Skeleton-aware Text Image Super-Resolution

Shimon Nakaune
nakaune@cvlab.cs.tsukba.ac.jp
Satoshi Iizuka
iizuka@cs.tsukuba.ac.jp
Kazuhiro Fukui
kfukui@cs.tsukuba.ac.jp

University of Tsukuba
Tsukuba, Japan

Abstract

We present a novel structure-aware loss function for text image super-resolution to improve the recognition accuracy of text recognizers in natural scenes. Text image super-resolution is a particular case of general image super-resolution, where our primary goal is to improve the readability of characters in a low-resolution image by increasing the resolution of the text image. In this scenario, general loss functions usually used in previous super-resolution models are insufficient to learn character shapes precisely and stably as it often leads to blurring and breaking of the shapes. In this paper, we propose a *skeleton loss* for training text super-resolution networks. Skeleton loss enables the networks to generate more readable characters by considering the detailed structural formation of character skeletons, in the optimization process. The key idea of the skeleton loss is to measure the differences between two types of character skeletons, where one is obtained from a high-resolution image and another is from the super-resolved image generated from a given low-resolution image. To implement this idea in an end-to-end form, we introduce a skeletonization network that can generate skeletons from an input text image. Quantitative analysis shows that our method outperforms existing super-resolution models with modern text recognizers in terms of recognition accuracy. Furthermore, our experiments show that our skeleton loss can boost generating readable text images of existing super-resolution networks without modifying their structures.

1 Introduction

Scene text recognition is an important technique, as it obtains useful textual information on various objects such as signboards and license plates in the real world. It is challenging to recognize scene texts accurately since each character of a text can have many different fonts, orientations, and resolutions in natural scenes. Recently, Convolutional Neural Networks (CNNs) have shown high performance in this task [17, 26, 28, 33]. However, even the performances of the CNN-based approaches can significantly drop when the text images are blurred and in low-resolution [1]. This paper aims to address this issue by increasing the resolution of an input scene text image based on a novel structure-aware loss function, called *skeleton loss*, for the text image super-resolution.

(a) Input (b) SAN [9] (c) TSRN [34] (d) Ours (e) Ground Truth

Figure 1: Comparison of our approach for text super-resolution with the state-of-the-art on a real scene text image. The upper row shows the text images, and the lower row shows their skeletons, the output of the skeletonization network. Existing approaches miss important strokes for text recognition, while our method generates strokes plausibly.

The task of converting a given Low-Resolution (LR) text image to a High-Resolution (HR) text image is referred to as text image Super-Resolution (SR). The popular approaches to this task are non-text-oriented SR methods that are CNN-based and do not consider the property of the characters [23]. In these methods, for a pair of LR and HR text images used in training, a corresponding LR image is generated from a given HR image by applying a simple down-sampling method such as the bicubic down-sampling. However, recent studies have revealed that CNNs trained with image sets generated in such a simple way do not work on actual images in the real-world due to the gap between the property of synthetic and actual text images [12]. Wang *et al*. [34] constructed the TextZoom dataset of paired actual LR and HR scene text images to focus on more practical text SR, and then proposed a new Text Super-Resolution Network (TSRN). This network outperforms the conventional non-text-oriented methods.

The loss functions widely used for the SR networks such as L1, L2 (MSE) losses, and local image gradient losses, are based on pixel-wise differences in the image space. Thus, they tend to generate deformed characters from highly blurred text images (Figure 1). To overcome this issue, several types of perceptual losses were proposed [7, 14, 25]. However, perceptual losses are still not suitable to reconstruct the shapes of characters in the text super-resolution, as they are designed to mainly handle natural images in the feature spaces produced by pre-trained image classification networks.

In this paper, we design a loss function to enhance the readability of the super-resolved text characters. We introduce a *skeleton loss* that measures the differences between two kinds of skeletonized characters of HR images and SR images in addition to the conventional loss functions. More concretely, in our skeleton loss, HR images and SR images are converted into skeleton images, and then the difference between the generated HR and SR skeletons is measured using cross-entropy loss combined with dice loss. Since a scene text image contains various degradations, a standard thinning method [36] may deform or distort the strokes composing the characters. To avoid this problem, we construct a skeletonization network with a fully convolutional network to convert input characters into smooth anti-aliased skeletons. Finally, our network is trained with pairs of text and skeleton images without using any additional manually annotated data. For the implementation of the SR network, we employ the sequential residual block-based network [34], combined with an attention mechanism [7, 22] that can capture global information of the whole text image.

We evaluate and compare our approach with various existing SR methods on a real scene text dataset and demonstrate that our model achieves the best performance in boosting the

recognition accuracy of LR images of four text recognizers [17, 26, 28, 33]. Furthermore, our experiment shows that our skeleton loss can improve the performance of existing models without changing their network architectures.

To summarize, our contributions are as follows: (1) a novel skeleton-based loss function that can improve the readability of the super-resolved text characters, (2) a text skeletonization network and its training strategy, and (3) an in-depth evaluation of our model with comparisons and ablation studies on challenging datasets.

2 Related Work

Image Super-Resolution. Estimating a high-resolution (HR) image from a given low-resolution (LR) image is called Single Image Super-Resolution (SISR), which is an ill-posed problem. The most basic approach to this problem uses image interpolation with bilinear and bicubic functions to generate an overall smooth texture with a low computational cost. Moreover, recently, deep Convolutional Neural Networks (CNNs) also have demonstrated excellent performances in SISR tasks [3, 4, 10, 13, 14, 16, 20, 37, 38]. Most of these methods are optimized with pixel-wise L1/L2 loss to maximize the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). As another approach, several losses that implement perceptual constraints have been proposed [7, 14, 25] to reflect human perception in the loss function. The perceptual loss proposed by Johnson *et al*. [7] is robust to local pixel-level changes as they are calculated by using the Euclidean distance between the feature vectors of the VGG network [30]. It can work effectively with high visual quality when they work in combination with adversarial training [14, 25].

Text Image Super-Resolution. In early studies, filtering-based super-resolution approaches were used for text images [18]. In ICDAR 2015 competition on the text image super-resolution [23], CNN-based methods showed high performance compared to other conventional methods, although they are not specifically designed for text images. However, their performances decrease significantly when used with real-world LR text images due to the domain gaps between the synthetic and real LR images [12]. For this reason, Wang *et al*. proposed TextZoom, which is a real paired LR and HR text image dataset, and developed a Text Super-Resolution Network (TSRN) [34]. Additionally, they proposed a gradient profile loss which is defined as the error of the image gradient field. In our work, we further improve the CNN-based methods by designing a loss function that focuses on the important character shapes for text recognition.

Text Recognition. Traditional text recognition approaches usually employ a bottom-up framework, which detects and classifies characters first, and then converts them into a word using lexicons [31]. Recently, various deep neural network-based methods have been studied to further improve text recognition accuracy [17, 26, 28, 33]. CRNN [26] combines a CNN and a Recurrent Neural Network (RNN) to extract sequential features, and then use a Connectionist Temporal Classification (CTC) [5] decoder to predict the character with the highest conditional probability at the current step based on the results of previous sequences. Afterward, attention-based decoders for the text recognition task were developed rapidly [15, 27]. ASTER [28] and MORAN [17], which have attention-based decoders, added a rectification module at the beginning of the recognition sequence to remove the negative effects of scene text distortion. DAN [33] uses a fully convolutional network (FCN) with attention mechanisms for encoders, and an RNN decoder that predicts characters from features automatically aligned with the attention map. We use ASTER, DAN, MORAN, and

CRNN as text recognizers to evaluate the performance of our method.

3 Proposed Method

An SR network is used to convert low-resolution (LR) images into super-resolved images. An additional network, which we call the skeletonization network, is used to train the SR network to recover accurate shapes of the text characters. An overview of our proposed method is shown in Figure 2.

The SR network is based on the Text Super-Resolution Network (TSRN) [34], which has a central alignment module and stacked Sequential Residual Blocks (SRBs). The alignment module rectifies the spatial misalignment of input text images to match their ground truth high-resolution (HR) images, while the SRB leverages bi-directional LSTMs [6] to extract sequential dependent features of the text images. Additionally, in order to capture more global dependencies, we employ a self-attention mechanism [35] at the end of the SRBs, which can learn spatially distant but important relationships such as the color differences between the background and each character. The inputs of this network are an RGB text image and a binary mask computed by thresholding the luminance of the input image using its average luminance following the TSRN. The outputs are an RGB text image and its mask with increased resolution.

The skeletonization network is a loss network based on TSRN. This network converts the text images into skeleton images, where all the strokes are normalized, and the background texture is removed. We use the skeletonization network to define a *skeleton loss* that measures the difference between the skeletonized characters in HR images and SR images. To minimize the skeleton loss, the skeleton of the SR image should be matched with its correct skeleton as closely as possible. This constraint allows the SR network to concentrate more on the character shape itself, thus generating text that is easy to recognize (Figure 1). For example, the network trained with the MSE loss tends to generate a character mixture of multiple characters due to the pixel-wise average of possible characters, *e.g.* "θ" mixed with "e" and "o". In this case, the skeleton loss can penalize such mixed characters as their skeletons are more clearly different from the ground truth skeletons. In addition, the loss network pre-trained for skeletonization has already learned semantic features of text characters, which allows transferring the semantic character prior from the loss network to the SR network. During training, the SR network is updated by taking the loss gradients through the pre-trained skeletonization network in backpropagation.

3.1 Skeletonization Network

Our skeletonization network consists of five sequential residual blocks [34] and two convolutional layers, as shown in Figure 2. The input of the skeletonization network is an RGB text image, and the output is a grayscale skeleton image. Unlike other skeleton extraction networks [19, 21], this network does not use a pooling operation that decreases the resolution to avoid losing the detailed features. For the last layer of the model, we use a Sigmoid function to ensure that the output is in the range of [0, 1].

We train the network using synthetic supervised data composed of pairs of texts and skeletons without manual annotations. For the skeleton dataset, we generate various text patterns by randomly combining characters with two types of fonts and background images with random color (Figure 3). The corresponding ground truth skeleton images are generated

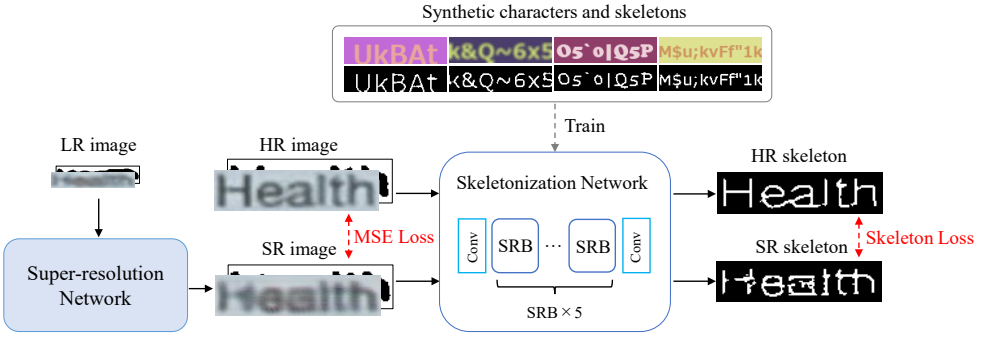

Figure 2: An overview of our approach. We use a pre-trained skeletonization network to define a *skeleton loss* that measures the difference between skeletonized characters in HR images and SR images. The skeleton loss combined with the MSE loss is used to train a super-resolution network.

by a line width normalization network proposed by Simo-Serra *et al.* [29], which can remove the background and thin the lines without deforming the strokes in clean text images. Since our skeletonization network should handle not only clean SR text images but also SR images partially blurred, we apply Gaussian blur to the synthetic texts to mimic the imperfect SR texts. Training is done with the Adam optimizer [11] with momentum term 0.9. The learning rate is 0.001 consistently. The input patches are all generated to be 128×32 pixels, and we use a batch size of 30 for 66,800 iterations. After the training, the skeletonization network is used to define the skeleton loss for training the SR network, where learnable weights of the skeletonization network remain fixed.

We note that this framework can be easily extended to other languages and fonts, such as Chinese characters, allowing the framework applied to various text images in the wild.

3.2 Skeleton Loss

The skeleton loss L_{Sk} is defined by the following equation:

$$L_{\mathrm{Sk}} = \alpha L_{\mathrm{BCE}} + (1-\alpha)(1-L_{\mathrm{Dice}}), \tag{1}$$

where L_{BCE} is the Binary Cross Entropy (BCE) loss and L_{Dice} is the Dice loss. We use $\alpha = 0.8$. In order for the skeletonization network to accurately predict skeletons and backgrounds, we use the BCE loss, which has a higher loss when misclassified, and the Dice loss, which focuses more on the skeleton region and addresses the class imbalance between the skeleton and background regions. This BCE loss can be expressed by the following equation:

$$L_{\mathrm{BCE}} = -\{T(\phi(I^{HR}))\log\phi(\psi(I^{LR})) + (1-T(\phi(I^{HR})))\log(1-\phi(\psi(I^{LR})))\}, \tag{2}$$

where ψ is the SR network, ϕ is the skeletonization network, T is the binarization operation with a threshold of 0.5, I^{LR} and I^{HR} are the low-resolution and high-resolution images, respectively. The Dice loss is expressed by the following equation:

$$L_{\mathrm{Dice}} = \frac{2 \times \phi(\psi(I^{LR})) \odot T(\phi(I^{HR}))}{\phi(\psi(I^{LR})) + T(\phi(I^{HR}))}, \tag{3}$$

Figure 3: Examples of generated supervised data for training the skeletonization network. The top row shows synthetic text images, the middle row shows ground truth skeleton images, and the bottom row shows outputs of the trained skeletonization network.

where $\odot$ denotes the adamantine product.

Total loss used to train the SR network can be defined as:

$$L = L_{\mathrm{MSE}} + \beta L_{\mathrm{Sk}}, \tag{4}$$

where L_{MSE} is the Mean Squared Error (MSE), which is expressed as $L_{\mathrm{MSE}} = \|\psi(I^{LR}) - I^{HR}\|^2_{\mathrm{FRO}}$ using the Frobenius distance $\|\cdot\|_{\mathrm{FRO}}$. We set the coefficient $\beta = 0.008$.

4 Experimental Results

We evaluate our approach on the TextZoom dataset [34]. The TextZoom dataset contains paired real low-resolution and high-resolution images that are captured with different focal lengths. The paired data is annotated and divided into three subsets by difficulty: easy, medium, and hard. The image size is 64×16 pixels or less for the low-resolution image, and 128×32 pixels or less for the high-resolution image, which is data for $2\times$ super-resolution. TextZoom is a challenging real scene text dataset as it includes various fonts, character orientations, and degradations.

We use text recognition accuracy as our evaluation metric, which is tested by ASTER [28], a CNN-based text recognizer. For the text image super-resolution task, the text recognition accuracy is one of the most practical metrics [34]. We used a Pytorch version trained model of ASTER that is publicly available[1]. Additionally, to further validate our approach, we evaluate the recognition accuracy using additional text recognizers: CRNN [26], MORAN [17], and DAN [33].

Our SR network is trained with Adam [11] optimizer by setting the momentum term to 0.9. The batch size is 30, and the learning rate is initialized as 0.001 and then reduced to half every 2×10^5 iterations. Following the work of [34], the models are trained for 500 epochs, and the one with the best accuracy is used.

4.1 Comparison with Existing Approaches

We compared the proposed method with 10 state-of-the-art SR models such as SRCNN [4], VDSR [10], SRResNet [14], EDSR [16], RDN [38], LapSRN [13], RCAN [37], SAN [3], HAN [20], and TSRN [34]. All the models are evaluated in the same manner, where the batch size and optimizer setting are the same as the original papers.

Results of the text recognition accuracy of ASTER [28] are shown in Table 1. Our proposed model has the highest recognition accuracy in all subsets and achieved an average

[1] ASTER Pytorch

Table 1: Text recognition accuracy of SR images tested by ASTER [28]. SA denotes Self-Attention. Highest accuracy in **bold**, second in underline.

Model	Loss Function	Accuracy of ASTER [28] [%] easy	medium	hard	**average**
Bicubic	-	64.7	42.4	31.2	47.2
SRCNN[4]	L_2	69.7	44.7	33.4	50.5
VDSR[10]	L_2	65.0	47.1	33.2	49.6
SRResNet[14]	L_2	68.9	47.5	34.6	51.5
EDSR[16]	L_1	69.1	51.5	37.1	53.6
RDN[38]	L_1	65.8	46.0	33.4	49.5
LapSRN[13]	*Charbonnier*	67.8	46.5	32.8	50.2
RCAN[37]	L_1	69.2	49.2	35.9	52.5
SAN[3]	L_1	71.2	51.5	37.5	54.5
HAN[20]	L_1	64.7	48.0	35.0	50.2
TSRN[34]	$L_2 + L_{gp}$	74.2	57.2	40.1	58.2
TSRN + SA (**Ours**)	$L_2 + \boldsymbol{L_{sk}}$	**77.3**	**59.6**	**42.7**	**60.9**

Table 2: Text recognition accuracy of SR images tested by DAN [33], MORAN [17] and CRNN [26].

Model	Accuracy of DAN [33] [%] easy	medium	hard	average	Accuracy of MORAN [17] [%] easy	medium	hard	average	Accuracy of CRNN [26] [%] easy	medium	hard	average
Bicubic	65.2	41.3	30.2	46.7	60.6	37.9	30.8	44.1	36.4	21.1	21.1	26.8
SRCNN [4]	67.5	42.8	31.1	48.3	62.7	41.1	31.6	46.2	39.8	22.5	21.7	28.7
VDSR [10]	63.9	44.7	33.4	48.3	61.2	43.9	31.4	46.5	40.3	25.5	23.5	30.3
SRResNet [14]	65.8	46.2	33.5	49.6	62.9	45.4	32.3	47.9	43.9	28.8	25.6	33.1
EDSR [16]	66.8	47.8	36.1	51.3	65.6	47.5	34.8	50.3	49.3	33.7	27.3	37.5
RDN [38]	64.7	44.9	31.9	48.3	62.1	43.2	32.6	46.9	41.8	27.6	23.5	31.6
LapSRN [13]	67.0	44.2	31.9	48.8	63.3	42.1	30.5	46.4	40.0	24.8	23.0	30.0
RCAN [37]	68.5	45.9	33.8	50.5	67.0	46.3	33.5	50.0	50.8	33.2	28.2	38.2
SAN [3]	69.4	48.1	36.3	52.3	68.6	45.6	36.7	51.4	54.2	34.8	29.0	40.2
HAN [20]	62.7	46.1	33.6	48.4	60.8	43.1	33.1	46.6	44.1	28.9	26.4	33.8
TSRN [34]	71.7	53.7	38.3	55.6	69.7	53.9	38.4	54.9	55.0	40.5	31.8	43.2
Ours	**74.3**	**57.3**	**40.5**	**58.5**	**72.9**	**55.9**	**40.6**	**57.5**	**56.5**	**42.3**	**32.5**	**44.5**

of 2.7% higher than the baseline model [34]. The recognition accuracy of the other three text recognizers ([17, 26, 33]) are shown in Table 2. Our proposed method outperforms all the existing models for all text recognizers, which shows the robust performance of our model.

We provide a qualitative comparison of the SR images in Figure 4. The existing methods tend to produce partially collapsed characters that are hard to recognize. In contrast, our method generates plausible characters, closer to the ground truth characters.

In addition to evaluating text recognition accuracy, we further evaluate our approach using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which are standard quantitative metrics for image super-resolution. The results of these metrics are shown in Table 3. In contrast to the results of text recognition accuracy shown in Tables 1 and 2, our approach tends to be slightly lower than the existing models in terms of PSNR and SSIM. These results are due to the nature of PSNR and SSIM: they tend to rate high for overly smoothed images in contrast to human perception, as discussed in [14]. We found this limitation is common for evaluating text image super-resolution, where they fail to accurately assess SR text image quality in terms of the ease of recognition.

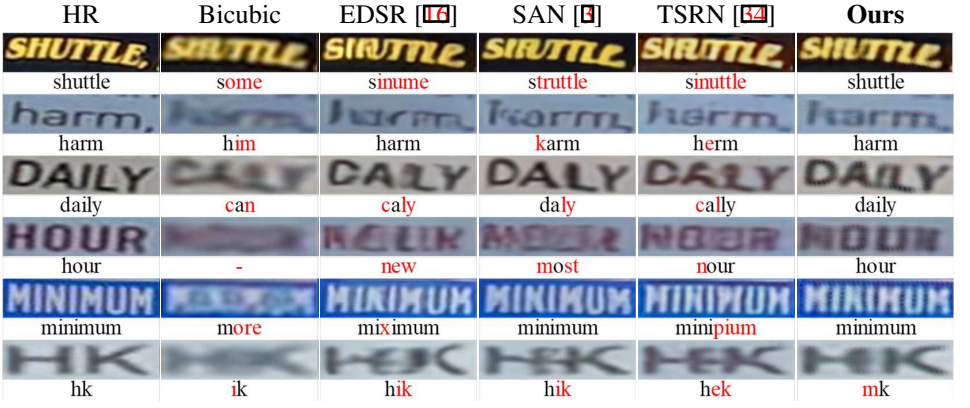

Figure 4: Visual comparison with existing methods. Text recognition results by ASTER [28] are shown under the images, where misrecognized characters are shown in red.

Table 3: Comparison with exiting methods using the PSNR and SSIM metric. Highest measures in **bold**, second in underline.

	PSNR			SSIM		
Model	easy	medium	hard	easy	medium	hard
Bicubic	22.35	18.98	19.39	0.7884	0.6254	0.6592
SRCNN [9]	23.32	**19.57**	20.13	0.8238	0.6409	0.6928
VDSR [10]	23.30	19.42	20.00	0.8244	0.6424	0.7012
SRResNet [14]	23.01	19.53	20.01	0.8339	0.6449	0.7097
EDSR [16]	23.42	18.47	19.48	0.8520	0.6395	0.7206
RDN [33]	23.43	19.56	20.10	0.8445	0.6466	0.7153
LapSRN [13]	23.32	19.26	19.86	0.8339	0.6411	0.7046
RCAN [32]	22.67	18.77	20.04	0.8659	0.6479	0.7331
SAN [3]	**24.14**	19.52	**20.40**	**0.8709**	0.6519	**0.7375**
HAN [20]	23.00	19.34	20.09	0.8364	0.6424	0.7198
TSRN [34]	23.41	19.05	19.85	0.8688	**0.6726**	0.7351
Ours	23.92	18.78	19.81	0.8697	0.6676	0.7311

4.2 Ablation Study

Comparison with Different Loss Functions. We compare our skeleton loss with several loss functions typically used for image super-resolution. The results are shown in Table 4. In this experiment, we use a TSRN [34] model as an SR network, and the coefficient of the skeleton loss combined with L1 loss is the same as the one with L2 loss. In the case of using adversarial loss, we pre-train the SR network without using the adversarial loss and then train it with all losses, including the adversarial loss. The results show that our skeleton loss further encourages the network to improve the readability of SR characters compared to other loss functions such as gradient field loss and perceptual loss. The adversarial loss makes the difference between text and background clearer but does not recover the correct characters. Therefore, the combination of L2 loss and adversarial loss degrades the text recognition accuracy. In contrast, the combination of our skeleton loss and adversarial loss tends to improve the accuracy because the adversarial loss clarifies the characters correctly recovered by the skeleton loss.

Effectiveness of Skeleton Loss. To further verify the effectiveness of the skeleton loss,

Table 4: Comparison with different loss functions. L_{gp}, L_{tv} , L_p, L_{adv} and L_{sk} denote gradient profile loss [34], total variation loss, perceptual loss, adversarial loss [7] and our skeleton loss, respectively.

Model	Loss Function	Accuracy of ASTER [28] [%]			
		easy	medium	hard	**average**
TSRN	L_1	72.1	55.1	37.0	55.8
TSRN	L_2	74.2	57.8	40.3	58.5
TSRN [34]	L_2+L_{gp}	74.2	57.2	40.1	58.2
TSRN	$L_2+L_{tv}+L_p$	74.6	57.0	39.0	58.0
TSRN	L_2+L_{adv}	74.2	55.7	41.0	58.0
TSRN	$L_2+\boldsymbol{L_{sk}}$	<u>75.5</u>	**60.1**	42.3	<u>60.3</u>
TSRN	$L_2+\boldsymbol{L_{sk}}+L_{gp}$	<u>75.5</u>	59.3	**42.5**	60.1
TSRN	$L_2+\boldsymbol{L_{sk}}+L_{adv}$	**75.8**	<u>59.6</u>	**42.5**	**60.4**

Table 5: Effectiveness of skeleton loss on different super-resolution models.

Model	Loss Function	Accuracy of ASTER [28] [%]			
		easy	medium	hard	**average**
EDSR [16]	L_1	69.1	51.5	**37.1**	53.6
EDSR	$L_1+\boldsymbol{L_{sk}}$	**70.0**	**52.5**	36.5	**53.9**
RCAN [37]	L_1	**69.2**	49.2	35.9	52.5
RCAN	$L_1+\boldsymbol{L_{sk}}$	68.1	**50.9**	**36.8**	**52.9**
SAN [3]	L_1	71.2	51.5	37.5	54.5
SAN	$L_1+\boldsymbol{L_{sk}}$	**71.8**	**52.4**	**38.4**	**55.3**
HAN [20]	L_1	64.7	48.0	35.0	50.2
HAN	$L_1+\boldsymbol{L_{sk}}$	**67.0**	**49.5**	**36.4**	**52.0**
Ours	L_2	75.2	57.4	41.7	59.2
Ours	$L_2+\boldsymbol{L_{sk}}$	**77.3**	**59.6**	**42.7**	**60.9**

we experiment by adding the skeleton loss to train five SR models: EDSR [16], RCAN [37], SAN [3], HAN [20], and our model. As shown in Table 5, the skeleton loss boosts the text recognition accuracy of all the baseline models consistently.

4.3 Results on Other Scene Text Recognition Datasets

We further validate the generalization capability of our model on several well-known scene text recognition datasets SVT [32], SVT-P [24], IC13 [8],and IC15 [9]. In this experiment, we generate LR text images in two ways: by down-sampling the images to 64×16 pixels and applying Gaussian blur, and by selecting low-resolution images (smaller than 64×16) from the datasets and upsampling them to 64×16 pixels. We super-resolve the LR images using the SR models and test the SR results with ASTER [28] to measure text recognition accuracy. All the SR models are trained on TextZoom dataset. As shown in Table 6, our method achieved the best or comparable accuracy compared to the existing approaches on all the datasets.

4.4 Discussion

The skeletonization network may not be able to extract accurate skeletons from text images with strong blur or special fonts, which can cause adverse effects on text super-resolution.

Table 6: Results on other text recognition datasets. Results show text recognition accuracy of ASTER [28]. Syn and Real denote down-sampled LR images algorithmically and real LR images, respectively. All the SR models are trained on TextZoom dataset.

Dataset	SVT [32]	SVT-P [24]	IC13 [8]	IC15 [9]	
LR type	Syn	Syn	Syn	Syn	Real
No. of imges	647	645	1081	2077	128
Bicubic	17.9	8.1	31.4	13.5	60.9
EDSR [16]	18.1	10.1	33.5	14.0	56.3
SAN [3]	16.9	9.8	30.9	11.4	52.3
TSRN [34]	20.9	**14.7**	37.1	16.1	60.9
Ours	**24.1**	14.3	**41.3**	**17.1**	**67.2**

Figure 5: Failure cases of our skeletonization. Strong blur and text with outlines, multiple text colors, and background colors are factors that cause skeletonization failure.

For example, in Figure 5, the skeletonization network fails to identify the text region and outputs its broken text boundary. Note that this is a minor case in the entire dataset and actually has little impact on the training.

Training Cost. Our approach requires the use of the skeletonization network as a loss network to train an SR network, which incurs an additional computational cost. Specifically, the skeletonization network has 627k parameters with recursive blocks. Our SR network with the skeleton loss takes 0.35 seconds for a batch of 30 images with 64×16 pixels for training, while the one without the skeleton loss takes 0.11 seconds. Overall training time is 50 hours for our model and 36 hours for the model without the skeleton loss. Note that the skeletonization network is used only for training the SR network and is not used during the testing, i.e., no additional cost is required for image super-resolution with our model.

5 Conclusion

In this paper, we presented a novel structure-aware loss function, called *skeleton loss*, for text image super-resolution to improve the recognition accuracy of text recognizers in natural scenes. The key idea of our skeleton loss is to measure the differences between two types of character skeletons in high-resolution and super-resolved images. To implement this idea in an end-to-end form, we introduced a skeletonization network based on our skeleton loss and a super-resolution network. The former converts input characters into smooth anti-aliased skeletons. The latter is based on an existing sequential residual block-based network with an attention mechanism to capture the global structure of the text images. Quantitative analysis showed that *skeleton loss* enables the super-resolution networks to generate more readable characters by extracting the detailed structural information from characters through skeletonization. Moreover, the analysis showed that the proposed method outperforms various existing super-resolution models with modern text recognizers in terms of recognition accuracy.

References

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

[2] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561.

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

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

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

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[8] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

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

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

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, Andreas Maier, and Christian Riess. Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE transactions on pattern analysis and machine intelligence*, 42(11): 2944–2959, 2019.

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

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

[15] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2231–2239, 2016.

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

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

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

[19] Sabari Nathan and Priya Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

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

[21] Oleg Panichev and Alona Voloshyna. U-net based convolutional neural network for skeleton extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[22] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, 2016.

[23] Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. Icdar2015 competition on text image super-resolution. In *13th International Conference on Document Analysis and Recognition*, pages 1201–1205, 2015.

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

[25] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[26] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[27] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016.

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

[29] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Real-Time Data-Driven Interactive Rough Sketch Inking. *ACM Transactions on Graphics*, 37(4), 2018.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[31] Kai Wang and Serge Belongie. Word spotting in the wild. In *European conference on computer vision*, pages 591–604, 2010.

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

[33] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12216–12224, 2020.

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

[35] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363, 2019.

[36] TY Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

[37] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision*, pages 286–301, 2018.

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