Scene Text Image Super-Resolution via Parallelly Contextual Attention Network

Cairong Zhao*†
Shuyang Feng*
Tongji University
Shanghai, China
fengshuyang@tongji.edu.cn
zhaocairong@tongji.edu.cn

Brian Nlong Zhao
Viterbi School of Engineering
University of Southern California
California, USA
briannlongzhao@gmail.com

Zhijun Ding
Tongji University
Shanghai, China
dingzj@tongji.edu.cn

Jun Wu
Computer Science School, Fudan University
Shanghai, China
wujun@fudan.edu.cn

Fuming Shen
University of Electronic Science and Technology of China
Chengdu, China
fumin.shen@gmail.com

Heng Tao Shen
University of Electronic Science and Technology of China
Chengdu, China
shenhengtao@hotmail.com

ABSTRACT

Optical degradation blurs text shapes and edges, so existing scene text recognition methods have difficulties in achieving desirable results on low-resolution (LR) scene text images acquired in real-world environments. The above problem can be solved by efficiently extracting sequential information to reconstruct super-resolution (SR) text images, which remains a challenging task. In this paper, we propose a Parallelly Contextual Attention Network (PCAN), which effectively learns sequence-dependent features and focuses more on high-frequency information of the reconstruction in text images. Firstly, we explore the importance of sequence-dependent features in horizontal and vertical directions *parallelly* for text SR, and then design a parallelly contextual attention block to adaptively select the key information in the text sequence that contributes to image super-resolution. Secondly, we propose a hierarchically orthogonal texture-aware attention module and an edge guidance loss function, which can help to reconstruct high-frequency information in text images. Finally, we conduct extensive experiments on TextZoom dataset, and the results can be easily incorporated into mainstream text recognition algorithms to further improve their performance in LR image recognition. Besides, our approach exhibits great robustness in defending against adversarial attacks on seven mainstream scene text recognition datasets, which means it can also improve the security of the text recognition pipeline. Compared with directly recognizing LR images, our method can respectively improve the recognition accuracy of ASTER, MORAN, and CRNN by 14.9%, 14.0%, and 20.1%. Our method outperforms eleven state-of-the-art (SOTA) SR methods in terms of boosting text recognition performance. Most importantly, it outperforms the current optimal text-orient SR method TSRN by 3.2%, 3.7%, and 6.0% on the recognition accuracy of ASTER, MORAN, and CRNN respectively.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Reconstruction**; **Object recognition**; • **Computer systems organization** → **Neural networks**.

KEYWORDS

Scene Text, Super-Resolution, Sequential Information, Attention, Boundary.

ACM Reference Format:
Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fuming Shen, and Heng Tao Shen. 2021. Scene Text Image Super-Resolution via Parallelly Contextual Attention Network. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475469

1 INTRODUCTION

In recent years, research topics around scene text have been very active [37, 43, 47]. Scene text-related research plays a very important role in many computer vision tasks [3, 48]. However, imperfect imaging conditions often hinder the progress of these fields. Under real-world circumstances, due to the large variation in depth of field, low-resolution (LR) text images abound. And it is almost impossible to avoid them completely. This has posed a great challenge for scene text recognition [1, 7], as the shapes and contours of the text in LR images are often blurred, crippling the performance of many existing text recognition algorithms.

Scene text super-resolution (SR) can effectively alleviate the above-mentioned problem. Previous works [13, 41, 51] propose the use of SR methods to improve the performance of text recognition.

*Both authors contributed equally to this research.
†Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475469

并行上下文注意块

分层正交纹理感知注意模块

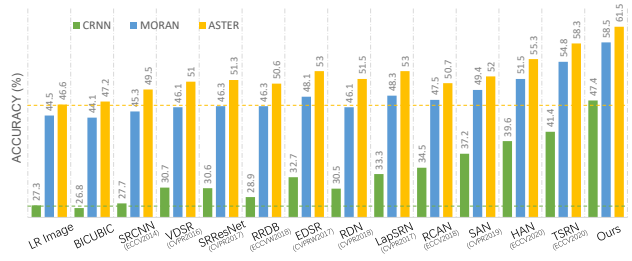

Figure 1: The average recognition accuracy of the three mainstream text recognition methods on the TextZoom dataset [43]. By comparing with eleven SOTA SR methods, our method is able to achieve competitive performance.

However, their methods focus on synthetic data, showing limited generalization capability to real-world text images. Specifically, they [13, 41, 51] conduct many experiments on the ICDAR2015 TextSR competition [34] (only synthetic LR-HR data provided), which means that their model may simply learn the inverse mapping algorithm. This training strategy is more common for early generic SR studies that lack real paired data. However, the trained models often fail in real scenarios. To address this problem, [43] first propose a real paired scene text SR dataset, and their proposed text-orient model achieves the best results in terms of modeling sequence contextual information. However, most of the above methods still have the following problems. Firstly, compared to the generic SR, text SR is characterized by sequential correlation, while the existing methods do not effectively investigate sequence contextual information. Secondly, as a significant high-frequency component, the edge contour information of the text has not been taken into sufficient consideration. Thirdly, the existing methods fail to fully exploit the hierarchical features of the different layers in the network.

To address the above problems, we propose a text-specific end-to-end SR method, named PCAN. First of all, we explore the importance of vertical and horizontal contextual modeling steps. Intuitively, horizontal contextual information focuses on high-level text semantics between characters, and vertical contextual information emphasizes the relationship between stroke features within characters. Based on our experimental analysis, although both contextual features contribute to the text SR task, sequential modeling of both leads to less promising performance. Therefore, we propose a parallelly contextual attention block (PCAB) which captures the horizontal and vertical sequence information *separately* and applies the attention mechanism to weight the above two information. In addition, to make full use of the edge profile information of the text, we design a hierarchically orthogonal texture-aware attention module to obtain a larger receptive field at a lower computational cost, allowing the network to re-weight features to focus on high-frequency information adaptively. Also, inspired by the Sobel operator, we use convolutional operation to generate a Sobel-like edge map, which can supervise the network to effectively capture this high-frequency information. Finally, we exhibit the robustness of our approach in terms of defending against adversarial attacks. To the best of our knowledge, we are the first to find that text SR not only improves recognition accuracy on LR images, but also defends against adversarial perturbations to a certain extent, i.e., it can improve the security of the text recognition pipeline. In summary, our method achieves very competitive performance against state-of-the-art methods (see in Fig. 1).

Our contributions can be summarized as follows:

- We propose a novel module that can effectively capture the horizontal and vertical correlations of text images in a *parallel* way. To achieve the optimal aggregation of horizontal and vertical information, we propose an attention mechanism to weight these two types of information.
- We propose a hierarchically orthogonal texture-aware attention module and an edge guidance loss function to boost our model for high-frequency information of the reconstruction.
- We make the first attempt to prove that text SR can defend against text adversarial attacks. It is a feasible solution to improve the security of text recognition algorithms.
- Extensive experiments on text SR in real scenarios have demonstrated the effectiveness of our proposed method. Our proposed PCAN algorithm achieves very competitive performance against state-of-the-art methods.

2 RELATED WORK

2.1 Super-Resolution

Image super-resolution (SR), an algorithm that generates plausible high-resolution (HR) images from the information provided by low-resolution (LR) images, has been a very popular research field in recent decades. Traditional SR algorithms, such as interpolation-based methods [52], sparse representation methods [50] , etc., usually reconstructed SR images using information such as pixel color values, image gradients, etc., according to predefined formulas.

In recent years, deep learning based SR algorithms have made a breakthrough and significantly outperform traditional algorithms on various benchmarks. In the deep learning era, single image super-resolution (SISR) is seen as a regression task or a generative task [6, 10, 12, 15, 23–26, 32, 44, 54]. These methods usually required training on a large number of LR-HR image pairs. Today, there is a rich variety of datasets available, including a wide range of targets in natural scenes, such as animals, buildings, food, etc. Some of them acquire real LR-HR image pairs by adjusting the focal length of the cameras [4, 5, 53], while others only provide HR images, and LR images need to be synthesized using some default functions in MATLAB (i.e., BICUBIC interpolation with anti-aliasing) [2, 18, 40]. Since images acquired in real-world scenarios suffer from unknown degradation, the methods proposed for synthetic LR-HR images are difficult to achieve desirable results. Therefore, real SR methods urgently need to receive extensive attention.

2.2 Text Recognition

Traditional text recognition methods usually use a bottom-up strategy, i.e., segmenting and recognizing individual characters first, and then aggregating the results into whole words or sentences. These methods [11, 39] were usually based on predefined formulas to extract hand-crafted features for subsequent classification tasks.

In the era of deep learning, scene text recognition further evolves into a top-down strategy. This strategy has many advantages, as it

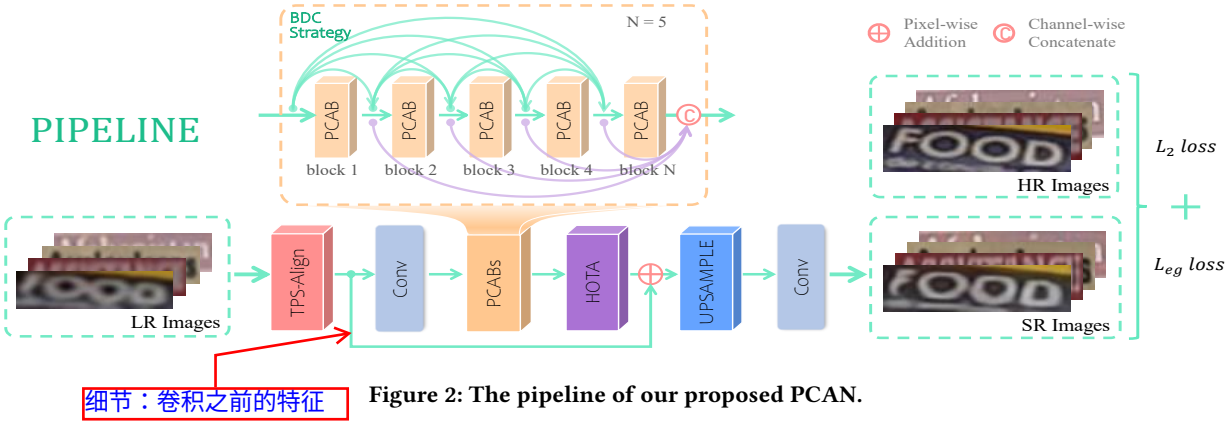

Figure 2: The pipeline of our proposed PCAN.

allows end-to-end processing of variable-length text images, significantly reducing the complexity of the model and speeding up processing. According to the different ways of computing the loss function, existing methods can be classified into the following two paradigms: CTC-based approach and attention-based approach. CRNN [37] was the first to introduce RNN into an end-to-end text recognition method and use the CTC loss function [14] to solve the character feature alignment problem, thus making a sequence prediction scheme with variable-length text images as input feasible. Recently, attention-based recognition methods, represented by ASTER [38] and MORAN [28], have made great progress. These two methods [28, 38] respectively proposed new text rectify modules, thus, their recognition performance was significantly improved for curved text images. Furthermore, many approaches had been proposed to improve the shortcomings of existing methods in various aspects such as font information interference, attention drift problems, etc. [8, 45]. Same as [43], we also choose state-of-the-art methods, ASTER [38], MORAN [28] and CRNN [37] as our baseline recognizers to evaluate the recognition accuracy of the SR images.

2.3 Scene Text Image Super-Resolution

The purpose of scene text SR is to enhance the performance of text recognition algorithms on LR images. [30] analyzed the effect of using different artificial operators on text image SR. [33] proved that the convolution-transposed convolution structure is useful for document SR. [19] proposed a lightweight text SR framework and deployed the model on the edge device. [13, 46, 51] continuously refreshed the results of the ICDAR2015 TextSR competition [34] by improving the existing generic SR model. However, these methods may not generalized in complex real scenarios, i.e., what the model learns may only be the inverse mapping of the down-sampling function [43].

Recently, [43] constructed a real text SR dataset, which provided real LR images sampled at different focal lengths. They demonstrated the necessity of text SR research, proposed a reasonable evaluation metric (the accuracy of mainstream recognition methods on SR images), and designed a text-specific SR network, which can significantly improve the performance of text recognition methods on LR images.

3 METHOD

In this section, we will introduce in detail our proposed PCAN. Specifically, we start with an overall introduction to the network architecture, then, focus on our proposed basic block (PCAB) and adopt a more efficient block-stacking strategy. Besides, we introduce the attention mechanism to text SR, building the hierarchically orthogonal texture-aware attention on the hierarchical features extracted from the previous basic blocks. Finally, we focuses on the optimization of the model. An edge guidance loss is proposed to explicitly extract edge contour information from HR images to assist network training.

3.1 Network Architecture

As shown in Fig. 2, the proposed PCAN mainly consists of three parts: Thin-Plate-Spline (TPS) Align module, Parallelly Contextual Attention Block (PCAB), and Hierarchically Orthogonal Texture-aware Attention (HOTA) module.

First of all, the pre-processed LR input is first rectified by a TPS-align module, which is first introduced to text SR by TSRN [43] and has been widely known in the field of scene text recognition [1, 38]. After that, we extract low-dimensional texture information through several CNN layers, and contextual information through a set of PCABs. Inspired by DenseNet [17], we cascade several PCABs by blocks-level dense connection (BDC) strategy. Finally, we construct the attention mechanism on all PCABs' output features and upsample these features with the magnification factor of $\times 2$. The SR image can be obtained by the last convolution.

3.2 Parallelly Contextual Attention Block

The text SR algorithm can be used as a text image enhancement step to improve the recognition accuracy of text recognition algorithms on LR images. In contrast to natural images, text images contain distinct contextual relationships between characters. Previous work [43] inspires us that contextual information in two orthogonal directions (horizontal and vertical), is both helpful for reconstructing text images. However, they sequentially model two sets of context dependencies, i.e., first construct recurrent connections of visual features in the vertical direction using RNN, and then similarly construct recurrent connections in the horizontal direction on the features obtained from the previous processing. The above two operations are separated by only one layer of CNN

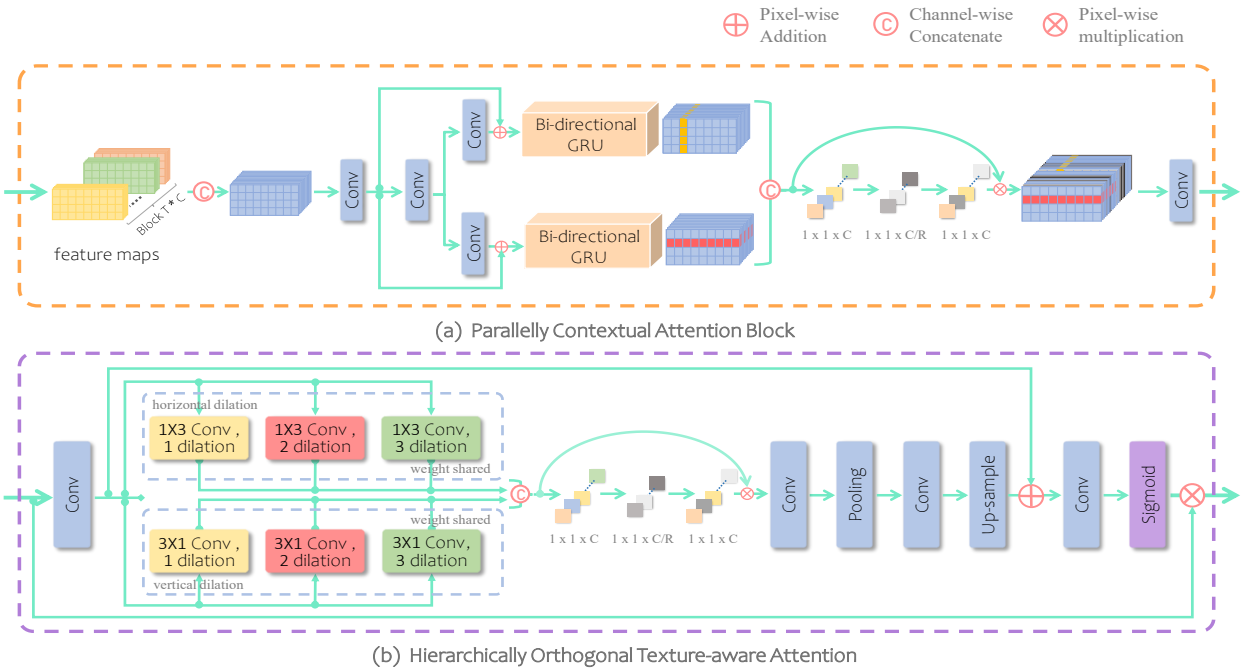

Figure 3: The illustration of PCAB and HOTA modules. In Fig. (a), the PCAB module aims to model the contextual dependencies between orthogonal visual features. In Fig. (b), the HOTA module aims to construct the global attention for larger perception fields.

with limited nonlinear capability. Empirically, we find that a two-branch structure, which parallelly constructs feature dependencies in the vertical and horizontal directions, provides better results than sequentially modeling way. Besides, the use of channel attention adaptively weighting the above two features can further improve the performance of our algorithm (see in 4.3).

Based on the above findings, we propose the parallelly contextual attention block (PCAB), shown in Fig. 3 (a). To balance the computational efficiency and performance, we choose the GRU unit [9] as the basic unit of RNN. Bidirectional RNNs are used to construct sequence-dependent features in the vertical and horizontal directions, respectively, and updates its internal state recurrently in the hidden layer.

$$H_{t_1} = \varphi\left(H_{t_1-1}, (X_r + X_h)\right), \qquad t_1 = 1, 2, ..., W \tag{1}$$

$$H_{t_2} = \varphi\left(H_{t_2-1}, (X_r + X_v)\right), \qquad t_2 = 1, 2, ..., H \tag{2}$$

where H_t denotes RNN hidden layers, X_r, X_h and X_v respectively denote the trunk and branch features processed by the CNN layers. Intuitively, the horizontal direction modeling is used to construct character-to-character dependencies, while the vertical direction modeling is used for textural contexts within characters. Inspired by SENet [16], we use a channel attention mechanism to weight the features in both directions, which gives the network the ability to adaptively adjust the importance of different contexts, thus capturing contextual information in text images more effectively.

$$F_{out} = f_c\left(Concat\left(O_h, O_v\right)\right) \tag{3}$$

O_h and O_v respectively denote the RNN output features of recurrent connection in horizontal and vertical directions. The function f_c represents the channel attention mechanism, which adaptively generates a weighting descriptor $w \epsilon \mathbb{R}^c$ constrained to $[0, 1]$ by Sigmoid. Unlike the SE module [16], our approach generates a channel attention vector that can be used to control the flow of information from the two-branch context-dependent features, while the former only makes a selection for visual patterns extracted from the same CNN layer.

By stacking blocks, the network can usually achieve better performance [25, 43, 55]. Inspired by [17], we adopt a block-level dense connection (BDC) strategy (see in Fig. 2), which not only maximizes the information flow between modules but also make good use of the hierarchical features between modules. In the experimental section (see in 4.3), we will analyze the impact of the number of blocks and the connection strategy on the performance.

3.3 Hierarchically Orthogonal Texture-aware Attention

Previous text SR methods [13, 33, 43, 51] treat features equally in their network layers. However, attention has already caused a profound impact in generic SR research, which can assist the network in adaptively distinguishing features and significantly improving the capacity of model representation. We aim to make the model more focused on the text itself, recovering as much as possible the high-frequency information that is beneficial for the subsequent recognition task. In this section, we propose a hierarchically orthogonal texture-aware attention (HOTA) for scene text SR, which requires less computation to achieve high performance.

改进：
LCT -> SENet

As shown in Fig. 3 (b), since features at different depths have different receptive fields and focus on patterns at different semantic levels, we first collected the output of all PCABs and concatenate them together to obtain the hierarchical features. After one convolutional layer for channel reduction, we will immediately model orthogonal texture features. [47] empirically demonstrate that convolutional operations in the orthogonal direction can efficiently extract texture information of the text. Inspired by them [47], we design convolution kernels in orthogonal directions ($k \times 1$ *and* $1 \times k$ *convolution kernels*) to capture high-frequency information such as text texture and text edges. We assign different dilated coefficients to the $k \times 1$ and $1 \times k$ convolution kernels and share the weights between convolution kernels of the same shape. Thus, we have fewer parameters and significantly larger receptive fields than a single $k \times k$ convolution. After that, the above features are concatenated together and the key features are selected using the channel attention mechanism. Then, we feed the above features to a series of convolution, max pooling, and up-sampling operations, and sum with the residual features after channel reduction. Finally, after one convolution recovering the channel dimension, the attention map of the input hierarchical features is obtained by Sigmoid, which can be multiplied with the input features to achieve our purpose.

3.4 End-to-end Optimization

Text shape and edge as a more discriminative characteristic can help text SR, detection, and recognition to achieve better results [29, 43, 47]. When we use Sobel high-pass filtering to process LR and HR text images separately, we find that there are large high-frequency information differences between them. Inspired by Sobel operator, we propose edge guidance loss for auxiliary training of our network, which can be formulated as follows:

$$f(x) = \sqrt{(Conv(x, W_h))^2 + (Conv(x, W_v))^2 + \varepsilon} \tag{4}$$

$$L_{EG} = E_y ||f_{sr}(\hat{y}) - f_{hr}(y)||^2 \tag{5}$$

here, we use convolution to equivalently implement the Sobel operator. W_v and W_h respectively denote kernels of the Sobel operator in the vertical and horizontal directions. The function $f(x)$ integrates the edge profile information of the image x. Finally, our L_{EG} computes the squared term expectation of the difference between the $f(x)$-function of the network output SR images and HR images, respectively. The total optimization function can be formalized as follows:

$$L_{total} = \lambda_1 L_2(\hat{y}, y) + \lambda_2 L_{EG}(\hat{y}, y) \tag{6}$$

here, λ_1 and λ_2 are pre-specified hyper-parameters that can balance the importance of the above two terms L_2 and L_{EG}, the former being the overall information that needs to be reconstructed by the SR network, and the latter focusing on the edge contour details.

4 EXPERIMENTS

4.1 Datasets and Evaluation

TextZoom [43] including 17367 LR-HR image pairs as the training set with the magnification factor of ×2. It is cropped from two paired SISR datasets: RealSR [4] and SRRAW [53] and additionally provides text annotation. The TextZoom divides the test data into 1619 easy samples, 1411 medium samples, and 1343 hard samples according to their different shooting focal lengths.

Other 7 Real-world STR Datasets. They are CUTE80 [36], ICDAR2003 [27], ICDAR2013 [22], ICDAR2015 [21], IIIT5K-Words (IIIT5K) [31], Street View Text (SVT) [42], and SVT Perspective (SP) [35]. The above data are used in [1] for a fair comparison between different text recognition algorithms.

Evaluation. As in [43], this task focuses on improving the accuracy of text recognition on real LR images by designing a text-specific SR algorithm as the image processing step of the text recognition algorithm. We mainly evaluate the effectiveness of our approach using the recognition accuracy of open-source text recognition algorithms, and the metrics PSNR and SSIM as reference.

Table 1: Different contextual feature modeling approaches and experimental results. 'S' and 'P' respectively denote sequential and parallel connection. 'H' and 'V' respectively denote horizontal and vertical directions.

Configuration			Accuracy of ASTER [38]			
approach	manner	attention	easy	medium	hard	average
2 × V	S	×	73.4%	53.4%	38.7%	56.3%
2 × H	S	×	74.1%	55.5%	40.1%	57.7%
H + V	S	×	74.8%	55.7%	39.6%	57.8%
H + V	S	√	75.2%	57.1%	40.7%	58.8%
H + V	P	×	75.1%	57.3%	40.2%	58.6%
H + V	P	√	**75.4%**	**58.0%**	**41.3%**	**59.3%**

4.2 Implementation Details

In the image processing stage, we follow the example of [43, 46], adding a binary mask to the RGB images. All the LR and HR images are respectively up-sampled to 64×16, 128×32 for avoiding down-sample degradation.

During training, the learning rate is fixed at 1e-3, the trade-off weight of L2 loss and Edge-guidance Loss is set to 1 and 1e-4, respectively. We use Adam optimizer with momentum term 0.9. When evaluating recognition accuracy, we use the official released model of ASTER[1] [38], CRNN[2] [37], MORAN[3] [28], and TRBA[4] [1]. All the SR models are trained by 500 epochs with a single NVIDIA RTX 2080ti GPU.

4.3 Ablation Study

Comparison of Sequential Feature Modeling Approaches. As show in Table 1, we compared several different approaches to modeling sequence context information. In the above experiments, we directly cascade five basic blocks without BDC strategy, within which we use different contextual modeling approaches. We can see that modeling sequential features in both horizontal and vertical directions is better than modeling in only a single direction. Moreover, decoupling the recurrent dependence in the orthogonal

[1] https://github.com/ayumiymk/aster.pytorch
[2] https://github.com/meijieru/crnn.pytorch
[3] https://github.com/Canjie-Luo/MORAN_v2
[4] https://github.com/clovaai/deep-text-recognition-benchmark

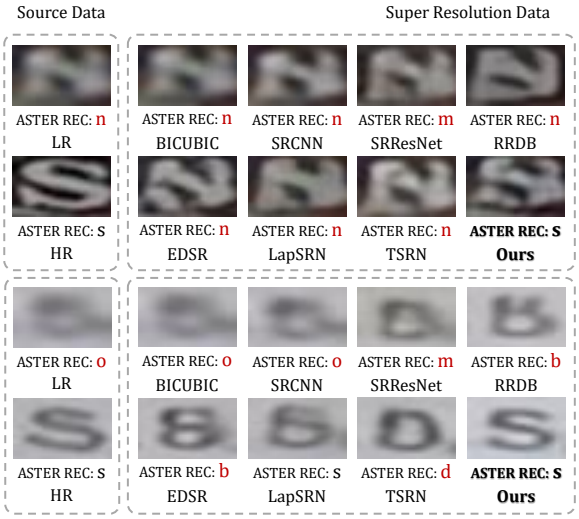

Figure 4: SR image local patches generated by PCAN.

Table 2: Ablation study for different settings of our method PCAN. B denotes BDC strategy and HA denotes HOTA module. L_{EG} denotes edge guidance loss function.

Configuration		Accuracy of ASTER [38]			
Method	Loss	easy	medium	hard	average
SRResNet [25]	L_2+L_{tv}+L_p	69.6%	47.6%	34.3%	51.3%
TSRN [43]	L_2+L_{GP}	75.1%	56.3%	40.1%	58.3%
5×PCAB	L_2	75.4%	58.0%	41.3%	59.3%
5×PCAB+B	L_2	76.8%	57.6%	41.3%	59.7%
5×PCAB+HA	L_2	76.4%	57.8%	41.0%	59.5%
5×PCAB+B+HA	L_2	76.8%	60.2%	42.6%	60.9%
5×PCAB+B+HA	L_2+L_{EG}	**77.5%**	**60.7%**	**43.1%**	**61.5%**

direction, using parallel modeling way will significantly outperform sequential connections, bringing a performance gain of average 0.8% to ASTER [38]. Furthermore, we find that using channel attention to weight the features after modeling in both direction led to better experimental results. Therefore, the contextual information between characters in the horizontal direction is beneficial for text super-resolution, while the sequence-related features in the vertical direction are also important due to the highly similar texture details between strokes within characters. However, the importance of the above features may be different for different images, so it is necessary for the network to learn effective ways to combine these two types of features, i.e., feature fusion using the idea of channel weighting. In Fig. 4, our method can reconstruct more edge and contour details, thus helping text recognition methods to avoid such bad cases.

Connection Strategy for PCABs. In this part, we discuss how the number of blocks and the different connection strategies can affect the performance of our model. In this experiment, we used both the HOTA module and the Edge Guidance Loss function and our results can be found in Fig. 5. On the one hand, comparing the blue and

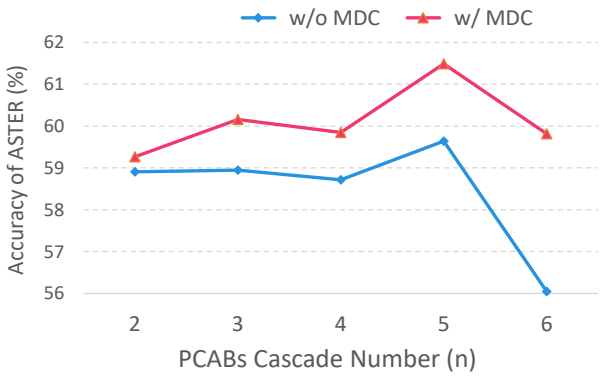

Figure 5: Performance differences in the number of modules and connection methods.

Table 3: PSNR and SSIM results of different SR methods on TextZoom [43].

Method	PSNR			SSIM		
	easy	medium	hard	easy	medium	hard
BICUBIC	22.35	18.98	19.39	0.7884	0.6254	0.6592
SRCNN [12]	23.48	19.06	19.34	0.8379	0.6323	0.6791
VDSR [23]	24.62	18.96	19.79	0.8631	0.6166	0.6989
SRResNet [25]	24.36	18.88	19.29	0.8681	0.6406	0.6911
RRDB [44]	22.12	18.35	19.15	0.8351	0.6194	0.6856
EDSR [26]	24.26	18.63	19.14	0.8633	0.6440	0.7108
RDN [55]	22.27	18.95	19.70	0.8249	0.6427	0.7113
LapSRN [24]	24.58	18.85	19.77	0.8556	0.6480	0.7087
RCAN [54]	22.15	18.81	19.83	0.8525	0.6465	0.7227
SAN [10]	22.69	18.77	19.82	0.8597	0.6477	0.7280
HAN [32]	23.30	19.02	20.16	0.8691	0.6537	0.7387
TSRN [43]	**25.07**	18.86	19.71	**0.8897**	0.6676	0.7302
PCAN (ours)	24.57	**19.14**	**20.26**	0.8830	**0.6781**	**0.7475**

red line, we can find that the method using the blocks-level dense connection (BDC) strategy is significantly better than a simple and direct connection. When the network depth reaches a certain level, BDC strategy can make the performance degradation smoother. One possible reason is that the BDC strategy allows better use of the hierarchical features, and makes the deep network adaptively degradable so that the model can be easily optimized. On the other hand, looking only at one of the blue or red line, we can find that as the number of PCABs increases, the SR images reconstructed by our model are more effective. However, deeper networks inevitably increase the complexity of optimization, and the performance gains from cascade modules are not endless. Therefore, we can get the best results when using five PCABs.

Comprehensive analysis. As shown in Table 2, we ablate each of our proposed modules, in turn, to demonstrate that all the modules can work simultaneously. At the same time, we compare our method with SRResNet [25] and TSRN [43], which are more close to our network design philosophy. Inspired by TSRN [43], the PCAN uses four channel inputs (RGB+Mask) and uses TPS alignment modules,

Table 4: Performance of mainstream SISR algorithms on the three subsets in TextZoom. L_{tv} denotes Total Variation Loss. L_p denotes Perceptual Loss proposed in [20]. $L_{Charbonnier}$ denotes the Charbonnier Loss proposed in LapSRN [24]. L_{GP} denotes the Gradient Prior Loss proposed in [43]. L_{EG} denotes our proposed Edge-guidance Loss. Improvement of PCAN in the last line represents the accuracy increase of our SR compared to LR.

Method	Loss	Accuracy of ASTER [38]				Accuracy of MORAN [28]				Accuracy of CRNN [37]			
		easy	medium	hard	average	easy	medium	hard	average	easy	medium	hard	average
BICUBIC	–	64.7%	42.4%	31.2%	47.2%	60.6%	37.9%	30.8%	44.1%	36.4%	21.1%	21.1%	26.8%
SRCNN [12]	L_2	69.4%	43.4%	32.2%	49.5%	63.2%	39.0%	30.2%	45.3%	38.7%	21.6%	20.9%	27.7%
VDSR [23]	L_2	71.7%	43.5%	34.0%	51.0%	62.3%	42.5%	30.5%	46.1%	41.2%	25.6%	23.3%	30.7%
SRResNet [25]	L_2+L_{tv}+L_p	69.4%	47.3%	34.3%	51.3%	60.7%	42.9%	32.6%	46.3%	39.7%	27.6%	22.7%	30.6%
RRDB [44]	L_1	70.9%	44.4%	32.5%	50.6%	63.9%	41.0%	30.8%	46.3%	40.6%	22.1%	21.9%	28.9%
EDSR [26]	L_1	72.3%	48.6%	34.3%	53.0%	63.6%	45.4%	32.2%	48.1%	42.7%	29.3%	24.1%	32.7%
RDN [55]	L_1	70.0%	47.0%	34.0%	51.5%	61.7%	42.0%	31.6%	46.1%	41.6%	24.4%	23.5%	30.5%
LapSRN [24]	$L_{Charbonnier}$	71.5%	48.6%	35.2%	53.0%	64.6%	44.9%	32.2%	48.3%	46.1%	27.9%	23.6%	33.3%
RCAN [54]	L_1	67.3%	46.6%	35.1%	50.7%	63.1%	42.9%	33.6%	47.5%	46.8%	27.9%	26.5%	34.5%
SAN [10]	L_1	68.1%	48.7%	36.2%	52.0%	65.6%	44.4%	35.2%	49.4%	50.1%	31.2%	28.1%	37.2%
HAN [32]	L_2	71.1%	52.8%	39.0%	55.3%	67.4%	48.5%	35.4%	51.5%	51.6%	35.8%	29.0%	39.6%
TSRN [43]	L_2+L_{GP}	75.1%	56.3%	40.1%	58.3%	70.1%	53.3%	37.9%	54.8%	52.5%	38.2%	31.4%	41.4%
PCAN (ours)	L_2+L_{EG}	**77.5%**	**60.7%**	**43.1%**	**61.5%**	**73.7%**	**57.6%**	**41.0%**	**58.5%**	**59.6%**	**45.4%**	**34.8%**	**47.4%**

Table 5: Text SR defense against adversarial perturbations on STR benchmark, where the recognition accuracy is provided by TRBA [1].

Method	IIIT5k	SVT	IC03_860	IC03_867	IC13_857	IC13_1015	IC15_1811	IC15_2077	SVTP	CUTE80	Average
original	87.36%	87.32%	95.11%	94.69%	92.99%	92.21%	78.24%	75.39%	80.15%	74.21%	85.24%
attacked	0.367%	0.309%	0.465%	0.346%	0.117%	0.296%	0.055%	0.052%	0.310%	0.348%	0.243%
+ SRCNN [12]	36.90%	32.92%	45.12%	40.60%	47.96%	40.30%	19.99%	19.72%	22.33%	24.74%	32.21%
+ SRResnet [25]	36.90%	31.22%	46.86%	45.21%	52.39%	45.52%	21.20%	20.81%	23.26%	23.00%	33.71%
+ LapSRN [24]	41.97%	37.87%	49.07%	42.45%	**56.24%**	49.85%	25.18%	24.51%	27.13%	26.13%	37.44%
+ **PCAN (ours)**	**42.90%**	**39.88%**	**52.44%**	**49.71%**	55.31%	**50.15%**	**27.83%**	**27.11%**	**29.61%**	**31.36%**	**39.59%**

which can handle the pixel misalignment problem during optimization. Simply connecting five PCABs can outperform TSRN by 1.0%. Together with our proposed HOTA attention module, BDC strategy, and Edge Guidance Loss function, our approach can outperform SRResNet [43] by 10.2%, and TSRN [43] by 3.2% on ASTER [38].

4.4 Comparison with State-of-the-Art

In Table 4, we exemplify the experimental results achieved by eleven different SR algorithms on the TextZoom [43] dataset, including SRCNN [12], VDSR [23], SRResNet [25], RRDB [44], EDSR [26], RDN [55], LapSRN [24], RCAN [54], SAN [10], HAN [32], and TSRN [43]. Compared to generating SR images with BICUBIC upsampling only, our designed real text SR algorithm, PCAN, can bring 14.3%, 14.4%, and 20.6% improvement in recognition accuracy to ASTER [38], MORAN [28], and CRNN [37], respectively. Besides, our approach can bring about 6~8% performance gain to the text recognizer compared to the mainstream 10 generic SR algorithms, and the proposed method is still able to bring 3~6% recognition performance gain compared to the text-orient SOTA method TSRN. Finally, we also provide generic SR evaluation metrics SSIM and PSNR, as shown in Table 3. Since PSNR is calculated pixel-to-pixel, our method uses a TPS alignment module that produces a certain pixel offset, so our metrics are not always the best. Usually, PSNR and SSIM could not represent the visual quality of the images [25], in this task, they are also not so important compared to recognition accuracy.

Some qualitative results are shown in Fig. 6, due to the more reasonable sequence context modeling approach and the increased ability to express high-frequency information, we can intuitively see that more intra- and inter-character contextual information can be retained.

4.5 Defense against Adversarial Perturbations

An effective adversarial attack method is proposed by [49] which generates adversarial perturbations against attention-based and CTC-based sequence classification methods, respectively. It almost completely misleads the mainstream text recognition algorithms. Their untargeted attack method aims to minimize the posterior probability between the true predicted characters by applying the inverse gradient [49]. However, our method adequately models the contextual relationships on the images so that it is robust to the adversarial perturbations. To the best of our knowledge, we are the first to study text SR for defense against text adversarial attacks.

对抗性扰动

通过应用逆梯度最小化真实预测字符之间的后验概率

In this section, we demonstrate the effectiveness of our approach to defense against adversarial attacks. As shown in Table 5, all SR models are trained from the TextZoom [43] and demonstrate the

Figure 6: Visualization results of state-of-the-art SR methods on TextZoom dataset, where the recognition results is provided by ASTER [38].

Figure 7: The illustration of the defense against adversarial attacks. 'Rec' is the recognition result of TRBA [1].

effectiveness in defending against adversarial attacks on seven mainstream text recognition datasets. Same as [49], we use the recognition accuracy of TRBA [1] to get quantitative results of our method. Firstly, we obtained the original recognition accuracy on seven mainstream datasets with an average value of 85.24% on gray-scale images with only scaling and normalization image pre-processing steps. Then, we overlap untargeted adversarial perturbations on the input images and the average recognition accuracy drops massively to less than 1%, i.e., the recognizer almost completely failed. However, the reconstruction of adversarial perturbations images by our text SR algorithm can recover the performance of the text recognition algorithm to some extent, with an average recognition accuracy of 39.59%.

In Fig. 7, we visualize the images where the perturbation is more obvious so that can see the impact of the text SR approach on combating the perturbation more intuitively. On the one hand, we can see from the local patch in the red box that the text SR algorithm is able to remove the noise on the attacked image to some extent. On the other hand, from the local image in the blue box, we can see that the perturbation noise is encoded by the SR network into the same domain of the original image, thus being able to break the distribution of the perturbed data.

5 CONCLUSION

In this paper, we verify the effectiveness of contextual information in orthogonal directions for text SR and propose a parallelly contextual modeling approach that can achieve better experimental results. Meanwhile, we focus on the text edge contour information to improve the model's ability to express high-frequency components. Our proposed model significantly improves the performance of text recognition algorithms on real LR images, which can outperform the eleven mainstream SOTA SR methods. Finally, we make the first attempt to defend against adversarial attacks with the text SR algorithm, proving the robustness of our approach. Also, we propose a new idea that text SR can not only improve the accuracy of text recognition on LR images but also enhance data source security. Real text SR methods have greater potential than methods designed for synthetic data. As a new problem to be studied, scene text SR should be given more attention in the future.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62076184, 61673299, 61976160 and 61573255. This work was also supported by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education. This work is also supported by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4715–4723.

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. (2012).

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1563–1570.

[4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3086–3095.

[5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. 2019. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1652–1660.

[6] Rong Chen, Yuan Xie, Xiaotong Luo, Yanyun Qu, and Cuihua Li. 2019. Joint-attention discriminator for accurate super-resolution via adversarial training. In *Proceedings of the 27th ACM International Conference on Multimedia*. 711–719.

[7] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. 2020. Text recognition in the wild: A survey. *arXiv preprint arXiv:2005.03492* (2020).

[8] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*. 5076–5084.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[10] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11065–11074.

[11] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2015), 295–307.

[13] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. 2015. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211* (2015).

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[15] Jingcai Guo, Shiheng Ma, Jie Zhang, Qihua Zhou, and Song Guo. 2020. Dual-view Attention Networks for Single Image Super-Resolution. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2728–2736.

[16] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5197–5206.

[19] Dhruval Jain, Arun D Prabhu, Gopi Ramena, Manoj Goyal, Debi Prasanna Mohanty, Sukumar Moharana, and Naresh Purre. 2020. On-Device Text Image Super Resolution. *arXiv preprint arXiv:2011.10251* (2020).

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[21] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1156–1160.

[22] 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. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 1484–1493.

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1646–1654.

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 624–632.

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

[27] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. 2005. ICDAR 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)* 7, 2-3 (2005), 105–122.

[28] Canjie Luo, Lianwen Jin, and Zenghui Sun. 2019. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition* 90 (2019), 109–118.

[29] Canjie Luo, Qingxiang Lin, Yuliang Liu, Lianwen Jin, and Chunhua Shen. 2021. Separating content from style using adversarial learning for recognizing text in the wild. *International Journal of Computer Vision* (2021), 1–17.

[30] Céline Mancas-Thillou and Majid Mirmehdi. 2007. An introduction to super-resolution text. In *Digital document processing*. Springer, 305–327.

[31] Anand Mishra, Karteek Alahari, and CV Jawahar. 2012. Scene text recognition using higher order language priors. In *BMVC-British Machine Vision Conference*. BMVA.

[32] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. 2020. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*. Springer, 191–207.

[33] Ram Krishna Pandey, K Vignesh, AG Ramakrishnan, et al. 2018. Binary Document image super resolution for improved readability and OCR performance. *arXiv preprint arXiv:1812.02475* (2018).

[34] Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. 2015. ICDAR2015 competition on text image super-resolution. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1201–1205.

[35] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*. 569–576.

[36] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications* 41, 18 (2014), 8027–8048.

[37] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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* 39, 11 (2016), 2298–2304.

[38] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2035–2048.

[39] Bolan Su and Shijian Lu. 2014. Accurate scene text recognition based on recurrent neural network. In *Asian Conference on Computer Vision*. Springer, 35–48.

[40] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 114–125.

[41] Hanh TM Tran and Tien Ho-Phuoc. 2019. Deep laplacian pyramid network for text images super-resolution. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 1–6.

[42] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*. IEEE, 1457–1464.

[43] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. 2020. Scene text image super-resolution in the wild. In *European Conference on Computer Vision*. Springer, 650–666.

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*.

[45] Yizhi Wang and Zhouhui Lian. 2020. Exploring Font-independent Features for Scene Text Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1900–1920.

[46] Yuyang Wang, Feng Su, and Ye Qian. 2019. Text-attentional conditional generative adversarial network for super-resolution of text images. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1024–1029.

[47] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. 2020. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11753–11762.

[48] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2020. Handwritten mathematical expression recognition via paired adversarial learning. *International Journal of Computer Vision* (2020), 1–16.

[49] Xing Xu, Jiefu Chen, Jinhui Xiao, Lianli Gao, Fumin Shen, and Heng Tao Shen. 2020. What machines see is not what they get: Fooling scene text recognition models with adversarial text images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12304–12314.

[50] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. 2008. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[51] Haochen Zhang, Dong Liu, and Zhiwei Xiong. 2017. Cnn-based text image super-resolution tailored for ocr. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.

[52] Lei Zhang and Xiaolin Wu. 2006. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing* 15, 8 (2006), 2226–2238.

[53] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. 2019. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3762–3770.

[54] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*.

[55] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2472–2481.