR model is proposed [24], which uses the transformer to decode twice at the end of the encoder. The second decoder is fine-tuned and optimized on the basis of the result of the first decoding, which can effectively improve the recognition performance. Moreover, the Bi-STET model is proposed [25] to solve the problem of information loss in the process of converting visual features to textual features. After extracting visual features from the ResNet, the encoder of the transformer is used to enhance the visual features, to better integrate visual information and text information. Besides, text recognition on the decoder of the transformer also has better recognition effect.

However, integrating the transformer into the STR models greatly increases the number of model parameters and training time. Therefore, researchers try to introduce auxiliary network modules with a relatively low number of parameters to improve the recognition accuracy when facing the problems of image blurring and low resolution. Wang proposes the TextSR model [13], which introduces a content-aware text super-resolution network SRGAN to restore low-resolution images with super-resolution images under the guidance of adversarial loss and then uses the ASTER model to identify the text content of super-resolution images. To solve the problems of low brightness in images and text occlusion, the SPIN model [26] proposes the structure preserving network (SPN) and the auxiliary inner-offset network (AIN), respectively. Specifically, SPN adjusts the intensity value between pixel points based on the structure-preserving transformation to alleviate the problem of low brightness in images. Based on the theory of offset from geometric transformation, the AIN introduces colour offsets to distinguish the colour intensity, to alleviate the problems of text occlusion and shadow. To solve the problem of complex background, the SCGAN model is put forward [27], which outputs binary images through the generator and inputs into the attention-based decoder to generate the attention feature maps. After the fusion of binary images and attention feature maps, the recognized texts are outputted to the discriminator and compared with the ground truth texts. The loss is propagated back to optimize the network parameters of the generator and to improve the recognition performance. The SEED model is proposed [28] to alleviate the problems of uneven illumination and incomplete characters. Based on the ASTER model, the SEED model innovatively introduces the pretrained language model FastText in the stage of visual features conversion to textual features. Moreover, the cosine embedding loss is calculated with semantic information and word embedding of target texts from the FastText, to supervise the effect of feature extraction in the encoder, to obtain more comprehensive text information and better recognition results.

3. HRNet Encoder and Dual-Branch Decoder Framework-Based Scene Text Recognition Model

This paper proposes a STR model based on HRNet encoder and dual-branch decoder framework, as shown in Figure 1. A single scene text image is taken as the input, and after the process of TPS network and Gaussian blur, the encoder module and dual-branch decoder module are adopted, in which the super-resolution image and recognized text are outputted by the super-resolution branch and recognition branch, respectively. Specifically, HRNet is adopted as the feature extraction network in the encoder module to output the feature maps with multiple resolutions. The SAM is acted on the feature maps with the highest resolution to strengthen the learning of important feature information. The input of the super-resolution branch is the feature maps with the highest resolution enhanced by the SAM, and the super-resolution image is generated by upsampling through Trans Conv2D. The input of the recognition branch is the feature maps with multiple resolutions. Through the Independent Trans Conv2D Layers, the lower resolution feature maps are expanded, so that the final multiscale feature maps can be fused in the channel dimension. The attention-based decoder is used to decode the fused feature maps to obtain text recognition results. In the parallel dual-branch decoder module, the super-resolution branch and the recognition branch together enhance the feature extraction effect of the encoder module and then improve the effect of STR. In the testing stage, the super-resolution branch is abandoned to simplify the model and reduce the complexity of the model.

3.1. Encoder Module. The encoder module of the model is shown in Figure 2, which innovatively adopts HRNet as the feature extraction network and maintains a high-resolution representation throughout the whole process. A high-resolution subnet is taken as the first stage, and multiresolution subnets from high to low are added one by one to form more stages. The multiresolution subnets are connected in parallel, and the information is repeatedly exchanged during the whole process to perform the multifeature fusion. At the end of the encoder module, the SAM is used to strengthen the learning of important feature information of the feature maps with the highest resolution outputted by the HRNet encoder. The feature with less information is suppressed by using the attention mask, so that the encoder module can transfer the most effective learned features to the super-resolution branch and the recognition branch. Finally, different connection operations are adopted according to the different purposes of the super-resolution branch and the recognition branch. The feature maps with the highest resolution enhanced by the SAM are inputted to the super-resolution branch, and four kinds of feature maps with different resolutions are inputted to the recognition branch.

The SAM is constituted by a series of convolution operation and sigmoid activation function, as shown in Figure 3. The feature maps with the highest resolution are added to the input image after the 1×1 convolution operation; that is, the feature maps are supervised by the input image. Then, the attention maps are obtained by the activation function and then are acted on the feature maps by weighted summation. In this way, important features can be enhanced and features with less information can be suppressed.

3.2. Super-Resolution Branch. The super-resolution branch of the model employs the Trans Conv2D for upsampling on feature maps with the highest resolution enhanced by the SAM, to restore the super-resolution images. No extra super-resolution network is introduced, the super-resolution branch is directly connected to the encoder module, and a simple upsampling recovery operation is adopted, so the super-resolution branch of the proposed model is more dependent on the feature maps outputted from the encoder module. The effect of feature extraction of the HRNet encoder is strengthened through the supervision of the super-resolution branch. Meanwhile, the super-resolution branch is only used in the training stage and is abandoned in the testing stage, which helps to reduce the model complexity. The Trans Conv2D is composed of 3×3 transposed convolution operation, BatchNorm layer, and ReLu layer. The average absolute error loss L_{sr} of the restored super-resolution image and the original image is calculated, as shown in

$$L_{\mathrm{sr}} = \frac{1}{W \times L} \sum_{i=1}^{W} \sum_{j=1}^{L} \left\| O^{i,j} - I^{i,j} \right\|, \tag{1}$$

where W and L represent the width and length of the image, respectively, O represents the super-resolution image restored by the super-resolution branch, and I represents the original scene text image.

3.3. Recognition Branch. The recognition branch of this proposed model consists of a multiscale fusion structure and an attention-based decoder structure. Specifically, in the multiscale fusion structure, in contrast to expanding the size of the feature maps by bilinear interpolation, the Independent Trans Conv2D Layer is used on all low-resolution feature maps, to obtain the feature maps with the same size as the feature maps with the highest resolution. The resolutions of feature maps decrease from the top to bottom, and the number of input channels and output channels of a single Independent Trans Conv2D Layer is determined according to the size of corresponding feature maps. Furthermore, the multiscale fusion is carried out in the channel dimension through the splicing operation, as shown in Figure 4. Then, by employing the channel attention mechanism [29], the weights on different channels of the multiscale feature maps are calculated and important channels of the feature maps are adaptively selected to help the network obtain more effective information.

After obtaining the structure of multiscale fusion, the attention-based decoder is connected to achieve complete text recognition. To realize effective sequence conversion from visual features to textual features, the multiscale feature maps are processed by a 3×3 basic convolution module in

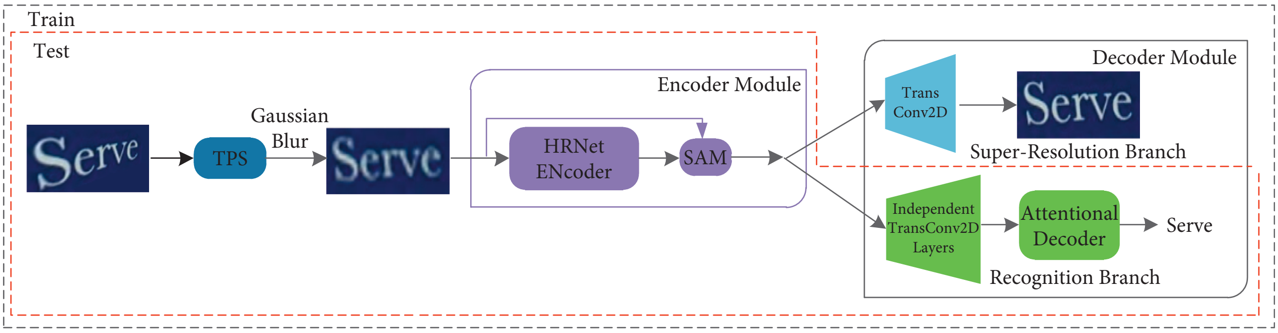

Figure 1: Structure of the proposed model.

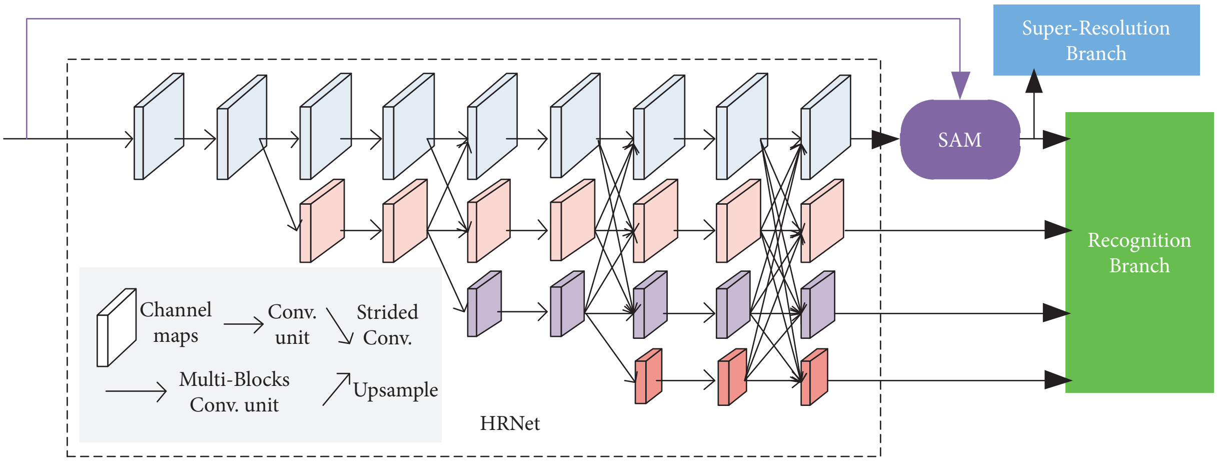

Figure 2: Structure of the encoder module.

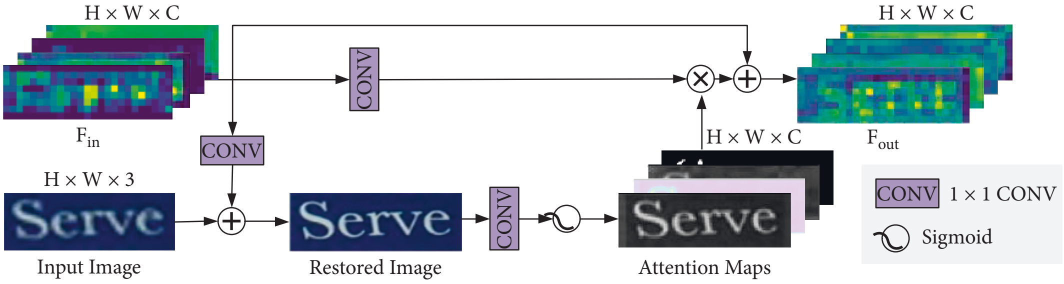

Figure 3: Structure of the SAM.

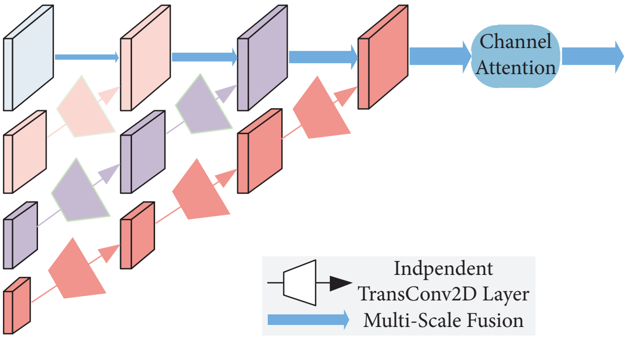

Figure 4: Structure of multiscale fusion in the recognition branch.

the HRNet, to adjust the channel numbers without changing the size of the feature maps and then rearrange the dimension of the feature maps. In other words, the channel dimension and the width dimension of feature maps are converted, to transform the two-dimensional visual feature maps into one-dimensional textual feature vectors. Then, the semantic information of one-dimensional feature vectors is strengthened through the Bi-LSTM network. Finally, the textual feature vectors are decoded by the GRU based on the attention mechanism to recognize the characters, as described in the ASTER [10]. The structure of the attention-based decoder is shown in Figure 5, and <EOS> represents the last character of the text sequence.

The sequence cross entropy loss L_{SCE} is calculated between the recognized text and the ground truth text, as shown in equation (2), to improve the decoding effect of the decoder module and the feature extraction effect of the encoder module and then improve the accuracy of STR.

$$L_{\mathrm{SCE}} = -\frac{1}{\mathrm{MN}} \sum_{i=1}^{M} \sum_{j=1}^{N} y_{i,j} \log\left(s_{i,j}\right), \tag{2}$$

where M represents the number of samples in a batch, N represents the number of text characters, y represents the ground truth text, and s represents the recognized text of the proposed model.

The loss function of the proposed model is shown in (3), where λ_1 is the corresponding weight parameter of the super-resolution branch loss and λ_2 is the corresponding weight parameter of the recognition branch loss.

$$L = \lambda_1 L_{\mathrm{sr}} + \lambda_2 L_{\mathrm{SCE}}. \tag{3}$$

4. Experiments and Results

The experimental environment of the proposed model is based on Pycharm integrated development environment, the PyTorch deep learning framework is adopted, and hardware is based on 1 NVIDIA GeForce GTX 2080Ti 11GB GPU. According to the unified experimental data and effective comparison models advocated by Baek et al. [9], the training data are the public synthetic datasets Synth90K [30] and SynthText [31] and the testing data are the testing set of seven natural scene datasets. The verification data are the training set of seven natural scene datasets. The seven natural scene datasets are as follows: IIIT5K-Words (IIIT5k) [32] refers to the regular scene text images such as billboards and posters in Google image search. Street View Text (SVT) [33] refers to the regular outdoor images in Google street view. ICDAR 2003 (IC03) [34] is a competition-based regular dataset published by the ICDAR conference, excluding scene text images of less than three characters or non-alphanumeric. ICDAR 2013 (IC13) [35] is a regular dataset, which is mostly taken from the IC03 dataset and expands some clear scene text images such as road signs and book covers. ICDAR 2015 (IC15) [36] is an irregular dataset, which mostly consists of some random images of blurred and occluded in streets or shopping malls. SVT-Perspective (SVTP) [37] refers to the irregular scene text images with perspective interference in Google street view. CUTE80 [38] mainly contains distorted and irregular scene text images.

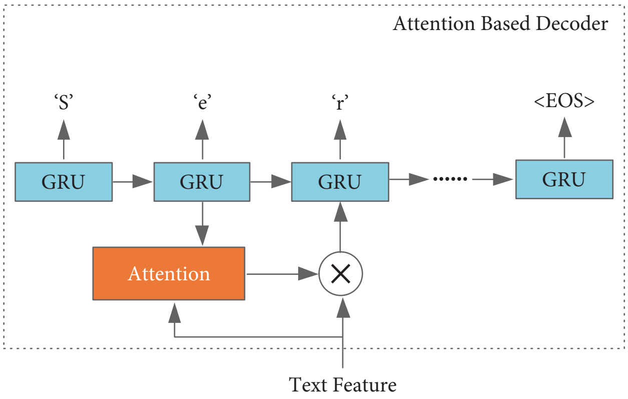

Figure 5: Structure of the attention-based decoder in the recognition branch.

The scene text images as input of the network are three-channel RGB images with a unified size of 64 × 256, and the size of the images is unified to 32 × 100 after TPS. The four kinds of feature maps with different resolutions outputted by the encoder module are 8 × 25, 4 × 13, 2 × 7, and 1 × 4, from the feature maps with the highest resolution to the feature maps with the lowest resolution, respectively. Due to the setting of super-resolution branch, pairs of low-resolution images and high-resolution images are required. Therefore, to simulate the recovery process of super-resolution network, the original image after random Gaussian blur is used as a low-resolution image and the original image is used as a high-resolution image. The Adadelta optimizer is used to update the network parameters, the weight attenuation factor is set as 0.1, the initial training learning rate is 1, and the fine-tuned training learning rate is 0.1. To ensure that the values of L_{sr} and L_{SCE} are in the same magnitude, λ_1 and λ_2 is set as 0.1 and 1 and the word accuracy is used as the evaluation metric.

4.1. Experiments of Model Comparison. To evaluate the effect of the proposed model, an experiment is performed to compare with other recent models, as shown in Table 1. For the fairness of comparison, the models using additional datasets for training are not compared. Synth90K and SynthText are used as training sets in all comparative experiments, and no lexicon is provided in the experiments. Word accuracy is taken as the evaluation metric. Meanwhile, the speed of the proposed model is 4.3 ms and 54 ms per image in the training stage and in the testing stage, respectively. Specifically, the proposed model innovatively introduces the HRNet, which combines with some methods such as the super-resolution branch, the SAM, and the Independent Trans Conv2D Layers. Compared with the ASTER and TextSR, the accuracy of the proposed model is improved in most datasets, especially in IC15, SVTP, and CUTE80, which are irregular and blurry, and the accuracy is improved by more than 3%. Compared with the Bi-STET, which uses the transformer to enhance and decode

Table 1: The accuracy comparison between the proposed model and recent models (%).

Model	Benchmark							Average	
	IIIT5k	SVT	IC03	IC13	IC15	SVTP	CUTE80	Regular	Irregular
ASTER	93.4	89.5	94.5	91.8	76.1	78.5	79.5	92.3	78.0
TextSR	92.5	87.2	93.2	91.3	75.6	77.4	78.9	91.0	77.3
ESIR	93.3	90.2	—	91.7	76.9	79.6	83.3	91.7	79.9
2DOCR	94	90.1	94.3	92.7	76.3	82.3	86.8	92.7	81.8
Bi-STET	94.7	89	**96**	93.4	75.7	80.6	82.5	**93.2**	79.6
SEED	93.8	89.6	—	92.8	80	81.4	83.6	92.0	81.6
DAN	94.3	89.2	95	93.9	74.5	80	84.4	93.1	79.6
SPIN	94.7	87.6	93.4	91.5	79.1	79.7	85.1	91.8	81.3
RobustScanner	**95.3**	88.1	—	**94.8**	77.1	79.5	**90.3**	92.7	82.3
SCGAN	94	90	95.6	93.3	81.6	**85.1**	78.1	**93.2**	81.6
Proposed model	93.7	**91.3**	93.3	94.3	**82.8**	83.1	83.0	93.1	**82.9**

Note: bold font is the optimal value in each column, and the underline font is the suboptimal value in each column.

Table 2: Comparison of accuracy of ablation models (%).

Model	IIIT5k	SVT	IC03	IC13	IC15	SVTP	CUTE80
Baseline (HRNet)	91.7	88.4	93.4	92.2	78.6	80.2	80.9
Baseline + SR (Bilinear Interpolation)	93.0	89.5	92.7	92.7	81.1	81.1	78.1
Baseline + SR (Bilinear Interpolation) + SAM	93.0	92.1	91.9	93.2	81.7	83.3	81.2
Baseline + SR (Trans Conv2D) + SAM	93.4	91.8	93.3	93.6	81.8	82.6	81.6
Proposed model	93.7	91.3	93.3	94.3	82.8	83.1	83.0

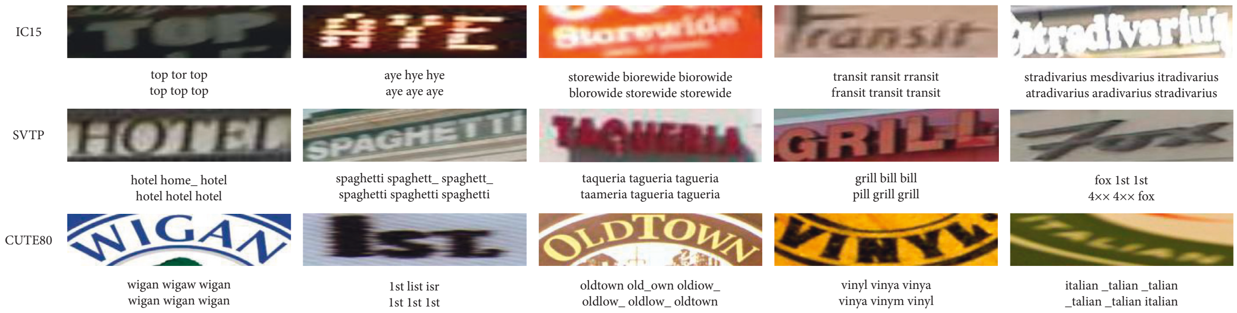

Figure 6: Qualitative comparison of recognition results.

information, the accuracy of the proposed model is also improved in three kinds of irregular datasets. Compared with the SCGAN, which introduces GAN to alleviate the background interference, the recognition accuracy of the proposed model is more balanced for different datasets. In addition, compared with other recent models, the proposed model can achieve better performance in average accuracy for irregular datasets as well as good performance for regular datasets.

4.2. Validity Experiments of the Proposed Methods. To verify the effectiveness of the proposed methods, such as the super-resolution branch, the SAM, and the Independent Trans Conv2D Layers, several comparative experiments are set up. The HRNet is used as the feature extraction network in the baseline model, and the proposed methods are gradually added to fine-tune in ablation models. The baseline model is trained for up to 3 epochs, and the fine-tuned models are trained for up to 4 epochs based on the baseline model. On the whole, the proposed model is trained for up to 13 epochs. The setting of hyperparameters is consistent all time.

Quantitative comparison is made based on the testing sets, and the results are shown in Table 2. Compared with the classical ASTER model in Table 1, which uses ResNet as the feature extraction network, the recognition accuracy of IC15, SVTP, and CUTE80 is improved by 2.5%, 1.7%, and 1.4%, respectively, by using the HRNet in the baseline model. The recognition accuracy is also improved in natural scenes by adding the super-resolution branch composed of bilinear interpolation to the baseline, which verifies that HRNet can be used as both a feature extraction network and a super-resolution network to provide effective high-resolution feature maps. In addition, the recognition accuracy can be further improved by the addition of the SAM, and instead of bilinear interpolation, we use Trans Conv2D as the upsampling method to recover super-resolution images in the super-resolution branch.

As shown in Figure 6, the qualitative comparison is given based on the irregular testing sets, such as IC15, SVTP, and

Table 3: PSNR results of restored images by Trans Conv2D and bilinear interpolation (dB).

	Bilinear interpolation	Trans Conv2D	Improved
IIIT5k	27.57	30.88	+3.31
SVT	31.93	36.05	+4.12
IC03	27.79	31.68	+3.89
IC13	27.82	32.06	+4.24
IC15	33.08	38.11	+5.03
SVTP	32.76	37.86	+5.10
CUTE80	24.98	28.39	+3.41

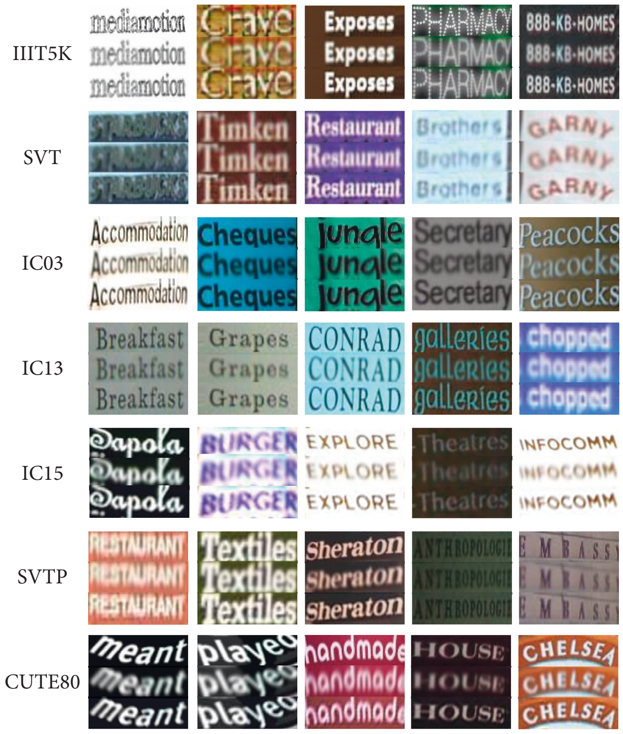

Figure 7: Performance of the adoption of Trans Conv2D in the super-resolution branch. In each image, from the top to bottom are the original image, blurred low-resolution image, and super-resolution image, respectively.

CUTE80. The text content below each picture is in lower case. The first line is the ground truth text, the recognition results of Baseline and Baseline + SR (Bilinear Interpolation), respectively. The second line is the recognition results of the Baseline + SR (Bilinear Interpolation) +SAM, Baseline + SR (Trans Conv2D) + SAM, and the proposed model (Baseline + SR (Trans Conv2D) + SAM + Independent Trans Conv2D Layers), respectively. It can be seen that the baseline model has some problems of misrecognition for individual characters. However, the proposed methods, such as the super-resolution branch, the SAM, and the Independent Trans Conv2D Layers, can be used gradually to effectively recognize the characters, which are relatively difficult to recognize, and then the proposed model can obtain better recognition results.

4.3. Validity Experiments of the Dual-Branch Structure. In the super-resolution branch of the proposed model, the methods of Trans Conv2D and bilinear interpolation are used to compare the effect of image recovery, respectively. The values of PSNR metric of the restored images are calculated, as shown in Table 3. Compared with the bilinear interpolation, the Trans Conv2D could increase the PSNR by more than 3 dB, which verifies the effectiveness of adopting Trans Conv2D in the super-resolution branch. Moreover, qualitative comparison is carried out with regard to the adoption of Trans Conv2D in the super-resolution branch, as shown in Figure 7. Experimental results on seven natural scene datasets verify that super-resolution branch can better accomplish the super-resolution task and assist the feature extraction network to effectively encode the scene text images; therefore, the accuracy of STR can be improved.

In the recognition branch of the proposed model, the Independent Trans Conv2D Layers are used for size expansion. The comparison between the feature maps generated by the Independent Trans Conv2D Layers and bilinear interpolation is shown in Figure 8, and the generated feature maps of five channels are randomly selected. The brighter regions in the feature maps represent the higher feature values of the regions and the more information contained. Four kinds of feature maps with different resolutions are outputted in the encoder module, with sizes of 8×25, 4×13, 2×7, and 1×4, respectively. The single Independent Trans Conv2D Layer is used to expand the size of feature maps with lower resolutions, so that the size of each resolution feature map is the same, that is, 8×25. From Figure 8, it can be seen that the feature maps generated by the Independent Trans Conv2D Layers contain more text information than the bilinear interpolation in the size of 4×13, which can reduce the loss of feature information in the process of size expansion. However, for the size of 2×7, the feature maps generated by the bilinear interpolation can only maintain some edge information, so very little information is transmitted to the recognition branch for text recognition. Meanwhile, the feature maps generated by the Independent Trans Conv2D Layers can retain some visual information even at the lowest resolution. Moreover, the multiscale fusion results transmitted to the attention-based decoder can contain more effective text information. In other words, the recognition effect of the model is significantly improved by several proposed methods on various testing sets, as shown in Table 2. As shown in Table 4, the three ablation models and the proposed model all use the super-resolution branch in the training stage and abandon it

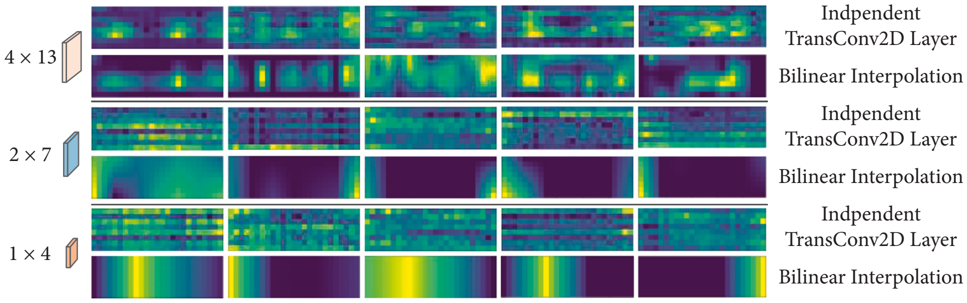

Figure 8: Comparison on the feature maps generated by the Independent Trans Conv2D layers and bilinear interpolation.

Table 4: Parameter comparison in ablation models during training and testing (M).

Model	Parameters	
	Training	Testing
Baseline (HRNet)	35.564	35.564
Baseline + SR (Bilinear Interpolation)	35.565	35.564
Baseline + SR (Bilinear Interpolation) + SAM	35.568	35.567
Baseline + SR (Trans Conv2D) + SAM	35.573	35.567
Proposed model (Baseline + SR (Trans Conv2D) + SAM + Independent Trans Conv2D Layers)	37.582	37.576

in the testing stage, which can reduce the model complexity. Moreover, the proposed model, which adds the effective methods, such as super-resolution branch, the SAM, and the Independent Trans Conv2D Layers, does not increase too many model parameters.

5. Conclusions

This paper proposes a HRNet encoder and dual-branch decoder framework-based STR model to recognize distortion, blurred, and low-resolution text in natural scenes. Based on the encoder-decoder framework, the model innovatively introduces the HRNet as feature extraction network and introduces the SAM to enhance the learning of important features. The feature maps with multiple resolutions extracted by the HRNet encoder are inputted to the dual-branch decoder module composed of the super-resolution branch and the recognition branch. Specifically, the feature maps with the highest resolution are inputted to the super-resolution branch to restore the super-resolution images and to strengthen the feature extraction effect of the encoder module. After multiscale fusion through the Independent Trans Conv2D Layers in the recognition branch, the four kinds of feature maps with different resolutions are decoded by the attention-based decoder and finally the recognized text is obtained. Through ablation experiments and comparative experiments, the effectiveness of the proposed methods such as the HRNet encoder, the super-resolution branch, and the Independent Trans Conv2D Layers is verified. Compared with the ASTER model and other recent models, the proposed model can better perform STR on multiple public natural scene datasets, especially for the text with distortion, blurring, and low resolution. In the future, STR for images with complex background and jitter imaging will be further studied.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61801159 and 61571174 and Scientific and Technological Innovation Plan for Undergraduates in Zhejiang Province under Grant no. 2021R426066.

References

[1] S. Long, X. He, and C. Yao, “Scene text detection and recognition: the deep learning era,” *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.

[2] C. Y. Liu, X. X. Chen, C. J. Luo, L. Jin, Y. Xue, and Y. Liu, “Deep learning methods for scene text detection and recognition,” *Journal of Image and Graphics*, vol. 26, no. 06, pp. 1330–1367, 2021.

[3] J. X. Wang, Z. Y. Wang, and X. Tian, “Review of natural scene text detection and recognition based on deep learning,” *Journal of Software*, vol. 31, no. 5, pp. 1465–1496, 2020.

[4] X. P. Wang, L. T. Yang, D. Meng, M. Dong, and H. Wang, “Multi-UAV cooperative localization for marine targets based on weighted subspace fitting in SAGIN environment,” *IEEE Internet of Things Journal*, vol. 9, no. 8, 2021.

[5] H. Wang, L. Wan, M. Dong, K. Ota, and X. Wang, “Assistant vehicle localization based on three collaborative base stations via SBL-based robust DOA estimation,” *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5766–5777, 2019.

[6] X. Wang, L. Wan, M. Huang, C. Shen, and K. Zhang, “Polarization channel estimation for circular and non-circular signals in massive MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 1001–1016, 2019.

[7] 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.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proceedings of the Neural Information Processing Systems*, pp. 3104–3112, Montreal, Canada, December 2014.

[9] J. Baek, G. Kim, J. Lee et al., “What is wrong with scene text recognition model comparisons? dataset and model analysis,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4715–4723, Seoul, Korea (South), November 2019.

[10] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, “ASTER: an attentional scene text recognizer with flexible rectification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.

[11] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[12] L. Christian, T. Lucas, H. Ferenc et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4681–4690, Honolulu, HI, USA, July 2017.

[13] W. Wang, E. Xie, P. Sun et al., “TextSR: Content-Aware Text Super-resolution Guided by Recognition,” 2019, https://arxiv.org/abs/1909.07113.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, July 2018.

[15] Y. Mou, T. Lei, H. Yang et al., “PlugNet: degradation aware scene text recognition supervised by a pluggable super-resolution unit,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 158–174, November 2020.

[16] W. Wang, E. Xie, X. Liu et al., “Scene text image super-resolution in the wild,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 650–666, November 2020.

[17] J. Wang, K. Sun, T. Cheng et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[18] F. Zhan and S. Lu, “ESIR: end-to-end scene text recognition via iterative image rectification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2059–2068, April 2019.

[19] H. Li, P. Wang, C. Shen, and G. Zhang, “Show, attend and read: a simple and strong baseline for irregular text recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8610–8617, 2019.

[20] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, “Focusing attention: towards accurate text recognition in natural images,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5076–5084, September 2017.

[21] T. Wang, Y. Zhu, L. Jin et al., “Decoupled attention network for text recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12216–12224, 2020.

[22] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, “RobustScanner: dynamically enhancing positional clues for robust text recognition,” *Computer Vision - ECCV 2020*, vol. 12364, pp. 135–151, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” *Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

[24] P. Lyu, Z. Yang, X. Leng, X. Wu, R. Li, and X. Shen, “2D Attentional Irregular Scene Text Recognizer,” 2019, https://arxiv.org/abs/1906.05708.

[25] M. Bleeker and M. de Rijke, “Bidirectional scene text recognition with a single decoder,” in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, December 2019.

[26] C. Zhang, Y. Xu, Z. Cheng et al., “SPIN: Structure-Preserving Inner Offset Network for Scene Text Recognition,” 2020, https://arxiv.org/abs/2005.13117.

[27] C. Luo, Q. Lin, Y. Liu, L. Jin, and C. Shen, “Separating content from style using adversarial learning for recognizing text in the wild,” *International Journal of Computer Vision*, vol. 129, no. 4, pp. 960–976, 2021.

[28] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, “SEED: semantics enhanced encoder-decoder framework for scene text recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13528–13537, Seattle, WA, USA, June 2020.

[29] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, July 2018.

[30] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, June 2014.

[31] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2315–2324, June 2016.

[32] A. Mishra, K. Alahari, and C. V. Jawahar, “Top-down and bottom-up cues for scene text recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2687–2694, Providence, RI, USA, June 2012.

[33] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1457–1464, Barcelona, November 2011.

[34] S. M. Lucas, A. Panaretos, L. Sosa et al., “ICDAR 2003 robust reading competitions: entries, results, and future directions,” *International Journal on Document Analysis and Recognition*, vol. 7, no. 2, pp. 105–122, 2005.

[35] D. Karatzas, F. Shafait, S. Uchida et al., “ICDAR 2013 robust reading competition,” in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1484–1493, Washington, DC, USA, August 2013.

[36] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou et al., “ICDAR 2015 competition on robust reading,” in *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 1156–1160, ICDAR, Tunis, Tunisia, August 2015.

[37] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, “Recognizing text with perspective distortion in natural scenes,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 569–576, Sydney, NSW, Australia, December 2013.

[38] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, “A robust arbitrary text detection system for natural scene images,” *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014.