Shi *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:82
https://doi.org/10.1186/s13634-023-01028-9

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

BCSR: toward arbitrarily oriented text image super-resolution via adaptive Bezier curve network

Mingzhu Shi[1,2*], Muxian Tan[2], Siqi Kong[2] and Bin Zao[2]

*Correspondence:
shimz@tjnu.edu.cn

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

Abstract

Although existing super-resolution networks based on deep learning have obtained good results, it is still challenging to achieve an ideal visual effect for irregular texts, especially spatially deformed ones. In this paper, we propose a robust Bezier Curve-based image super-resolution network (BCSR), which can efficiently handle the degradation caused by deformations. Firstly, the arbitrarily shaped text is adaptively fitted by a parameterized Bezier curve, aiming to convert a curved text box into an annotated text box. Then, we design a BezierAlign layer to calibrate between the extracted features and the input image. By importing the extracted text prior information, the accuracy of the super-resolution network can be significantly improved. It is worth highlighting that we propose a kind of text prior loss that enables the text prior image and the super-resolution text image to achieve cooperation enhancement. Extensive experiments on several standard scene text datasets demonstrate that our proposed model achieves desirable objective evaluation results and further immensely helps downstream tasks related to text recognition, especially in text instances with multi-orientation and curved shapes.

Keywords: Super-resolution, Bezier curve, Text prior, Scene text recognition, Arbitrary-shape text

1 Introduction

Rich and distinctive semantic information in the text serves as a key cue for visual recognition. Despite many promising approaches have been proposed, scene text often encounters various degradations during image processing resulting in low resolution or blurring. This problem dramatically degrades the performance of downstream tasks such as text detection, optical character recognition (OCR), and scene text recognition. Therefore, how to improve the resolution of text areas in natural scenes is what we urgently need to solve now.

Previous super-resolution (SR) [1–5] methods simply learn the degradation patterns of HR-LR pairs, e.g., L1 or L2 loss, to recover the resolution of scene text in images. However, these methods do not treat text as a specific task, so they perform poorly in downstream tasks. Recently, some methods [6–8], tailored for the scene text image

super-resolution (STISR) task, benefit from the superficial properties of the scene text. For example, Wang et al. [9] propose Text Super-Resolution Network (TSRN) to obtain sequence information of text lines through a sequence residual network. In addition, many methods started to try to learn low level-details and high-level semantics to get better features [10]. Zhao et al. [11] propose a parallel contextual attention network (PCAN) that can efficiently learn sequence-related features and pay more attention to high-frequency information for text image reconstruction. TPGSR [12] introduces possible text sequences based on TSRN and improves the quality of the generated text by continuous iteration. These CNN-based methods mentioned above have difficulties in dealing with spatial deformation, especially with severely curved or rotated text images, because it is a local operation. Ma et al. [13] propose TATT, which uses the global attention mechanism to align the text prior with the text image of spatial deformation and plays the role of semantic guidance of text in the process of text reconstruction.

By observation, semantic information in the text can help recover the shape and details of characters. Existing methods generally employ pre-recognizers such as CRNN [14] or MORAN [15] to extract text priors, which have achieved satisfactory results in recovering well-aligned text, but are still challenging when dealing with spatially deformed text. Figure 1 shows an example where the recent approach TSRN produces wrong images when processing curved text and the characters are incorrectly recognized by the recognizer. TPGSR is inferior to our model in terms of quantitative analysis. This is because TSRN ignores the semantic information that comes with the text, whereas TPGSR uses CRNN as a pre-trained recognizer to obtain text priors. CRNNs empirically tend to produce erroneous recognition results when recognizing severely distorted and curved text, which greatly affects the super-resolution reconstruction task. Thus, our motivation is to design a novel and effective pre-trained text recognition module for recognizing arbitrarily oriented and curved text in images, which ensures the generation of accurate text prior information to help guide low-resolution text reconstruction.

To achieve this goal, we propose a robust Bezier Curve-based image super-resolution network (BCSR) to reconstruct spatially deformed text images. We incorporate a text prior generation module (TPG) that introduces the Bezier curve into the super-resolution network. Unlike previous approaches that use rectangular bounding boxes to detect text positions, the Bezier curve can adaptively fit curved or oriented text and remove the influence of the surrounding irrelevant background. In our proposed module, we adopt the parametric Bezier curve to adaptively fit the spatially deformed text and then realize the feature sharing of the text instance through BezierAlign with negligible computation. Since the text prior information containing rich semantic information is introduced into the super-resolution (SR) network, the recognition accuracy of the reconstructed text

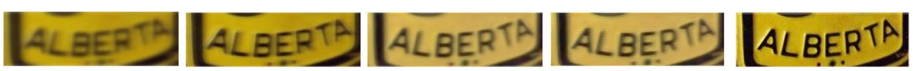


Fig. 1 The recognition results of super-resolution images reconstructed by different models. Below the image is the text recognition results, where red fonts mean different from HR image recognition results. 'P' and 'S' stand for PSNR and SSIM results

with satisfactory visual effects is also improved. To further beautify the text's appearance suffered from deformation degradation, we propose a TP loss that effectively measures the similarity between the TP features generated by TPG and the HR ground-truth images. Our work has made the following noteworthy contributions:

1) We design a toward arbitrarily oriented text image super-resolution network that employs a novel and effective TPG module feature-level to resolve LR images and serves as a prior to enhancing text recognition accuracy.
2) We innovatively use the Bezier curve to convert the curved text box into an annotated box. Simultaneously, the detection module and the recognition module are connected by a lightweight recognition head, greatly simplifying the difficulty of text recognition.
3) By adding the correct semantic information to the SR network, our model significantly boosts the recognition performance of irregular text in the TextZoom dataset and demonstrates its good generalization ability on other datasets.

2 Related works

2.1 Single image super-resolution

Single Image Super-Resolution (SISR) attempts to reconstruct LR images with missing features and associated partial distortions into ideal HR images. Previous approaches employ artificially handcrafted image priors, including statistical prior, self-similarity prior, and sparsity prior. Dong et al. [3] pioneered the application of deep convolutional neural networks in image super-resolution. Later on, the network structure based on CNN has been proposed successively, such as VDSR [16], DRCN [17], ESPCN [18], and FSRCNN [19]. These methods typically use L norms as objective functions but ignore human perception. In [7] embedded Generative Adversarial Networks (GANs) in the SR task to minimize the distance perception correlation distribution between LR and HR images to address this issue. Recently, SFT-GAN [20] and FSRNet [2] introduced semantic segmentation information based on the GAN network to generate visually more satisfying HR images. Although existing GAN-based SISR methods can generate realistic texture information, these tend to generate meaningless and irrelevant noise to the input image.

2.2 Scene text image super-resolution

Unlike SISR, Scene Text Image Super-Resolution (STISR) focuses on scene text images to improve text readability and produce the best quality images. There is no doubt that the STISR method can directly adopt the SISR method. Dong team [21] expanded SRCNN [3] to text images and received one of the top rankings in the 2015 ICDAR competition [22]. Wang et al. [23] proposed TextSR, using adversarial loss and text perception loss together as a loss function for super-resolution generation. Thus, it can more effectively focus image reconstruction on text regions rather than irrelevant background regions. To improve the performance of STISR in real-world scenarios, Wang et al. [9] constructed a real-world STISR image dataset from the real SISR dataset [24, 25], namely TextZoom. Nakao et al. [26] proposed SRRNN handle character and text-related

problems and employed two super-resolution training methods on text images, achieving continuous improvement in the accuracy of scene images containing text. Furthermore, in [27] and [28] proposed to enhance the network block structure by self-participating image features and participating channels to improve STISR performance.

2.3 Scene text recognition

Scene text recognition (STR) technology has a wide range of application scenarios, such as visual search [27], license plate recognition [9], and other image understanding tasks based on scene text [23, 29, 30]. Some early methods mostly follow the bottom-up principle, i.e., extracting low-level features to detect and recognize individual characters and then integrating these characters into words according to heuristic rules or local language models [25]. However, these methods are only suitable for weak representations of semantic information. Other methods follow the top-down principle [9], which maps the scene text image into a feature sequence and then performs word-level classification, typically including CRNN [14], RARE [31], etc. Although RARE can adapt to the recognition of curved text images, it requires expensive character-level or pixel-level annotations. In [32], Li et al. designed a model based on a 2D attention module, which can locate and identify irregular characters one by one without additional supervision information. Recently, Liu et al. [33] proposed an Adaptive Bezier Curve Network (ABCNet), which uses parameterized Bezier curves to reconstruct scene text images of arbitrary shapes, significantly saving computational time and resources. Although deep learning-based algorithms have attained the remarkable performance, recognizing arbitrarily oriented text in low-quality images remains hard. Inspired by ABCNet, we try to use cubic Bezier curves and BezierAlign as text prior generation modules to generate categorical text priors for STISR model training. The results show that introducing the text prior generation module into the STISR model can greatly improve the perceptual quality of the generated HR images, thereby enhancing the text recognition performance.

3 Methodology

3.1 Overall architecture

The whole structure of BCSR is depicted in Fig. 2, which consists of two major components: (1) The text prior generation module (TPG) for extracting the TP features of LR images; (2) The basal super-resolution (SR) network for reproducing HR text images from input LR images and TP guidance features.

Denote $X \in R^{h\times w\times 3}$ as the LR input image, $f_I \in R^{h\times w\times c}$ and f_{TP} as the image feature and estimated prior information by the TPG. First, we pass a convolutional layer whose kernel size is 9×9 to extract the image feature. f_I is defined in Eq. (1).

$$f_I = \mathrm{Conv}(X) \tag{1}$$

On the other hand, the LR input image is passed to bicubic interpolation with the aim of aligning f_{TP} and f_I to facilitate subsequent computation of their correlation. Since the text prior information can help reconstruct the HR image, the image after bicubic interpolation is transmitted to TPG to obtain the sequence of text prior features. The definition of f_{TP} is shown in Eq. (2).

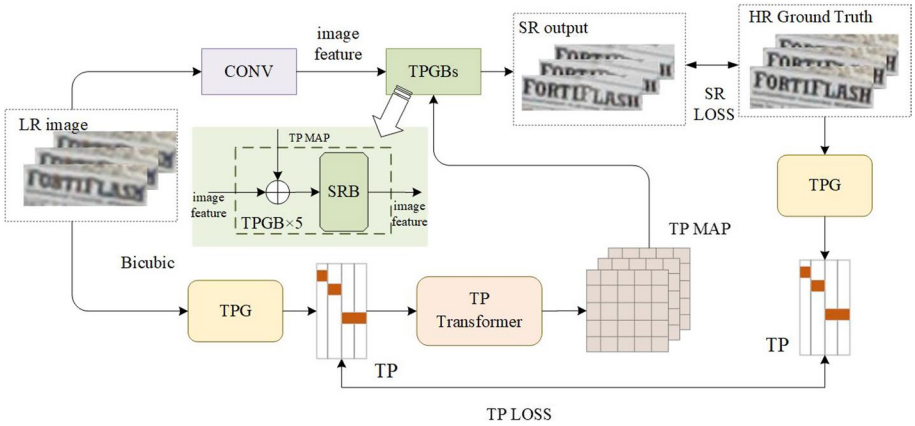


Fig. 2 The overall architecture of our proposed BCSR. TPGB represent text prior guided blocks, $\bigoplus$ denotes the element-wise addition. Accordingly, TP loss and SR loss are employed to train the whole network

$$f_{\mathrm{TP}} = \mathrm{TPG}(\mathrm{Bicubic}(X)) \tag{2}$$

Where $f_{\mathrm{TP}} \in R^{l\times|A|}$ is an l-length sequence with a classification probability vector of size $|A|$. A represents a total of 37 character including the numbers 0 through 9, the letters a through z, and a blank class. The semantic information in the generated f_{TP} will be assigned to the appropriate position in the spatial domain to help recover HR text images.

Then, the TP feature are passed to the TP Transformer. The TP Transformer contains four Deconvblocks, where every Deconvblock is composed of a deconvolution layer, a Batch Normalization (BN) layer and a rectified linear unit (ReLU) layer. The exportation of the TP Transformer will be a characteristic mapping with recovery space dimension $f_{\mathrm{TM}} \in R^{h\times w\times c}$. The TP map f_{TM} we obtain is a modulation mapping, and the semantics-specific part of the image feature can be improved by it.

Ultimately, the TP map f_{TM} and the image feature f_I are transmitted into TPGBs module, which consists of five text prior-guided blocks (TPGBs) where use element-wise addition to merge f_{TM} and f_I, and a Sequential-Recurrent Block (SRB) for reconstructing the HR image. Just like the previous super-resolution models, the output of the reconstruction module will evaluate the estimated HR images by some of these TPGBs.

3.2 The text prior generation module (TPG)

Researchers have become increasingly interested in mitigating the influence of imprecise prior information and combining useful prior information. Following [33], the text prior generation module (TPG) was proposed to assist the SR process in generating satisfactory high-quality images. The structure of TPG is shown in Fig. 3. Specifically, we adopt ResNet 50 [34] with FPN [35] as the backbone framework and highlight curved text regions by using the cubic Bezier curve and BezierAlign. Next, we present the three crucial sections of the TPG: (1) Bezier curve detection; (2) BezierAlign; (3) Lightweight Recognition Head.

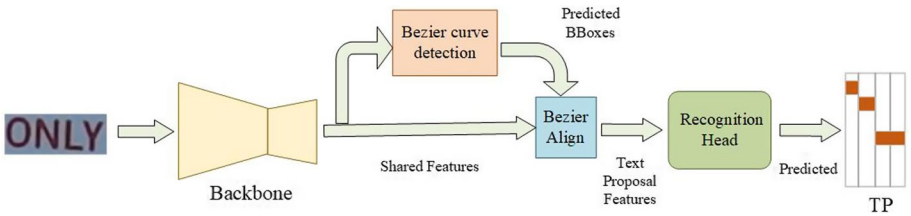


Fig. 3 The structure of the text prior generation module (TPG). First, we obtain the input image features through the backbone, then use the Bezier curve to annotate the text area and finally transmit the text features after BezierAlign to the recognition module

3.2.1 Bezier curve detection

There are two commonly used methods for text detection of arbitrary shapes: segmentation-based methods [36, 37] and regression-based methods [33, 38, 39]. However, existing regression-based methods [33, 37] require more parameterized predictions to fit text boundaries, so they are unsuitable for real-time detection. To solve this problem, Bezier curves are introduced.

The Bezier curve $B(t)$ is derived from Bernstein polynomials, which can form various shapes of curves by selecting control points. Given control points $b_0, b_1, \cdots, b_i$, the definition of the $n-th$ degree Bezier curve is illustrated in Eq. (3):

$$B(t) = \sum_{i=0}^{n} \binom{n}{i} b_i (1-t)^{n-i} t^i, i = 0, 1, 2 \ldots n \tag{3}$$

where $\binom{n}{i}$ is a binomial coefficient.

Inspired by [33], n is set to 3, that is, the cubic Bezier curves are sufficient for different types of arbitrary shape scene text images in practice. Using the cubic Bezier curve, text detection for arbitrarily shaped scenes can be simplified by a bounding box regression with a total of eight control points, the coordinates of which are the targets of the detection network predictions. We use a standard least squares fit for an arbitrarily shaped dataset with polygon annotations to calculate the best control points.

Assuming that the $i-th$ annotation point is p_i, annotation points on the boundary of the text image can be represented by $\{p_i\}_n^{i=1}$. All we have to do is find the optimal parameters of the Bezier curve in Eq. (3) using the standard square method, as described in Eq. (4).

$$\begin{bmatrix} B_{0,3}(t_0) & \cdots & B_{3,3}(t_0) \\ B_{0,3}(t_1) & \cdots & B_{3,3}(t_1) \\ \vdots & \ddots & \vdots \\ B_{0,3}(t_m) & \cdots & B_{3,3}(t_m) \end{bmatrix} \begin{bmatrix} b_{x_0} & b_{y_0} \\ b_{x_1} & b_{y_1} \\ b_{x_2} & b_{y_2} \\ b_{x_3} & b_{y_3} \end{bmatrix} = \begin{bmatrix} p_{x_0} & p_{y_0} \\ p_{x_1} & p_{y_1} \\ \vdots & \vdots \\ p_{x_m} & p_{y_m} \end{bmatrix} \tag{4}$$

where t is the ratio of the polyline segment to the perimeter of the entire curve, m stands for the number of points that need to be annotated on an edge. The original multi-segment line annotation is converted into a parameterized Bezier curve through Eqs. (3) and (4). Considering that only four control points are required for the straight text, we added two additional control points at the thirds of each long edge for consistency.

Following that proposal bounding boxes are generated, and the output offset of the model is shown in Eq. (5).

$$\Delta x = b_{ix} - x_{\min}, \Delta y = b_{iy} - y_{\min} \tag{5}$$

Here $x_{\min}$ and $y_{\min}$ correspond to the minimum of the fixed points, which decouples the relationship between the predicted fixed control point b_i and the image boundary and beyond. Here is a detection head that can densely predict the detection results by outputting feature maps. The head is composed of four stacked convolutional layers, where each convolutional layer stride is set to 1, and the padding is also 1, and 3×3 kernels. We only need to learn Δx and Δy through a convolutional layer with an output of 16 channels to get the correct prediction results. The visual comparison in Fig. 4 shows that the image corrected by the Bezier curve has a more satisfactory visual effect.

3.2.2 BezierAlign

The existing sampling methods, horizontal and quadratic sampling methods, sample the background information for curved text, leading to aligned features containing irrelevant background images that interfere with recognition. To address this issue, BezierAlign [33] is developed, where the columns in the BezierAlign are orthogonal to the boundaries of the Bezier curve for text areas.

Assuming that the text region features of the input image are given, the output feature map consists of pixel units of $h_{\text{out}} \times w_{\text{out}}$ rectangle size. Taking the pixel g_i at the position (g_{iw}, g_{ih}), t is calculated by the formula (6):

$$t = \frac{g_{iw}}{w_{\text{out}}} \tag{6}$$

Then, the upper boundary point t_p of the Bezier curve and b_p of the lower Bezier curve can be calculated by t and Eq. (3). With t_p and b_p, BezierAlign can apply bilinear

(a) Original polyline annotation (b) Generated Bezier annotation

Fig. 4 Visual comparison of rectified images under different annotations. **a** employs polyline annotations, using an STN-based method to rectify the original ground truth to an approximate rectangular text. **b** uses Bezier annotation, the red points are control points, and the red dashed lines form a control polygon for each curve boundary. We utilize the Bezier curve and BezierAlign to revise the results. The picture below is the outcome of warping

interpolation to obtain the sampling point o_p, and the calculation formula is shown in Eq. (7).

$$o_p = b_p \cdot \frac{g_{ih}}{h_{\text{out}}} + t_p\left(1 - \frac{g_{ih}}{h_{\text{out}}}\right) \tag{7}$$

Each pixel unit gets four o_p coordinate values, passed through the max pooling layer to get the output feature map. This method allows for accurately fitting text areas without introducing large amounts of invalid background information.

3.2.3 Lightweight recognition head

To better utilize the feature sharing in training, we propose a lightweight recognition head, equivalent to a simplified version of CRNN [14]. It consists of six convolutional layers with a padding size limit of 1, a bidirectional LSTM [40] layer for predicting feature sequences, and a fully connected layer with an output total class of 97. Based on the predicted probability of the sample points, the CTC loss [1] is adopted to align the text strings (GT) so that we can directly obtain the features of the region of interest through the GT of the generated Bezier curve.

3.3 Loss functions

During training, our BCSR model contains a super-resolution loss L_{SR} and a text prior loss L_{TP}.

SR loss function We make the following settings: $\hat{I_H}$ represents the estimated the HR image of the input LR image, I_H represents the actual value of HR image, and L_{SR} represents the loss function of SR, which is generally the L norm distance between $\hat{I_H}$ and I_H. The mathematical formula is shown in Eq. (8).

$$L_{\text{SR}} = \left|\hat{I}_H - I_H\right| \tag{8}$$

Text prior loss function In the network design, TP sequences generated by the TPG play a significant role in the final result of SR. What is more, TP sequences similar to authentic HR images are what we need, so we use TP sequences extracted from real HR images to supervise network learning. Text prior loss consists of L_1 loss and KL divergence loss. The TP extracted from LR images I_L and the ground truth images I_H, respectively, are denoted by t_L and t_H. With the text prior $t_H, t_L \in R^{L\times|A|}$ of the pair of LR and HR images, the $D_{KL}(t_L \| t_H)$ can be calculated as follows:

$$D_{KL}(t_L \| t_H) = \sum_{i=1}^{L}\sum_{j=1}^{|A|} t_H^{ij} \ln \frac{t_H^{ij} + \varepsilon}{t_L^{ij} + \varepsilon} \tag{9}$$

Where t_H^{ij} and t_L^{ij} represent the elements in the ith position and the jth dimension in t_H and t_L. ε is a small real number that avoids numeric errors in division and logarithms.

Combined with the SR loss function, the overall loss function of the network is defined as Eq. (10):

$$L = L_{\text{SR}} + \alpha|t_H - t_L| + \beta D_{KL}(t_L \| t_H) \tag{10}$$

where α and β are the balancing parameters.

4 Experiments

4.1 Implementation details

In this experiment, we adopted TSRN as the SR module. Inspired by FasterRCNN [41], the backbone of the proposed TPG module utilize ResNet 50 [34] and a Feature Pyramid Network (FPN) [35] for extracting image features. The detection head uses ROI alignment on feature maps with input resolutions of $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$, $\frac{1}{64}$, $\frac{1}{128}$, while the recognition branch calculates feature maps with the size of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ by BezierAlign. The pretrained model is executed on the Coco-text dataset [42]. We also performed data augmentation to make the model more robust to minor changes, e.g., rotate 90Â°, 180Â°, 270Â°; scale images according to the ratio of 0.5 and 0.8; and randomly crop (make sure that the size of the cropped image is not less than half of the original).

Our proposed model is trained on Telsa P100 GPUs with image batch size of 32. The optimizer takes Adam with a momentum of 0.9. Training lasts 500 epochs with a learning rate of 10^{-3}. m in Eq. (4) is set to 3 in TextZoom and 5 for other datasets. In Eq. (10), the weights α and β are both set to 0.9, while ε in Eq. (9) is set to 10^{-6}.

4.2 Dataset

TextZoom TextZoom dataset [9] contains 21,740 paired LR-HR images in real scenes, among which 17,367 samples are used for training. The remaining samples can be divided into three categories for testing, namely easy (1619 samples), medium (1411 samples), and difficult (1343 samples).

Total-text Total-text dataset [43] collects 1555 images from various scenes, which contain texts containing more than three categories: horizontal text, oblique text, and curved text. Each image in the dataset has its own annotated text.

ICDAR2015 ICDAR2015 [44] has 2077 scene text images for testing. Most of them are low-resolution images and perspective distorted images, making them extremely challenging.

CTW1500 CTW1500 [39] contains 1500 images, mainly taken from Google Open-Image and camera shots, with a large amount of horizontal and multi-directional text. There are various forms of text in the dataset, such as blur, perspective, distortion, and noise. In addition, this dataset is multilingual, mainly in Chinese and English.

4.3 Ablation studies

To better illustrate our model, we investigate the impact of TPG in SR reconstruction, the effect of tuning TPG and the effectiveness of the SR module in our BCSR framework. All evaluations in this section are performed on the dataset TextZoom.

Impact of TPG in SR reconstruction Since TPG aims to improve text recognition performance by generating probable text sequences, we compare it with other strategies, e.g., using CRNN as models of character recognition module. According to Table 1, the SR performance of pre-recognition classifiers like MORAN and ASTER [45], which can efficiently handle curved or rotated text, is better than that of CRNN. Our proposed TPG obtains the highest PSRN/SSIM (21.67/0.7991), which also implies the best performance. The experiments and comparisons demonstrate that our text prior generation

Table 1 Ablation study for TPG

Backbone	Strategies for extracting prior	PSNR/dB	SSIM
TSRN [9]	With CRNN	20.37	0.7719
TSRN [9]	With MORAN	20.92	0.7823
TSRN [9]	With ASTER	21.20	0.7916
TSRN [9]	With TPG (ours)	21.67	0.7991
TSRN [9]	Without	19.70	0.7157

Table 2 Tuning the TPG

Approach	Tuned	ACC (%)
TSRN [9]	–	41.4
Ours	×	46.5
Ours		54.9
HR	–	71.2

'Tuned' means whether the TPG is fine-tuned or not

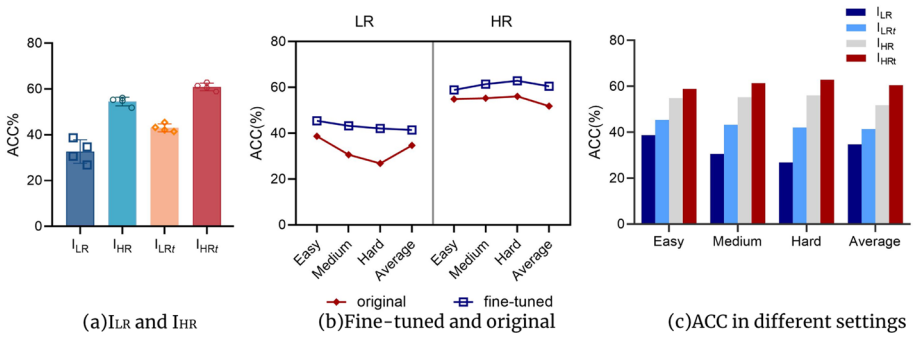
module improves the resolution of text images and reconstructs semantically correct texts.

Impact of tuning the TPG To demonstrate the significance of tuning the TPG, we evaluate the average correct recognition rate (ACC) by tuning the model's TPG. The text recognition accuracy is shown in Table 2. Compared with TSRN, the BCSR model without fine-tuning improves the recognition ability of SR images by 5.1%. The tuned model can increase the text recognition accuracy rate from 46.5 to 54.9%, achieving an 8.4% increase. It evidently shows that tuning the TPG can effectively improve the text recognition accuracy. We observe that the computation of our proposed method is more complex than that of TSRN due to the usage of ResNet50 and FPN in the TPG, which increases the complexity of convolution computation. This network makes certain computational efficiency sacrifices even if it produces better reconstruction outcomes.

The effectiveness of SR module in BCSR It is necessary to determine whether the expected generated HR images are helpful for the final text recognition tasks because one of the objectives of the SR module is to increase text recognition performance by HR image recovery. Thus, we use LR and HR images as inputs to evaluate the BCSR model with fixed and tuned TPG. From (a) of Fig. 5, we can see that the average ACC results with HR as input are higher than those with LR as input among all cases, and (b) emphasizes that adjusting TPG is beneficial for text recognition performance, and (c) illustrates that the more complex images become, the more noticeable they improve text recognition accuracy by adjusting the TPG on the LR images. Such results indicate that our SR module can effectively boost the final SR text recognition performance.

4.4 Comparison with state of the arts

In this section, we compare the proposed BCSR network with other excellent approaches, including SRCNN [3], SRResNet [7], HAN [46], TSRN [9], TBSRN [12], PCAN [47], TPGSR [12] and TPGSR-3 [12]. For a fair comparison, all models are trained



(a)I_{LR} and I_{HR} (b)Fine-tuned and original (c)ACC in different settings

Fig. 5 Ablation on the impact of SR module. I_{LR} and I_{HR} represent using the LR image and estimated HR images as recognizer input, respectively, while t in the subscript means the recognition with tuned TPG

on the same dataset with the same settings. The reconstructed high-resolution images are all used for text recognition by CRNN.

Results on TextZoom The objective evaluation index values of each model are shown in Table 3. From the two objective metrics of PSNR and SSIM, our model achieves relatively good results on all three types of data from TextZoom and even achieves the best average, proving our model's superiority in enhancing visual quality.

To further investigate the generalization ability of our model to irregular text images, we artificially selected 804 samples of rotated and curved shapes from TextZoom as inputs to compare models. As reflected in Table 4, our BCSR model obtains the highest

Table 3 Evaluation of competitive SISR and STISR models on TextZoom datasets

Approach	PSNR/dB				SSIM			
	Simple	Medium	Hard	Average	Simple	Medium	Hard	Average
Bicubic	22.35	18.98	19.39	20.35	0.7884	0.6254	0.6592	0.6961
SRCNN [3]	23.48	19.06	19.34	20.78	0.8379	0.6323	0.6791	0.7227
SRResnet [7]	24.36	18.88	19.29	21.03	0.8681	0.6406	0.6911	0.7403
HAN [46]	23.30	19.02	20.16	20.95	0.8691	0.6537	0.7387	0.7596
TSRN [9]	25.07	18.86	19.71	21.42	0.8897	0.6676	0.7302	0.7690
TBSRN [12]	23.46	**19.17**	19.68	20.91	0.8729	0.6455	0.7452	0.7603
PCAN [47]	24.57	19.14	20.26	21.49	0.8830	0.6781	0.7475	0.7752
TPGSR [12]	23.73	18.68	20.06	20.97	0.8805	0.6738	0.7440	0.7719
TPGSR-3 [12]	24.35	18.73	19.93	21.18	0.8860	**0.6784**	0.7487	0.7774
Ours	**25.35**	19.02	**20.93**	**21.67**	**0.8901**	0.6763	**0.7507**	**0.7991**

Bold represents the highest value

Table 4 PSNR and SSIM of competing SISR and STISR models for irregular samples on the TextZoom dataset

Method	PSNR/dB	SSIM	ACC (%)
Bicubic	19.68	0.6658	18.1
TSRN [9]	19.10	0.7066	26.6
TBSRN [12]	19.70	0.7157	38.3
TPGSR [12]	19.79	0.7293	42.5
Ours	20.31	0.7875	43.8

Fig. 6 Visual comparison of different STISR models on TextZoom. The text recognition result is at the top of each image, black for correct characters and red for missed or incorrect. Zoom in for better visualization

Table 5 Cost versus performance

Approach	Flops	Parameters	Eatency/s	ACC (%)
TSRN [9]	0.91G	885.83 MMac	3.71	41.4
TPGSR [12]	1.61G	897.25 MMac	4.35	49.8
Ours	1.72G	907.33 MMac	4.82	54.9

ACC means the average recognition accuracy

PSNR (20.31), SSIM (0.7875) and ACC (43.8%). It is evident that when encountering arbitrarily oriented and curved text images, our model will open a large gap with other models such as TBSRN and TPGSR.

We also provide a visual comparison result for regular samples as well as spatially deformed samples as shown in Fig. 6. The deep learning-based method has been significantly improved in terms of visual effects compared with the bicubic interpolation algorithm. STISR models such as TSRN and TPGSR are still unstable in recovering spatially deformed images. Notably, our proposed model performs best in recovering text semantics in all case samples, and the generated image visual effects are also closest to high-resolution images. This fully proves that introducing Bezier curves to the text recognition module is practical and beneficial in upgrading the performance and robustness of the model.

To further explore the value of our BCSR, we also compare the computational consumption with TSRN and TPGSR. In Table 5, the experimental results show that our model increases the complexity of the convolutional computation due to the use of ResNet 50 [34] with FPN [35] in the TPG module, but its reconstructed image performs 5.1% better when compared to TPGSR. It is humble to conclude that introducing the

Table 6 Evaluation of competitive STISR models on other datasets

	Method				
	Bicubic	**TSRN [9]**	**TBSRN [12]**	**TPGSR [12]**	**Ours**
Total-text [43]					
PSNR/dB	19.22	20.9	21.56	21.84	**22.25**
SSIM	0.5542	0.6664	0.6879	**0.7264**	0.7157
ICDAR2015 [44]					
PSNR/dB	21.75	22.86	23.12	23.94	**24.12**
SSIM	0.6712	0.7222	0.7654	0.7872	**0.7912**
CTW1500 [39]					
PSNR/dB	17.52	18.39	19.01	19.38	**21.67**
SSIM	0.6309	0.6831	0.7111	0.7635	**0.7663**

Bold represents the highest value

Fig. 7 Comparison of visual effects of different models in Total-text, where the right is the detail image magnified by ×2

TPG module to generate textual priors in our model is more valued than offset by the additional consumption it brings.

Generalization to other datasets We evaluate the robustness of our BCSR network to other datasets, including Total-text [43], ICDAR2015 [44] and CTW1500 [39]. These datasets contain a large amount of arbitrarily oriented and curved texts. As shown in Table 6, our model achieves better results on other datasets, which demonstrates that our model has good generalization capabilities despite being trained on TextZoom, and that the high-quality images generated by our model can boost the performance of downstream tasks.

Visual comparison of competing STISR models on Total-text is shown in Fig. 7, where two examples are exhibited: curved (top) and rotated (bottom). It can be observed that although the bicubic interpolation algorithm can roughly restore the original image, it cannot reconstruct more details and give a vague feeling. TPGSR can reconstruct a sharper image, but the output boundary is blurred. Obviously, our proposed model can

Table 7 Text recognition accuracy on the LR scene images in other datasets

Datasets	ICDAR2015 [44]	Total-text [43]	CTW1500 [39]
No. of images	241	20	159
Original	21.5%	18.1%	19.2%
TSRN [9]	24.5%	20.2%	23.1%
TPGSR [12]	27.1%	22.4%	29.8%
Ours	31.2%	25.3%	35.9%
Improve	+ 9.7%	+ 7.2%	+ 16.7%

Fig. 8 Comparison of recognition results of different models. The first two rows are taken from the ICDAR2015 dataset, and the last two rows are from the CTW1500 dataset. Zoom in for better visualization

generate sharper image edges and more satisfying visual effects, which indicates that our BCSR is more suitable for solving text images with arbitrary shapes.

To better understand the generalization ability of our model, we picked low-quality text images (i.e., recognition score≤0.6) from the testing set with 420 samples (241 from ICDAR2015, 20 from Total-text and 159 from CTW1500). In this test, we adopt the recognition accuracy of the SR results as the evaluation standard. As can be seen from Table 7, it is evident that although TSRN [9] and TPGSR [12] can improve the recognition accuracy of the original image, BCSR shows excellent performance in all types of datasets. In particular, our model grew by 16.7% compared to the original images in the CTW1500 dataset.

Most of the text images in the ICDAR2015 dataset are arbitrarily oriented text images generated from arbitrary shooting angles and perspective distortion. From Fig. 8, we can see that our model's generated image recognition results are similar to those of the original images. This is because the BCSR text annotation box is closer

to the shape of the text and therefore produces a more pronounced text edge. On the other hand, our model achieves more accurate recognition results on both curved and horizontal text in the CTW1500 dataset. These demonstrate that our model can generate text images with higher resolution to help the text recognition task.

4.5 Limitations

Some examples of failures are shown in Fig. 9, where we observe that our method has weaker performance when dealing with long and dense text, the former because its native character-to-character connections are then blurred, and in the latter case, there is crossover when text regression box annotation is performed, hence the problem of text sticking, which leads to degradation of the quality of the generated images. Additionally, the uneven illumination and more severe perspective also made the experiment difficult. We will mitigate these problems in our future work.

5 Future work

In the future, we will continue perfecting BCSR in followings directions:

Continuous improvement in the structure itself In this paper, we demonstrate the results of only one stage. There are many potential improvements, such as using progressive super-resolution reconstruction.

More lightweight According to tabel1, the model is still somewhat computationally intensive. Various lightweight techniques can be adopted into BCSR, such as network cropping, lightweight regularization, or lightweight activation functions.

More application scenarios The dataset we have used is primarily English characters, and it is worth exploring whether BCSR performs better in other texts, such as German and Chinese.

6 Conclusion

In this paper, we presented a super-resolution network model named BCSR for recognizing arbitrarily oriented text images in natural scenes. Text images differ from other natural scene images and have their unique text classification information. To better utilize the prior information of the text, a text prior generation module is developed, by introducing a parameterized Bezier curve to reformulate arbitrarily shaped scene text images. Subsequently, we add TP features generated by TPG based on reconstructing text images using image features, which can generate more precise details for text recognition. In addition, we proposed a TP loss to realize the collaborative enhancement

Fig. 9 Some examples of failures. **a** shows the visualization and text recognition results of our method on extremely compressed and severely distorted text samples, while (**b**) represents the results of Bezier curve generation. Zoom in for better visualization

of the text prior image and the restored super-resolution text image. With such a model, we can not only solve the low-resolution text problem but also significantly strengthen the readability of text, especially for those text images with arbitrary shapes such as distorted and stretched.

Abbreviations

SR	Super-resolution
LR	Low resolution
HR	High resolution
TP	Text prior

Acknowledgements
The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Author Contributions
All authors are involved in deriving the algorithm and making the validation experiments. All authors read and approved the final manuscript.

Funding
This work was supported by the National Science Foundation of China under grant numbers 61501328; Enterprise Joint Horizontal Science and Technology Project under grant numbers 53H21034.

Availability of data and materials
The public data set can be downloaded from the official website.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 27 October 2022 Accepted: 30 May 2023
Published online: 18 July 2023

References

1. Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, S. Zhou, Focusing attention: towards accurate text recognition in natural images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5076–5084 (2017)
2. Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, Fsrnet: end-to-end learning face super-resolution with facial priors. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2492–2501 (2018)
3. C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
4. W. Wang, J. Hu, X. Liu, J. Zhao, J. Chen, Single image super resolution based on multi-scale structure and non-local smoothing. EURASIP J. Image Video Process. **2021**(1), 1–15 (2021)
5. J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, Z. Xu, Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution, in *Computer Vision - ECCV 2020*. ed. by A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Springer, Cham, 2020), pp.208–224
6. W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep Laplacian pyramid networks for fast and accurate super-resolution. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 624–632 (2017)
7. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, Photo-realistic single image super-resolution using a generative adversarial network. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690 (2017)
8. B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144 (2017)
9. W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, C. Shen, X. Bai, Scene text image super-resolution in the wild. in *European Conference on Computer Vision*, pp. 650–666 (2020). Springer
10. X. Wu, D. Hong, J. Chanussot, Uiu-net: U-net in u-net for infrared small object detection. IEEE Trans. Image Process. **32**, 364–376 (2023). https://doi.org/10.1109/TIP.2022.3228497
11. C. Zhao, S. Feng, B.N. Zhao, Z. Ding, J. Wu, F. Shen, H.T. Shen, Scene text image super-resolution via parallelly contextual attention network. in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2908–2917 (2021)
12. J. Ma, S. Guo, L. Zhang, Text prior guided scene text image super-resolution. arXiv preprint arXiv:2106.15368 (2021)
13. J. Ma, Z. Liang, L. Zhang, A text attention network for spatial deformation robust scene text image super-resolution. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911–5920 (2022)
14. B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)

15. C. Luo, L. Jin, Z. Sun, MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recogn. **90**, 109–118 (2019). https://doi.org/10.1016/j.patcog.2019.01.020
16. J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637–1645 (2016)
17. J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654 (2016)
18. W. Shi, J. Caballero, F.Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883 (2016)
19. C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network. in *European Conference on Computer Vision*, pp. 391–407 (2016). Springer
20. X. Wang, K. Yu, C. Dong, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 606–615 (2018)
21. C. Dong, X. Zhu, Y. Deng, C.C. Loy, Y. Qiao, Boosting optical character recognition: A super-resolution approach. arXiv preprint arXiv:1506.02211 (2015)
22. C. Peyrard, M. Baccouche, F. Mamalet, C. Garcia, Icdar2015 competition on text image super-resolution. in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1201–1205 (2015). IEEE
23. W. Wang, E. Xie, P. Sun, W. Wang, L. Tian, C. Shen, P. Luo, Textsr: Content-aware text super-resolution guided by recognition. arXiv preprint arXiv:1909.07113 (2019)
24. X. Zhang, Q. Chen, R. Ng, V. Koltun, Zoom to learn, learn to zoom. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770 (2019)
25. J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, 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*, pp. 3086–3095 (2019)
26. R. Nakao, B.K. Iwana, S. Uchida, Selective super-resolution for scene text images. in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 401–406 (2019). IEEE
27. J. Chen, B. Li, X. Xue, Scene text telescope: Text-focused scene image super-resolution. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035 (2021)
28. C. Zhao, S. Feng, B.N. Zhao, Z. Ding, J. Wu, F. Shen, H.T. Shen, Scene text image super-resolution via parallelly contextual attention network. in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2908–2917 (2021)
29. F.Z. Ait Bella, M. El Rhabi, A. Hakim, A. Laghrib, An innovative document image binarization approach driven by the non-local p-laplacian. EURASIP J. Adv. Signal Process. **2022**(1), 1–18 (2022)
30. M. de Leeuw, G. den Bouter, T. Ippolito, R. O'Reilly, M. van Remis, A. Webb, Gijzen, Deep learning-based single image super-resolution for low-field mr brain images. Sci. Rep. **12**(1), 1–10 (2022)
31. X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, J. Yan, Fots: Fast oriented text spotting with a unified network. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5676–5685 (2018)
32. H. Li, P. Wang, C. Shen, G. Zhang, Show, attend and read: A simple and strong baseline for irregular text recognition. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8610–8617 (2019)
33. Y. Liu, H. Chen, C. Shen, T. He, L. Jin, L. Wang, Abcnet: Real-time scene text spotting with adaptive bezier-curve network. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 9809–9818 (2020)
34. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
35. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125 (2017)
36. W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9336–9345 (2019)
37. Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, X. Bai, Textfield: Learning a deep direction field for irregular scene text detection. IEEE Trans. Image Process. **28**(11), 5566–5579 (2019)
38. X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, S. Kim, Arbitrary shape scene text detection with adaptive text region representation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6449–6458 (2019)
39. Y. Liu, L. Jin, S. Zhang, C. Luo, S. Zhang, Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recogn. **90**, 337–345 (2019)
40. P. He, W. Huang, Y. Qiao, C.C. Loy, X. Tang, Reading scene text in deep convolutional sequences. in *Thirtieth AAAI Conference on Artificial Intelligence* (2016)
41. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
42. A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie, Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
43. C.K. Ch'ng, C.S. Chan, Total-text: A comprehensive dataset for scene text detection and recognition. in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 935–942 (2017). IEEE
44. D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V.R. Chandrasekhar, S. Lu, Icdar 2015 competition on robust reading. in *In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160 (2015). IEEE
45. B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, Robust scene text recognition with automatic rectification. in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 4168–4176 (2016)

46. B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, H. Shen, Single image super-resolution via a holistic attention network. In: ECCV (2020)
47. C. Zhao, S. Feng, B.N. Zhao, Z. Ding, J. Wu, F. Shen, H.T. Shen, Scene text image super-resolution via parallelly contextual attention network. in *Proceedings of the 29th ACM International Conference on Multimedia* (2021)

Publisher's Note