Scene Text Image Super-Resolution in the Wild

Wenjia Wang[⋆1], Enze Xie[⋆2], Xuebo Liu[1],
Wenhai Wang[3], Ding Liang[1], Chunhua Shen[⋆⋆4], and Xiang Bai[5]

[1] SenseTime Research [2] The University of Hong Kong [3] Nanjing University
[4] The University of Adelaide [5] Huazhong University of Science & Technology
wangwenjia@sensetime.com

Abstract. Low-resolution text images are often seen in natural scenes such as documents captured by mobile phones. Recognizing low-resolution text images is challenging because they lose detailed content information, leading to poor recognition accuracy. An intuitive solution is to introduce super-resolution (SR) techniques as pre-processing. However, previous single image super-resolution (SISR) methods are trained on synthetic low-resolution images (*e.g.* Bicubic down-sampling), which is simple and not suitable for real low-resolution text recognition. To this end, we propose a real scene text SR dataset, termed TextZoom. It contains paired real low-resolution and high-resolution images which are captured by cameras with different focal length in the wild. It is more authentic and challenging than synthetic data, as shown in Fig. 1. We argue improving the recognition accuracy is the ultimate goal for Scene Text SR. In this purpose, a new Text Super-Resolution Network, termed TSRN, with three novel modules is developed. (1) A sequential residual block is proposed to extract the sequential information of the text images. (2) A boundary-aware loss is designed to sharpen the character boundaries. (3) A central alignment module is proposed to relieve the misalignment problem in TextZoom. Extensive experiments on TextZoom demonstrate that our TSRN largely improves the recognition accuracy by over 13% of CRNN, and by nearly 9.0% of ASTER and MORAN compared to synthetic SR data. Furthermore, our TSRN clearly outperforms 7 state-of-the-art SR methods in boosting the recognition accuracy of LR images in TextZoom. For example, it outperforms LapSRN by over 5% and 8% on the recognition accuracy of ASTER and CRNN. Our results suggest that low-resolution text recognition in the wild is far from being solved, thus more research effort is needed. The codes and models will be released at: github.com/JasonBoy1/TextZoom

Keywords: Scene Text Recognition, Super-Resolution, Dataset, Sequence, Boundary

⋆ Equal Contribution.
⋆⋆ Corresponding Author.

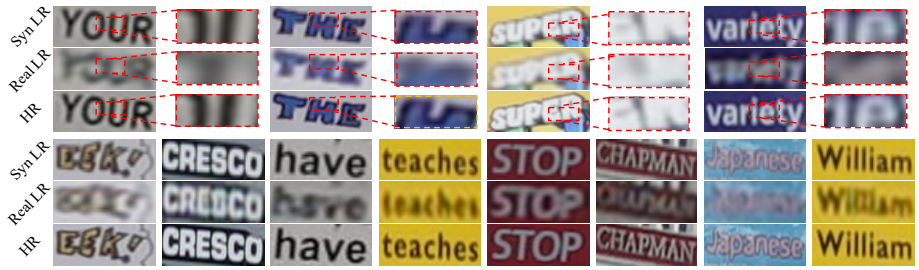

Fig. 1. Comparison between synthetic LR, real LR, and HR images in TextZoom. 'Syn LR' denotes BICUBIC down-sampled image of HR. 'Real LR' and 'HR' denotes LR and HR images captured by camera with different focal lengths. From the images we can find that the real LR images are much more challenging than the synthetic LR images.

Table 1. Statistics of TextZoom. The testing set is divided into 3 different subsets: easy, medium and hard. The recognition accuracy is tested by ASTER [37]. We see the recognition accuracy of LR images decreases when the difficulty increases. Our main purpose is to increase the recognition accuracy of the LR images by super-resolution.

TextZoom	train	test		
		easy	medium	hard
Image number	17367	1619	1411	1343
Accuracy(LR)	35.7%	62.4%	42.7%	31.6%
Accuracy(HR)	81.2%	94.2%	87.7%	76.2%
Gap	45.5%	31.8%	45.0%	44.6%

1 Introduction

Scene text recognition is a fundamental and important task in computer vision, since it is usually a key step towards many downstream text-related applications, including document retrieval, card recognition, license plate recognition, etc [35,34,43,3]. Scene Text recognition has achieved remarkable success due to the development of Convolutional Neural Network (CNN).

Many accurate and efficient methods have been proposed for most constrained scenarios (e.g., text in scanned copies or network images). Recent works focus on texts in natural scenes [25,26,6,28,37,44,41,42], which is much more challenging due to the high diversity of texts in blur, orientation, shape, and low-resolution. A thorough survey of recent advantages of text recognition can be found in [27] . Modern text recognizers have achieved impressive results on clear text images. However, their performances drop sharply when recognizing low-resolution text images [1]. The main difficulty to recognize LR text is that the optical degradation blurred the shape of the characters. **Therefore, it**

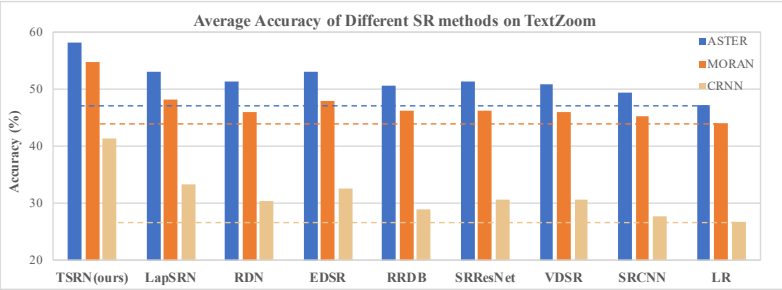

Fig. 2. Average recognition accuracy of the super-resolved images of LR images in TextZoom. We first super-resolve LR images with different SR methods, then directly test the SR results with the official released model of ASRER [37], MORAN [28] and CRNN [36]. We compare our TSRN with 7 state-of-the-art deep learning networks and show ours outperforms them clearly. Dotted lines means accuracy of LR inputs.

would be promising if we introduce SR methods as a pre-processing procedure before recognition. To our surprise, none of the real dataset and corresponding methods focus on scene text SR.

In this paper, we propose a paired scene text SR dataset, termed TextZoom, which is the **first dataset focus on real text SR**. Previous Super-Resolution methods [7,20,23,24,22,47,21] generate LR counterparts of the high-resolution (HR) images by simply applying uniform degradation like bicubic interpolation or blur kernels. Unfortunately, real blur scene text images are more varied in degradation formation. Scene texts are of arbitrary shapes, distributed illumination, and different backgrounds. Super-resolution on scene text images is much more challenging. Therefore, the proposed TextZoom, which contains paired LR and HR text images of the same text content, is very necessary. The TextZoom dataset is cropped from the newly proposed SISR datasets [4,46]. Our dataset has three main advantages. **(1)** This dataset is well annotated. We provide the direction, the text content and the original focal length of the text images. **(2)** The dataset contains abundant text from different natural scenes, including street views, libraries, shops, vehicle interiors and so on. **(3)** The dataset is carefully divided into three subsets by difficulty. Experiments on TextZoom demonstrate that our TSRN largely improves the recognition accuracy of CRNN by over 13% compared to synthetic SR data. The annotation and allocation strategy will be briefly introduced in section 3 and demonstrated in detail in supplementary materials.

Moreover, to reconstruct low-resolution text images, we propose a text-oriented end-to-end method. Traditional SISR methods only focus on reconstruct the detail of texture and only satisfy human's visual perception. However, scene text SR is quite a special task since it contains high-level text content. The fore-and-aft characters have information relations with each other. Obviously, a single blur character will not disable human to recognize the whole word if other characters are clear. To solve this task, firstly, we present a Sequential Residual Block

to model recurrent information in text lines, which enabling us to build a correlation in the fore-and-aft characters. Secondly, we propose a boundary-aware loss termed gradient profile loss to reconstructing the sharp boundary of the characters. This loss helps us to distinguish between the characters and backgrounds better and generate a more explicit shape. Thirdly, the misalignment of the paired images is inevitable due to the inaccuracy of the cameras. We propose a central alignment module to make the corresponding pixels more aligned. We evaluate the recognition accuracy by two steps: **(1)** Do super-resolution with different methods on LR text images; **(2)** Evaluate the SR text images with trained Text Recognizers *e.g.* ASTER, MOCAN and CRNN. Extensive experiments show our TSRN clearly outperforms 7 state-of-the-art SR methods in boosting the recognition accuracy of LR images in TextZoom. For example, it outperforms LapSRN by over 5% and 8% on recognition accuracy of ASTER and CRNN. Our results suggest that low-resolution text recognition in the wild is far from being solved, thus more research effort is needed.

The contributions of this work are therefore three-fold:

1. We introduce the first **real** paired scene text SR dataset TextZoom with different focal lengths. We annotate and allocate the dataset with three subsets: easy, medium and hard, respectively.
2. We prove the superiority of the proposed dataset TextZoom by comparing and analyzing the models trained on synthetic LR and proposed LR images. We also prove the necessity of scene text SR from different aspects.
3. We propose a new text super-resolution network with three novel modules. It surpasses 7 representative SR methods clearly by training and testing them on TextZoom for fair comparisons.

2 Related work

Super-Resolution. Super-resolution aims to output a plausible high-resolution image that is consistent with a given low-resolution image. Traditional approaches, such as bilinear, bicubic or designed filtering, leverage the insight that neighboring pixels usually exhibit similar colors and generate the output by interpolating between the colors of neighboring pixels according to a predefined formula. In the deep learning era, super-resolution is treated as a regression problem, where the input is the low-resolution image, and the target output is the high-resolution image [7,20,23,22,24,47,21]. A deep neural net is trained on the input and target output pairs to minimize some distance metric between the prediction and the ground truth. These works are mainly trained and evaluated on those popular datasets [2,45,30,14,31,40]. In these datasets, LR images are generated by a down-sample interpolation or Gaussian blur filter. Recently, several works capture LR-HR images pairs by adjusting the focal length of the cameras [4,46,5]. In [4,5], a pre-processing method is applied to reduce the misalignment between the captured LR and HR images While in [46], a contextual bilateral loss is proposed to leverage the misalignment. In this work, a new dataset TextZoom is proposed, which fills in the absence of paired scene text SR

dataset. It is well annotated and allocated with difficulty. We hope it can serve as a challenging benchmark.

Text Recognition. Early work adopts a bottom-up fashion [18] which detects individual characters firstly and integrates them into a word, or a top-down manner [16], which treats the word image patch as a whole and recognizes it as a multi-class image classification problem. Considering that scene text generally appears as a character sequence, CRNN [36] regard it as a sequence recognition problem and employs Recurrent Neural Network (RNNs) to model the sequential features. CTC [10] loss is often combined with the RNN outputs for calculating the conditional probability between the predicted sequences and the target [25,26]. Recently, an increasing number of recognition approaches based on the attention mechanism have achieved significant improvements [6,28]. ASTER [37] rectified oriented or curved text based on Spatial Transformer Network(STN) [17] and then performed recognition using an attentional sequence-to-sequence model. In this work, we choose state-of-the-art recognizer ASTER [37], MORAN [28] and CRNN [36] as baseline recognizers to evaluate the recognition accuracy of the SR images.

Scene Text Image Super-Resolution. Some previous works conducted on scene text image super-resolution are aimed at improving the recognition accuracy and image quality evaluation metrics. [29] compared the performance of several artificial filters on down-sampled text images. [32] propose a convolution-transposed convolution architecture to deal with binary document SR. [8] adapt SRCNN [7] in text image SR in the ICDAR 2015 competition TextSR [33] and achieved a good performance, but no text-oriented method was proposed.

These works take a step on low-resolution text recognition, but they only train on down-sampled images, learning to regress a simple mapping function of inverse-bicubic (or bilinear) interpolation. Since all the LR images are identically generated by a simple down-sample formulation, it is not well-generalized to real text images.

3 TextZoom Dataset

Data Collection & Annotation. Our proposed dataset TextZoom comes from two state-of-the-art SISR datasets: RealSR [4] and SRRAW [46]. These two newly proposed datasets consist of paired LR-HR images captured by digital cameras.

RealSR [4] is captured by four focal lengths with two digital cameras: Canon 5D3 and Nikon D810. In RealSR [4], these four focal lengths of images are allocated as ground truth, 2X LR images, 3X LR images, 4X LR images separately. For RealSR, we annotate the bounding box of the words on the 105mm focal length images. SR-RAW is collected by seven different focal lengths with SONY FE camera, range from 24-240mm. The images captured in shorted focal lengths could be used as LR images while those captured in longer lengths as corresponding ground truth. For SR-RAW, we annotate the bounding box of the words on the 240mm focal length images.

We labeled the images with the largest focal length of each group and cropped the text boxes from the rest following the same rectangle. So the misalignment is unavoidable. There are some top-down or vertical text boxes in the annotated results. In this task, we rotate all of these images to horizontal for better recognition. There are only a few curved text images in our dataset. For each pair of LR-HR images, we provide the annotation of the case sensitive character string (including punctuation), the type of the bounding box, and the original focal lengths. We demonstrate the detailed annotation principle of the text images cropped from SR-RAW and RealSR in detail in supplementary materials.

The size of the cropped text boxes is diverse, *e.g.* height from 7 to 1700 pixels, so it is not suitable to treat the text images cropped from the same focal lengths as a same domain. We define our principle following these considerations. **(1)No patching.** In SISR, data are usually generated by cropping patches from the original images [23,22,9,4,46]. Text images could not be cut into patches since the shape of the characters should maintain completed. **(2) Accuracy distribution.** We divide the text images by height and test the accuracy (Refer to the Tables showed in supplementary materials). We found that the accuracy does not increase obviously when the height is larger than 32 pixels. Setting images to 32 pixels height is also a customary rule in scene text recognition research [36,6,28]. The accuracy of the images smaller than 8 pixels are too low, which hardly has any value for super-resolution, so we discard the images the height of which is less than 8 pixels. **(3) Number.** We found that in the cropped text images, the height range from 8 to 32 claim the majority. **(4) No down-sample.** Since the interpolation degradation should not be introduced into real blur images, we could only up-sample the LR images to a relatively bigger size.

Following these 4 considerations, we up-sample the images ranging from 16-32 pixels height to 32 pixels height, and up-sample the images ranging from 8-16 pixels height to 16 pixels height. We conclude that (16, 32) should be a good pair to form a 2X train set for scene text SR task. For example, the text images taken from 150mm focal length and height sized in 16-32 pixels would be taken as a ground truth for the 70mm counterpart. So we selected all the images the height of which range from 16 pixels to 32 pixels as our ground truth image and up-sample them to the size of 128×32 (width×height), and the corresponding 2X LR images to the size of 64×16 (width×height). For this task, we only generate this 2X LR-HR pair dataset from the annotated text images mainly due to the special characteristics of text recognition. Other scale of factors of our annotated images could be used for different purpose.

Allocation of TextZoom. The SR-RAW and RealSR are collected by different cameras with different focal lengths. The distance from the objects also affect the legibility of the images. So the dataset should be further divided following their distribution.

The train-set and test-set are cropped from the original train-set and test-set in SR-RAW and RealSR separately. The author of SR-RAW used larger distance from the camera to the subjects to minimize the perspective shift [46]. So the accuracy of text images from SR-RAW is relatively lower under the similar focal

lengths compared to RealSR. The accuracy of the images cropped from 100mm focal lengths in SR-RAW is 52.1% tested by ASTER [37], while the accuracy of those from 105mm in RealSR is 75.0% tested by ASTER [37] (Refer to the Tables showed in supplementary materials). With the same height, the images of smaller focal lengths are more blurred. With this in mind, we allocate our dataset into three subsets by difficulty. The LR images cropped from RealSR render **easy**. The LR images from SR-RAW and the focal lengths of which larger than 50mm are viewed as **medium**. The rest are as **hard**.

In this task, our main purpose is to increase the **recognition accuracy** of the easy, medium and hard subsets. We also show the results of peak signal to noise ratio (PSNR) and structural similarity index (SSIM) in the supplementary materials.

Dataset Statistics The detailed statistics of TextZoom is shown in supplementary materials.

4 Method

In this section, we present our proposed method TSRN in detail. Firstly, we briefly describe our pipeline in section 4.1. Then we demonstrate the proposed Sequential Residual Block. Thirdly, we introduce our central alignment module. Finally, we introduce a new gradient profile loss to sharpen the text boundaries.

4.1 Pipeline

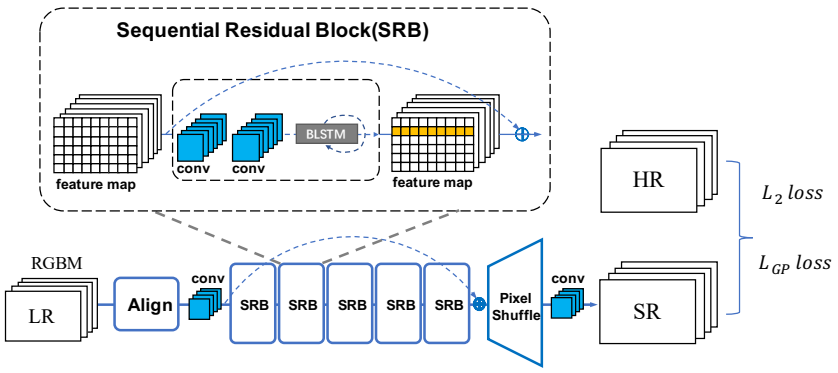

Fig. 3. The illustration of our proposed TSRN. We concatenate binary mask with RGB channels as a RGBM 4-channel input. The input is recitified by central alignment module and then fed into our pipeline. The output is the super-resolved RGB image. The outputs are supervised by L_2 *loss*. The RGB channels of the outputs are supervised by L_{GP} *loss*.

Our baseline is SRResNet [23]. As shown in Fig. 3, we mainly make two modifications to the structure of SRResNet: 1) add a central alignment module

in front of the network; 2) replace the original basic blocks with the proposed Sequential Residual Blocks (SRBs). In this work, we concatenate the binary mask with RGB image as our input. The binary masks are simply generated by calculating the mean gray scale of the image. The detailed information of masks is shown in supplementary materials. During training, firstly, the input is rectified by central alignment module. Then we use CNN layers to extract shallow features from the rectified image. Stacking five SRBs, we extract deeper and sequential dependent feature and do shortcut connection following ResNet [12]. The SR images are finally generated by up-sampling block and CNN. We also design a gradient prior loss (L_{GP}) aiming at enhancing the shape boundary of the characters. The output of the network is supervised by MSELoss (L_2) and our proposed gradient profile loss (L_{GP}).

4.2 Sequential Residual Block

Previous state-of-the-art SR methods mainly pursue better performance in PSNR and SSIM. Traditional SISR only cares about texture reconstruction while text images have strong sequential characteristics. In text recognition tasks, scene text images encode the context information for text recognition by Recurrent Neural Network (RNN) [13]. Inspired from them, we modified the residual blocks [23] by adding Bi-directional LSTM (BLSTM) mechanism. Inspired by [39], we build sequence connectionist in horizontal lines and fused the feature into deeper channels. Different from [39], we build the in-network recurrence architecture not for detecting but for low-level reconstruction, so we only adapt the idea of building text line sequence dependence. In Fig. 3, the SRB is briefly illustrated. Firstly, we extract feature by CNN. Then permute and resize the feature map as the horizontal text line can be encoded into sequence. Then the BLSTM can propagate error differentials [36], and invert the feature maps into feature sequences, and feed them back to the convolutional layers. To make the sequence dependent robust for tilted text images, we introduce the BLSTM from two directions, horizontal and vertical. BLSTM takes the horizontal and vertical convolutional feature as sequential inputs, and updates its internal state recurrently in the hidden layer.

$$\begin{aligned} H_{t_1} &= \phi_1(X_{t_1}, H_{t_1-1}), \quad t_1 = 1, 2, ..., W \\ H_{t_2} &= \phi_1(X_{t_2}, H_{t_2-1}), \quad t_2 = 1, 2, ..., H \end{aligned} \tag{1}$$

Here H_t denotes the hidden layers, X_t denotes the input features, t_1, t_2 separately denote the recurrent connection from horizontal and vertical direction.

4.3 Central Alignment Module

The misalignment make the pixel-to-pixel losses, such as L_1 and L_2 generate significant artifacts and double shadows. This mainly due to the misalignment of the pixels in training data. Sine some of the text pixels in LR images are in spatial corresponding to the background pixels in the HR images, the network could

learn a wrong pixel-wise counterpart information. As mentioned in Section. 3, the text regions in HR images are more central aligned compared to the LR images. So we introduce STN[17] as our central alignment module. The STN is a spatial transform network which can rectify the images and be learned end-to-end. To rectify spatial variation flexibly, we adopt TPS transformation as the transform manipulation. Once the text regions in LR images are aligned adjacent the center, the pixel-wise losses would make better performance and the artifacts could be relieved. We show more detailed information of central alignment module in supplementary materials.

4.4 Gradient Profile Loss

Gradient Profile Prior (GPP) is proposed in [38] to generate sharper edge in SISR task. Gradient field means the spatial gradient of the RGB values of the pixels.

Since we have a paired text super-resolution dataset, we could use the gradient field of HR images as ground truth. Generally, the color of characters in text images contrast strongly with the backgrounds.So sharpening the boundaries rather than smooth ones of characters could make the characters more explict.

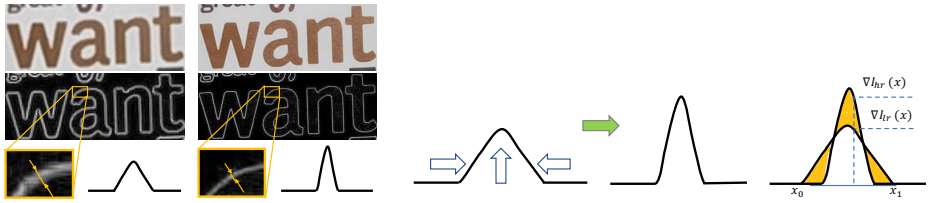

Fig. 4. The illustration of gradient field and Gradient Prior Loss.

We revisit the GPP and generate ground truth from HR images, then we define the loss function as below:

$$L_{GP} = \mathbb{E}_x ||\nabla I_{hr}(x) - \nabla I_{sr}(x)||_1 \qquad (x \in [x_0,\ x_1]) \tag{2}$$

$\nabla I_{hr}(x)$ denotes the gradient field of HR images, and $\nabla I_{sr}(x)$ denotes that of SR images.

Our proposed L_{GP} exhibits two advantageous properties: (1) The gradient field vividly show the characteristics of text images: the texts and backgrounds. (2) The LR images always come with wider curve of gradient field, while HR images mean thinner curve. And the curve of gradient field could be easily generated through mathematical calculation. This ensures a confidential supervision label.

5 Experiments

5.1 Datasets

We train the SR methods on our proposed TextZoom (see section 3.) training set. We evaluate our models on our three subsets **easy**, **medium** and **hard**. To avoid down-sample degradation, all the LR images are up-sampled to 64×16, and HR images to 128×32.

5.2 Implementation Details

During training, we set the trade-off weight of L_2 loss as 1 and L_{GP} as $1e-4$. We use the Adam optimizer with momentum term 0.9. When evaluating recognition accuracy, we use the official Pytorch version code and the released model of ASTER: `aster.pytorch`, MORAN: `MORAN_v2.pytorch`, CRNN: `crnn.pytorch` from github.

All the SR models are trained by 500 epochs with 4 NVIDIA GTX 1080ti GPUs. The batch-size is adapted as the setting in the original papers.

5.3 Is SR necessary for Text Recognition?

We further quantitatively analyzed the necessity of super-resolution from three aspects.

It is assumed that we could achieve better performance on recognizing low-resolution (LR) text images if we directly train the recognition networks on small size images, and then the super-resolution procedure could be removed. This query is reasonable because the deep neural networks have a strong robustness on the training domains. To refute this query and prove the necessity of super-resolution for text images, we compare the recognition accuracy of 4 methods:

- **Released.** Recognize with ASTER [37] model trained on customary size (no less than 32 pixels in height, We use official released model here).
- **ReIm.** Recognize with model trained on low-resolution images (In this work, we re-implemented ASTER [37] on Syn90K [15] and SynthText [11] at the size of 64×16, All the training details are the same as the original paper except the input sizes
- **Fine-tune.** Fine-tune released ASTER [37] model on our TextZoom training set.
- **Ours.** Choose the low resolution images by size, then use our proposed TSRN to generate the SR images and then recognize them with ASTER [37] official released model.).

To verify the robustness, we select all the images smaller than 64×16 from 7 common scene text testing sets, IC13, IC15, CUTE, IC03, SVT, SVTP, CUTE and IIIT5K and get 436 images in total. We term this testing set **CommonLR**. We compare these 4 methods on our dataset TextZoom and CommonLR. From Table 2, we can figure that the re-implemented model do increase the accuracy

Table 2. Comparison between different methods. **Released** means official released model from github. **ReIm** means our re-implemented model trained on Syn90K [15] and SynthText [11] at the size of 64×16.

Method	Recognition Accuracy	
	TextZoom	**CommonLR**
Released	47.2%	75.3%
ReIm	52.6%	79.3%
Fine-tune	**59.3%**	73.2%
Ours	58.3%	**80.3%**

sharply on the LR images. The average accuracy of TextZoom can be increased by 5.4%, from 47.2% to 52.6%. And the accuracy of CommonLR could also be improved for 5%. The result of re-implemented model is still lower than the accuracy of our results (TSRN(ours) + ASTER(Released)).

When we fine-tune the Aster on our TextZoom training set, the accuracy of TextZoom testing set would be even higher than our method. But TextZoom is a small sized dataset for recognition task, its different distribution would make the recognizer over-fit on it. The accuracy of CommonLR of fine-tune method is the lowest. Moreover, on this fine-tune Aster model the other testing sets like IC13, IC15, etc. would drop sharply for more than 10.0% points.

Actually, our method is superior to fine-tune and re-Im methods in following aspects. (1). The fine-tuned model over-fit on TextZoom. It achieves highest performance on TextZoom while lowest on CommomLR because the number of TextZoom is far from enough for text recognition task. Super-resolution ,a low-level task, usually needs less data to converge. Our method could directly choose SR or not by the size and get better overall performance.

(2).Our SR method can also produce better visual results for people to read (see Fig. 5). (3).While re-Im and fine-tune method need 2 recognition models for big and small size images separately, our method only need a tiny SR model, introducing marginal computation cost. This part could be found in supplementary materials.

So the SR methods could be a effective and convenient pre-processing procedure of scene text recognition.

5.4 Synthetic LR vs. TextZoom LR

To demonstrate the superiority of paired scene text SR images, we compare the performance of the models trained on synthetic datasets and our TextZoom dataset. The quantitative results are shown in the supplementary materials.

Table 3. Ablation study for different settings of our method TSRN. The recognition accuracies are tested by the official released model of ASTER [37].

Configuration			Accuracy of ASTER [37]			
	Method	Loss function	easy	medium	hard	**average**
0	SRResNet	$L_2 + L_{tv} + L_p$	69.6%	47.6%	34.3%	51.3%
1	5×SRBs	L_2	74.5%	53.3%	37.3%	56.2%
2	5×SRBs + align	L_2	74.8%	55.7%	39.6%	57.8%
3	5×SRBs + align **(Ours)**	$L_2 + L_{GP}$	**75.1%**	**56.3%**	**40.1%**	**58.3%**

5.5 Ablation Study on TSRN

In order to study the effect of each component in TSRN, we gradually modify the configuration of our network and compare their differences to build a best network. For brevity, we only compare the accuracy of ASTER [37].

Fig. 5. Visual comparisons for showing the effects of each component in our proposed TSRN. The recognition result strings of ASTER are displayed under each image. Those characters in red denote wrong recognition.

1) SRBs. We add BLSTM mechanism to the basic residual block in SRResNet [23] and get the proposed SRB. The SRB is the essential component in TSRN. Comparing # 0 and # 1 in Table 3 , stacking 5 SRBs, we can boost up the average accuracy by 4.9% compared to SRResNet [23].

2) Central Alignment Module. Central alignment module can boost the average accuracy by 1.5%, as shown in Table 3 method 2. From Fig. 5, we can find that without central alignment module, the artifacts are strong, and the characters are twisted. While with more appropriate alignment, we could generate higher quality images since the pixel-wise loss function could supervise the training better.

3) Gradient Profile Loss. From Table 3 method 3, we can find the proposed gradient profile loss can boost the average accuracy by 0.5%. Although the increase is slight, the visual results are better (Fig. 5 method 3).

In supplementary materials, we further discuss about the detailed component of our method.

5.6 Comparison with State-of-the-Art SR methods

Table 4. Performance of state-of-the-art SR methods on the three subsets in TextZoom. For better displaying, we calculated the average accuracy. L_1 denotes Mean Average Error (MAE) Loss. L_2 denotes Mean Squared Error (MSE) Loss. L_{tv} denotes Total Variation Loss. L_p denotes Perceptual Loss proposed in [19]. *Charbonnier* denotes the Charbonnier Loss proposed in LapSRN [21]. L_{GP} denotes our proposed Gradient Prior Loss. The recognition accuracies are tested by the official released model of ASTER [37], MORAN [28] and CRNN [36].

Method	Loss Function	Accuracy of ASTER [37]				Accuracy of MORAN [28]				Accuracy of CRNN [36]			
		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 [7]	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 [20]	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 [23]	$L_2 + L_{tv} + L_p$	69.6%	47.6%	34.3%	51.3%	60.7%	42.9%	32.6%	46.3%	39.7%	27.6%	22.7%	30.6%
RRDB [22]	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 [24]	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 [47]	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 [21]	*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%
TSRN(ours)	$L_2 + L_{GP}$	**75.1%**	**56.3%**	**40.1%**	**58.3%**	**70.1%**	**53.3%**	**37.9%**	**54.8%**	**52.5%**	**38.2%**	**31.4%**	**41.4%**
Improvement of TSRN		**10.4%**	**13.9%**	**8.9%**	**11.1%**	**9.5%**	**15.4%**	**7.1%**	**10.7%**	**16.1%**	**17.1%**	**10.3%**	**14.6%**

To prove the effectiveness of TSRN, we compare it with 7 SISR methods on our TextZoom dataset, including SRCNN [7], VDSR [20], SRResNet [23], RRDB [22], EDSR [24], RDN [47] and LapSRN [21]. All of the networks are trained on our TextZoom training set and evaluated on our three testing subsets.

In Table 4, we list the recognition accuracy tested by ASTER [37], MORAN [28], and CRNN [36] of all the mentioned 7 methods, along with BICUBIC and the proposed TSRN. In Table 4, it can be observed that TSRN outperforms all the 7 SISR methods in recognition accuracy sharply. Although these 7 SISR methods could achieve a relatively good accuracy, what we should pay attention to is the gap between SR results and BICUBIC. These methods could improve the average accuracy 2.3% $\sim$ 5.8%, while ours could improve 10.7% $\sim$14.6%. We can also find that our TSRN could improve the accuracy on all of the three state-of-the-art recognizers. In the supplementary materials, we show the results of PSNR and SSIM and show that our TSRN could also surpass most of the state-of-the-art methods in PSNR and SSIM.

6 Conclusion and Discussion

In this work, we verify the importance of scene text image super-resolution task. We proposed the TextZoom dataset, which is, to the best of our knowledge, the

Fig. 6. Visualization results of state-of-the-art SR methods on our proposed dataset TextZoom. The character strings under the images are recognition results of ASTER [37]. Those in red denote wrong recognition.

first real paired scene text image super-resolution dataset. The TextZoom is well annotated and allocated and divided into three subset:easy, medium and hard. Through extensive experiments, we demonstrated the superiority of real data over synthetic data. To tackle text images super-resolution task, we build a new text-oriented SR method TSRN. Our TSRN clearly outperforms 7 SR methods. It also shows low-resolution text SR and recognition is far from being solved, thus more research effort is needed.

In the future, we will capture more appropriately distributed text images. Extremely large and small images will be avoided. The images should also contain more kinds of languages, such as Chinese, French and Germany. We will also focus on new methods such as introducing recognition attention into the text super-resolution task.

Acknowledge Xiang Bai was supported by the Program for HUST Academic Frontier Youth Team 2017QYTD08.

References

1. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. arXiv preprint arXiv:1904.01906 (2019)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
3. Björklund, T., Fiandrotti, A., Annarumma, M., Francini, G., Magli, E.: Robust license plate recognition using neural networks trained on synthetic images. Pattern Recognition (2019)
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019)
5. Chen, C., Xiong, Z., Tian, X., Zha, Z., Wu, F.: Camera lens super-resolution. In: CVPR (2019)
6. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: Towards accurate text recognition in natural images. In: Proc. IEEE Int. Conf. Comp. Vis. (2017)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI (2015)
8. Dong, C., Zhu, X., Deng, Y., Loy, C.C., Qiao, Y.: Boosting optical character recognition: A super-resolution approach. arXiv preprint arXiv:1506.02211 (2015)
9. Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.): ECCV (2018)
10. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proc. Int. Conf. Mach. Learn. (2006)
11. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. He, P., Huang, W., Qiao, Y., Loy, C.C., Tang, X.: Reading scene text in deep convolutional sequences. In: AAAI (2016)
14. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
15. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
16. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV (2016)
17. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Proc. Adv. Neural Inf. Process. Syst. (2015)
18. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Proc. Eur. Conf. Comp. Vis. (2014)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proc. Eur. Conf. Comp. Vis. (2016)
20. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
21. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR (2017)

22. Leal-Taixé, L., Roth, S. (eds.): Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part V. Lecture Notes in Computer Science (2019)
23. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2017)
24. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR (2017)
25. Liu, W., Chen, C., Wong, K.Y.K., Su, Z., Han, J.: Star-net: a spatial attention residue network for scene text recognition. In: Proc. Brit. Mach. Vis. Conf. (2016)
26. Liu, Z., Li, Y., Ren, F., Goh, W.L., Yu, H.: Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network. In: Proc. AAAI Conf. on Arti. Intel. (2018)
27. Long, S., He, X., Ya, C.: Scene text detection and recognition: The deep learning era. arXiv preprint arXiv:1811.04256 (2018)
28. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognition (2019)
29. Mancas-Thillou, C., Mirmehdi, M.: An introduction to super-resolution text. In: Digital document processing (2007)
30. Martin, D.R., Fowlkes, C.C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
31. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications (2017)
32. Pandey, R.K., Vignesh, K., Ramakrishnan, A., et al.: Binary document image super resolution for improved readability and ocr performance. arXiv preprint arXiv:1812.02475 (2018)
33. Peyrard, C., Baccouche, M., Mamalet, F., Garcia, C.: Icdar2015 competition on text image super-resolution. In: ICDAR (2015)
34. Ray, A., Sharma, M., Upadhyay, A., Makwana, M., Chaudhury, S., Trivedi, A., Singh, A.P., Saini, A.K.: An end-to-end trainable framework for joint optimization of document enhancement and recognition. In: ICDAR, (2019)
35. Sánchez, J., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: A set of benchmarks for handwritten text recognition on historical documents. Pattern Recognition (2019)
36. Shi, B., Bai, X., Yao, C.: 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. (2017)
37. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. (2018)
38. Sun, J., Sun, J., Xu, Z., Shum, H.: Gradient profile prior and its applications in image super-resolution and enhancement. TIP (2011)
39. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: ECCV (2016)
40. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW (2017)
41. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: CVPR (2019)

42. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., Shen, C.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network (2019)
43. Wu, Y., Yin, F., Liu, C.: Improving handwritten chinese text recognition using neural network language models and convolutional neural network shape models. Pattern Recognition (2017)
44. Xie, E., Zang, Y., Shao, S., Yu, G., Yao, C., Li, G.: Scene text detection with supervised pyramid context network. In: AAAI (2019)
45. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces (2010)
46. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: CVPR (2019)
47. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)