Gradient-Based Graph Attention for Scene Text Image Super-resolution

Xiangyuan Zhu[1], Kehua Guo[1*], Hui Fang[2], Rui Ding[1], Zheng Wu[1], Gerald Schaefer[2]

[1]School of Computer Science and Engineering, Central South University, China
[2]Department of Computer Science, Loughborough University, U.K.
zhuxiangyuan@csu.edu.cn, guokehua@csu.edu.cn, h.fang@lboro.ac.uk, ruiding@csu.edu.cn,
wuzhenghuse@gmail.com, gerald.schaefer@ieee.org

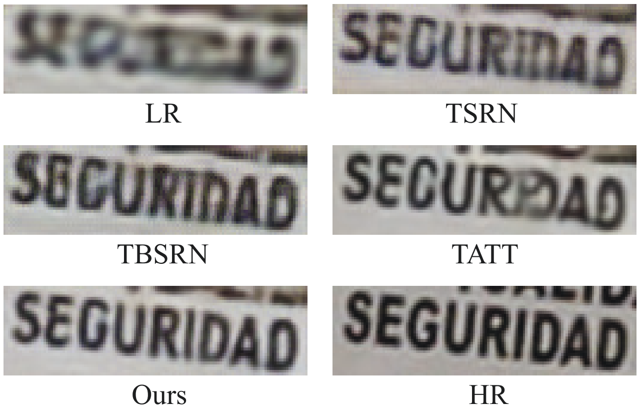
Abstract

Scene text image super-resolution (STISR) in the wild has been shown to be beneficial to support improved vision-based text recognition from low-resolution imagery. An intuitive way to enhance STISR performance is to explore the well-structured and repetitive layout characteristics of text and exploit these as prior knowledge to guide model convergence. In this paper, we propose a novel gradient-based graph attention method to embed patch-wise text layout contexts into image feature representations for high-resolution text image reconstruction in an implicit and elegant manner. We introduce a non-local group-wise attention module to extract text features which are then enhanced by a cascaded channel attention module and a novel gradient-based graph attention module in order to obtain more effective representations by exploring correlations of regional and local patch-wise text layout properties. Extensive experiments on the benchmark TextZoom dataset convincingly demonstrate that our method supports excellent text recognition and outperforms the current state-of-the-art in STISR. The source code is available at https://github.com/xyzhu1/TSAN.

Introduction

Scene text image super-resolution (STISR) is an emerging research task that can be applied to a wide range of real-world applications where low-resolution (LR) text images are observed due to various limitations such as low-quality capture devices, small text font sizes, or capturing at a distance (Wu, Yin, and Liu 2017; Shi, Bai, and Yao 2016). In addition, the uncontrolled capturing conditions in the wild lead to further deteriorations of the text quality in these images. Since the performance of state-of-the-art computer vision-based text recognition methods on these images are not satisfactory, image super-resolution techniques can be employed as a pre-processing step to reconstruct high-resolution (HR) images and thus improve the recognition accuracy (Cheng et al. 2018).

In one of the earliest attempts, Dong et al. (2015b) extend the classical SRCNN network for STISR, while Mou et al. (2020) introduce PlugNet to enhance feature-level representations of degraded images using a plugable super-resolution unit. However, these approaches are trained on synthetic LR-HR image pairs without any exploitation of text characteristics. Operating on TextZoom, the first STISR in-the-wild dataset, Wang et al. (2020) propose a central alignment module and gradient profile loss to both enforce text alignment and highlight the recovery quality of text contours to improve SR performance. To further exploit holistic text structures, Chen, Li, and Xue (2021) introduce a position-aware module and a content-aware module, whereas Chen et al. (2022) design a stroke-focussed module to embed text stroke-level structure into the learning representations for better text reconstruction. Despite the quality improvements achieved by these methods, how to make effective use of the well-structured repetitive patterns of text as prior knowledge for STISR task remains under-explored.



LR

TSRN

TBSRN

TATT

Ours

HR

Figure 1: Comparison of different STISR methods. Previous methods have difficulties reconstructing the texture details of an LR scene text image. In contrast, our method recovers clear text from the image.

In this paper, we exploit the local repetitive patterns and well-structured layout information of text to improve the performance of STISR in the wild. For this, we introduce a novel two-stage attention scheme (TSAN) to enhance text feature representations by highlighting well-structured patterns through a non-local group-wise attention module followed by a sparse graph attention module and a cascaded cross-channel attention module. With this elegant design, local salient text structure information is implicitly embedded

[*]Corresponding author.

into the model to improve the text SR reconstruction quality. As illustrated in Figure 1, our text attention mechanism significantly improves the text reconstruction quality to thus support better subsequent recognition. In an extensive set of experiments on the benchmark TextZoom dataset (Wang et al. 2020), we compare our approach to 19 state-of-the-art (SOTA) super-resolution methods, including both generic SR and specialised STISR techniques, and demonstrate it to outperform the current SOTA, while in a detailed ablation study we show how each of the proposed components contributes to this performance.

The main contributions of our work are as follows:

- We propose a novel text attention scheme to enhance text feature representations by exploring text structure information in an effective manner. Our approach leads to a significant improvement of model reliability based on several attention modules.
- We design a gradient-based graph attention module to capture regional text structure in order to enforce text layout prior embedding. To our best knowledge, this is the first work to investigate a graph attention scheme to embed text structure information into feature representations for STISR.
- We convincingly demonstrate the necessity of embedding the text layout prior via complementary text structure attentions in an effective manner, and show our proposed model to achieve superior performance compared to the current SOTA in STISR.

Related Work

We briefly review some related work on scene text recognition, single image super resolution, and scene text image super resolution.

Scene Text Recognition

Scene text recognition is a widely studied computer vision task that has a long history (Liu, Chen, and Wong 2018; Liao et al. 2019). Among the many approaches, some are commonly used as benchmarks to verify the performance of STISR methods (Zhao et al. 2021). In the CRNN model (Shi, Bai, and Yao 2016), feature extraction is integrated with sequence modelling and transcription into a unified framework. Shi et al. (2019) introduce an end-to-end neural network model named ASTER that uses a thin-plate spline algorithm to handle text irregularities and an attention enhanced seq2seq model to improve the scene text recognition. Benefitting from attention mechanisms, MORAN (Luo, Jin, and Sun 2019) improves the recognition reliability by training in a weakly supervised manner which only requires images and the corresponding text labels. Although these scene text recognition methods provide good performance, they still face difficulties, in particular when trying to recognise text from low-resolution or blurred images.

Single Image Super-Resolution

Single image super-resolution (SISR) aims to recover a high-resolution image from its low-resolution counterpart. Dong et al. (2015a) are the first to use a convolutional neural network, named SRCNN, for SISR. Since then, many deeper residual network architectures have been designed to improve SISR performance. For example, Kim, Lee, and Lee (2016) present a 20-layer network, VDSR, while Lim et al. (2017) introduce their EDSR model which expands the number of channels and layers of a conventional residual network. Attention mechanisms have also been extensively investigated to further boost model performance. Dai et al. (2019) propose SAN to explore the feature correlations of intermediate layers, while Hui et al. (2019) introduce IMDN to explore channel correlations. Compared to these attention-based models, HAN (Niu et al. 2020) enhances its feature representations from global image feature interdependencies in a more comprehensive manner. Although these approaches can be deployed for STISR, their designs do not exploit text structure information for better scene text image reconstruction.

Scene Text Image Super-Resolution

Since the release of TextZoom as the first realistic STISR dataset (Wang et al. 2020), STISR in the wild has attracted increased research attention. Wang et al. (2020) design a TSRN network with a boundary-aware loss term to explore the structured contours of text. Similarly, Zhao et al. (2021) present PCAN to learn sequence-dependent features along horizontal and vertical directions while preserving the block-shaped text layout information. Chen, Li, and Xue (2021) introduce their TBSRN model to explore both text-level and character-level contexts through a position-aware and a content-aware module. Inspired by gestalt psychology, which suggests that humans recognise objects with the guidance of similar object parts, Chen et al. (2022) propose a stroke-level-aware feature enhancement mechanism for STISR. Most recently, TATT (Ma, Liang, and Zhang 2022), employs a transformer-based module to emphasise text priors and proposes a structure consistency loss term to align features from regular and deformed texts. While they explore text priors from different perspectives, a combination of text priors from different levels should also be promising to further boost STISR performance.

Proposed Method

Motivation and Model Overview

Scene text images contain text and characters with well-structured layout and repetitive stroke-level patterns. These characteristics can be embedded into the model learning stage as prior knowledge to improve SR text reconstruction. Inspired by the text gestalt approach in (Chen et al. 2022), we assume that the perceptional improvement of text images is mainly driven by salient texture details and local repetitive structures in these images. We thus propose a novel attention enhancement method to learn these patterns in order to achieve better STISR performance. As depicted in Figure 2, in the first stage, we employ a group-wise attention module (GWAM) to enhance feature representations of the LR text image. Motivated by the intuitive observation that stroke-level patterns of text are represented well

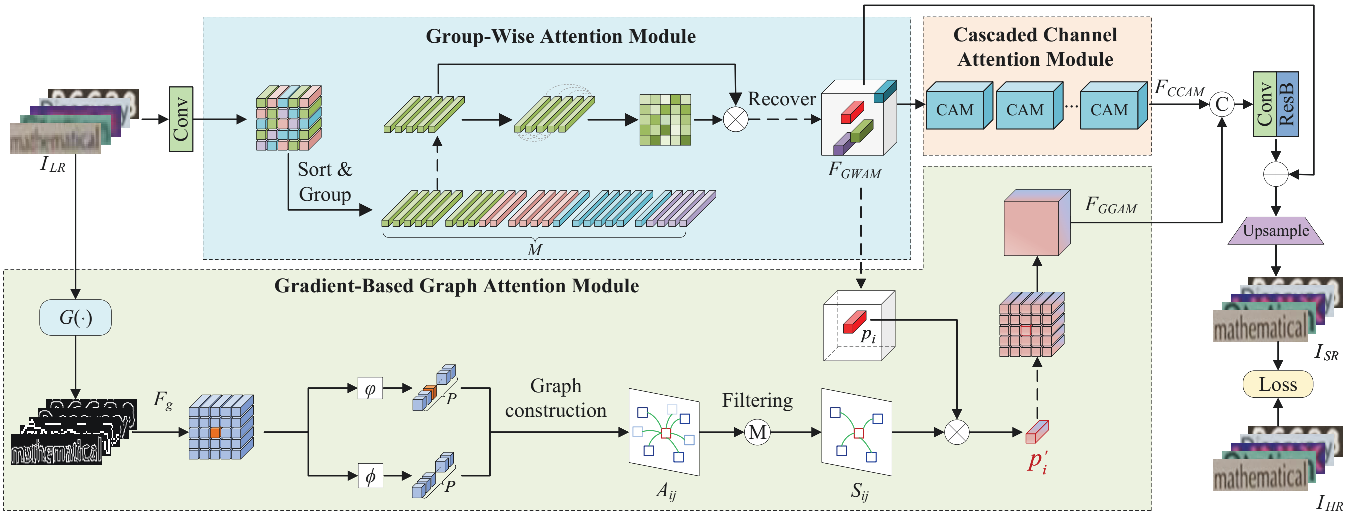


Figure 2: Overview architecture of the proposed method. Given a low-resolution input image, one branch of the model enhances the feature representations, while the other branch aims to refine boundary features.

by local edges, we introduce a novel gradient-based graph attention module (GGAM) to further refine the boundary features extracted from GWAM. Meanwhile, the GWAM features are also enhanced by a cascaded channel attention module (CCAM) to exploit feature correlations across salient regions in the channels. Importantly, GWAM and GGAM are complementary to each other although both explore contextual self-similarity for representational enhancement. GWAM focusses on enhancing generic feature representations due to its attention resource from low-level image patterns. In contrast, GGAM explores text contour attentions since the gradient maps divided into explicit patches are used to derive the attentions.

The overall network architecture of our proposed method is shown in Figure 2. Given a low-resolution image I_{LR} as input, a convolutional layer is used for initial feature extraction. This feature representation is then fed into GWAM for salient feature enhancement, expressed as

$$F_{GWAM} = H_{GWAM}\left(conv\left(I_{LR}\right)\right), \tag{1}$$

where $H_{GWAM}(\cdot)$ denotes the GWAM module and $conv(\cdot)$ represents the convolutional layer.

Subsequently, F_{GWAM} is input into both GGAM and CCAM for attention enhancement. For GGAM, we first compute the gradient map from the input image as

$$F_g = G\left(I_{LR}\right) = \|(I^x_{LR}(\cdot), I^y_{LR}(\cdot))\|_2, \tag{2}$$

with

$$I^x_{LR}(\cdot) = I_{LR}(x+1,) - I_{LR}(x-1,), \tag{3}$$

and

$$I^y_{LR}(\cdot) = I_{LR}(,y+1) - I_{LR}(,y-1), \tag{4}$$

where $G(\cdot)$ denotes the gradient calculation function, and F_g represents the obtained gradient map.

Next, F_{GWAM} is enhanced by GGAM, giving the output

$$F_{GGAM} = H_{GGAM}\left(F_{GWAM}, F_g\right), \tag{5}$$

where $H_{GGAM}(\cdot)$ denotes the gradient-based graph attention module.

For CCAM, we use the attention architecture from (Hu, Shen, and Sun 2018) to enhance features via cross-channel attention, yielding

$$F_{CCAM} = H_{CCAM}\left(F_{GWAM}\right). \tag{6}$$

We then merge F_{CCAM} and F_{GGAM} through concatenation and refine the concatenated representation via a convolutional layer and a residual block, giving

$$F_{cat} = ResBs\left(conv\left(H_{cat}\left(F_{CCAM}, F_{GGAM}\right)\right)\right) \tag{7}$$

as the output of this feature fusion.

Finally, we add F_{cat} and F_{GWAM} to obtain the super-resolution image

$$I_{SR} = H_{up}\left(F_{cat} + F_{GWAM}\right) \tag{8}$$

as final output, where $H_{up}(\cdot)$ denotes the up-sampling function.

Group-Wise Attention Module

High recurrence of small similar patterns in natural images is a favourable chacteristic for low-level image reconstruction tasks (Park et al. 2020). Consequently, non-local attention modules have been widely applied in image super-resolution approaches (Dai et al. 2019). However, this also leads to unrelated and noisy contents when calculating the similarity matrix across the whole image. Considering that enriched texture regions in scene text images are fairly sparse, we introduce a group-wise attention module to enhance pixel-level image representations. For this, we utilise the locality-sensitive hashing attention introduced in (Kitaev, Kaiser, and Levskaya 2019) for efficient computation of this component. Since locality-sensitive hashing is well suited to finding nearest neighbours in high-dimensional spaces, it can be employed to divide elements into M groups, with the

($K = \frac{N}{M}$) elements in each group exhibiting strong similarity. Once these partitions are generated on a spherical projected space, the elements with the K-th highest correlations in image feature space can be quickly located. These elements are then sorted for calculating similarities between them to enhance each other as

$$y_i = \frac{1}{C(x_i)} \sum_j f(x_i, x_j) g(x_j), \tag{9}$$

where $f(\cdot,\cdot)$ denotes the similarity calculation function, $C(\cdot)$ is a normalisation function, $g(\cdot)$ represents the feature representation function, and x_i is the feature vector for location i.

Gradient-Based Graph Attention Module

Regular structure and repetitive patterns are more frequently observed at edges and boundaries of characters in scene text images due to the well-structured layout. To exploit this, we propose GGAM as a novel module to enhance feature representations of these patches, supporting much clearer text contours and boundary recovery. As illustrated in Figure 2, we extract the gradient map F_g and divide it into small patches to construct the graph $G = (p, \varepsilon, A)$ with $p \in \mathbb{R}^{P\times H_p \times W_p}$ denoting a node, $P = |p|$ the number of graph nodes, $\varepsilon \subseteq p \times p$ representing the edges of the graph, and $A \in \mathbb{R}^{P\times P}$ denoting the adjacency matrix defining the edge weights. The number of nodes is decided by the patch size $H_p \times W_p$ whose impact we evaluate in our ablation study.

We build graph connections by computing the similarities between the nodes after a linear transformation as

$$A_{ij} = \langle \varphi(p_i), \phi(p_j) \rangle, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, p_i and p_j are the i-th and j-th node, and $\varphi(\cdot)$ and $\phi(\cdot)$ represent the learnable convolution operation.

Since some of the nodes, having low similarities to target node p_i, will mislead on how close different nodes in the graph are and introduce redundant information that can have a negative impact on the quality of graph reconstruction, we black out these insignificant nodes. For this, we first calculate the average similarity between each node with all its neighbours, and then use this to remove its linkage to those nodes that have different patterns. In particular, we obtain masks between the i-th node and other nodes as

$$M_i = \frac{\gamma(F_g)}{P} \sum_{j=1}^{P} A_{ij} + \psi(F_g), \tag{11}$$

where $\gamma(\cdot)$ and $\psi(\cdot)$ denote affine transformation functions which can embed deep features into specific transformation parameters. Next, we use the obtained mask to filter the previous adjacency matrix to remove those irrelevant nodes and arrive at the filtered adjacency matrix

$$W_{ij} = \begin{cases} A_{ij} & \text{if } A_{ij} \geq M_i \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

After obtaining the updated adjacency matrix, we normalise each row using the sotfmax function

$$S_{ij} = \frac{\exp(W_{ij})}{\sum_{j=1}^{N} \exp(W_{ij})}, \tag{13}$$

where $\exp(\cdot)$ is the exponential function, and S_{ij} is the edge weight between node p_i and node p_j after normalisation.

We then use the obtained adjacency matrix to update the interior of the node as

$$p_i' = \langle p_i, S_{i,\cdot} \rangle. \tag{14}$$

Next, we reshape each node back to the corresponding image patch and obtain the output. To stabilise the training process, we build multiple graph networks in parallel. These networks concurrently update multi-node information but do not share parameters with each other. At the same time, to reduce the computational burden, we reduce the channels of the input features when building a graph network. Finally, we concatenate the updated features of each graph network to obtain the output F_{GGAM}.

Loss Function

We apply two loss terms in our loss function, which is composed of a classical reconstruction loss term and a character perceptual loss term. To minimise the pixel difference between the SR and HR images, the reconstruction loss $\mathcal{L}_{rec}$ is obtained as

$$\mathcal{L}_{rec} = \|I_{SR} - I_{HR}\|_2^2, \tag{15}$$

where I_{SR} and I_{HR} are the generated SR image and the HR image, respectively.

The character perceptual loss $\mathcal{L}_{cha}$ is used to enhance the perceptual quality of text characters. Following (Chen, Li, and Xue 2021), we make use of a transformer recognition model trained on the Syn90k (Jaderberg et al. 2016) and SynthText (Gupta, Vedaldi, and Zisserman 2016) datasets to obtain attention maps of the HR and SR images and calculate the loss as

$$\mathcal{L}_{cha} = \tau \sum_{i=1}^{H} \sum_{j=1}^{W} \left| M_{SR}^{ij} - M_{HR}^{ij} \right| + \psi \sum \ln A, \tag{16}$$

where M_{SR} and M_{HR} are the feature maps of SR and HR images from the transformer model, and A denotes the weighted activation.

The overall loss is then calculated as

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{cha} \mathcal{L}_{cha}, \tag{17}$$

where λ_{cha} allows balancing between the two loss terms.

Experimental Results

Dataset

In our experiments, we use the benchmark TextZoom dataset (Wang et al. 2020) to evaluate the performance of our proposed TSAN model and to compare it to the state-of-the-art. TextZoom is composed of two real text image

	CRNN				MORAN				ASTER			
	easy	medium	hard	all	easy	medium	hard	all	easy	medium	hard	all
Bicubic	36.4	21.1	21.1	26.8	60.6	37.9	30.8	44.1	64.7	42.4	31.2	47.2
SRCNN$_{\text{TPAMI2015}}$	38.7	21.6	20.9	27.7	63.2	39.0	30.2	45.3	69.4	43.4	32.2	49.5
VDSR$_{\text{CVPR2016}}$	41.2	24.6	23.3	30.7	62.3	42.5	30.5	46.1	71.7	43.5	34.0	51.0
SRResNet$_{\text{CVPR2017}}$	45.2	32.6	25.5	35.1	66.0	47.1	33.4	49.9	69.4	50.5	35.7	53.0
LapSRN$_{\text{CVPR2017}}$	46.1	27.9	23.6	33.3	64.6	44.9	32.2	48.3	71.5	48.6	35.2	53.0
EDSR$_{\text{CVPRW2017}}$	42.7	29.3	24.1	32.7	63.6	45.4	32.2	48.1	72.3	48.6	34.3	53.0
RDN$_{\text{CVPR2018}}$	41.6	24.4	23.5	30.5	61.7	42.0	31.6	46.1	70.0	47.0	34.0	51.5
RRDB$_{\text{ECCV2018}}$	40.6	22.1	21.9	28.9	63.9	41.0	30.8	46.3	70.9	44.4	32.5	50.6
RCAN$_{\text{ECCV2018}}$	46.8	27.9	26.5	34.5	63.1	42.9	33.6	47.5	67.3	46.6	35.1	50.7
CARN$_{\text{ECCV2018}}$	40.7	27.4	24.3	31.4	58.8	42.3	31.1	45.0	62.3	44.7	31.5	47.1
SAN$_{\text{CVPR2019}}$	50.1	31.2	28.1	37.2	65.6	44.4	35.2	49.4	68.1	48.7	36.2	52.0
SRFBN$_{\text{CVPR2019}}$	44.5	30.7	25.8	34.3	59.9	44.0	33.0	46.5	63.5	47.5	34.3	49.4
IMDN$_{\text{MM2019}}$	45.1	29.5	25.2	34.0	60.7	43.6	34.3	47.1	65.8	49.0	34.3	50.7
HAN$_{\text{ECCV2020}}$	51.6	35.8	29.0	39.6	67.4	48.5	35.4	51.5	71.1	52.8	39.0	55.3
TSRN$_{\text{ECCV2020}}$	52.5	38.2	31.4	41.4	70.1	55.3	37.9	55.4	75.1	56.3	40.1	58.3
TBSRN$_{\text{CVPR2021}}$	59.6	47.1	35.3	48.1	74.1	57.0	40.8	58.4	75.7	59.9	41.6	60.1
PCAN$_{\text{MM2021}}$	59.6	45.4	34.8	47.4	73.7	57.6	41.0	58.5	77.5	60.7	43.1	61.5
TPGSR$_{\text{arXiv2021}}$	63.1	52.0	38.6	51.8	74.9	60.5	44.1	60.5	78.9	62.7	44.5	62.8
Gestalt$_{\text{AAAI2022}}$	61.2	47.6	35.5	48.9	75.8	57.8	41.4	59.4	77.9	60.2	42.4	61.3
TATT$_{\text{CVPR2022}}$	62.6	**53.4**	**39.8**	52.6	72.5	60.2	43.1	59.5	78.9	63.4	**45.4**	63.6
Ours	**64.6**	53.3	38.8	**53.0**	**78.4**	**61.3**	**45.1**	**62.7**	**79.6**	**64.1**	45.3	**64.1**

Table 1: Recognition accuracy results (in %) on TextZoom for all methods and using the three recognition models, CRNN, MORAN, and ASTER. Results are given for the three defined groups (east, medium, hard) as well as overall results on the complete test dataset. The best result for each group is bolded.

super-resolution datasets, RealSR (Cai et al. 2019) and SR-RAW (Zhang et al. 2019), in total containing 21,740 image pairs where the size of the low-resolution and high-resolution images are 16×64 and 32×128, respectively. Of these, 17,367 image pairs are used for training, and the remaining 4,373 pairs for testing which are further divided (based on different focal lengths) into 1,619 easy, 1,411 medium, and 1,343 hard pairs.

Evaluation Measures

Since the purpose of STISR is to improve scene text recognition, we take the recognition accuracies of three SOTA scene text recognition models on the reconstructed scene text images as the main evaluation measure. In particular, we use the official released code and pre-trained models of CRNN (Shi, Bai, and Yao 2016), ASTER (Shi et al. 2019) and MORAN (Luo, Jin, and Sun 2019) for this purpose. In addition, we also evaluate the reconstructed image quality in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

Implementation Details

We adopt the Adam optimiser (Kingma and Ba 2014) with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and train with a batch size of 16. We set the learning rate to 0.0001, and train the model for 300 epochs. All experiments are run, using PyTorch, on NVIDIA GeForce RTX 2080Ti GPUs and an Intel i9-10940X processor @3.30 GHz.

Results and Comparison with SOTA

To validate the effectiveness of our TSAN model, we compare our results to, apart from standard bicubic interpolation as a baseline, 13 state-of-the-art single image super-resolution methods, namely SRCNN (Dong et al. 2015a), VDSR (Kim, Lee, and Lee 2016), SRResNet (Huang et al. 2017), LapSRN (Lai et al. 2017), EDSR (Lim et al. 2017), RDN (Zhang et al. 2018b), RRDB (Wang et al. 2018b), RCAN (Zhang et al. 2018a), CARN (Ahn, Kang, and Sohn 2018), SAN (Dai et al. 2019), SRFBN (Li et al. 2019), IMDN (Hui et al. 2019), and HAN (Niu et al. 2020), and 6 specialised scene text image super-resolution methods, namely TSRN (Wang et al. 2020), TBSRN (Chen, Li, and Xue 2021), PCAN (Zhao et al. 2021), TPGSR (Ma, Guo, and Zhang 2021), Gestalt (Chen et al. 2022), and TATT (Ma, Liang, and Zhang 2022). To allow a fair comparison, we use the codes released by the authors and train all models with the same configuration. Following our ablation studies, we set $M = 8$, $\lambda_{cha} = 0.5$, and use a patch size of 7×7.

The obtained text recognition results, for all super-resolution models and all three text recognition approaches, are given in Table 1. From there, we can see that, as expected, while generic super-resolution approaches give a significant boost compared to standard bicubic interpolation, the resulting recognition accuracy is still relatively modest. Out of these approaches, HAN performs best, giving an average recognition accuracy of 39.6%, 51.5%, and 55.3% based on CRNN, MORAN, and ASTER, respectively.

STISR methods outperform generic SISR methods by a significant margin, with the worst performing STISR model

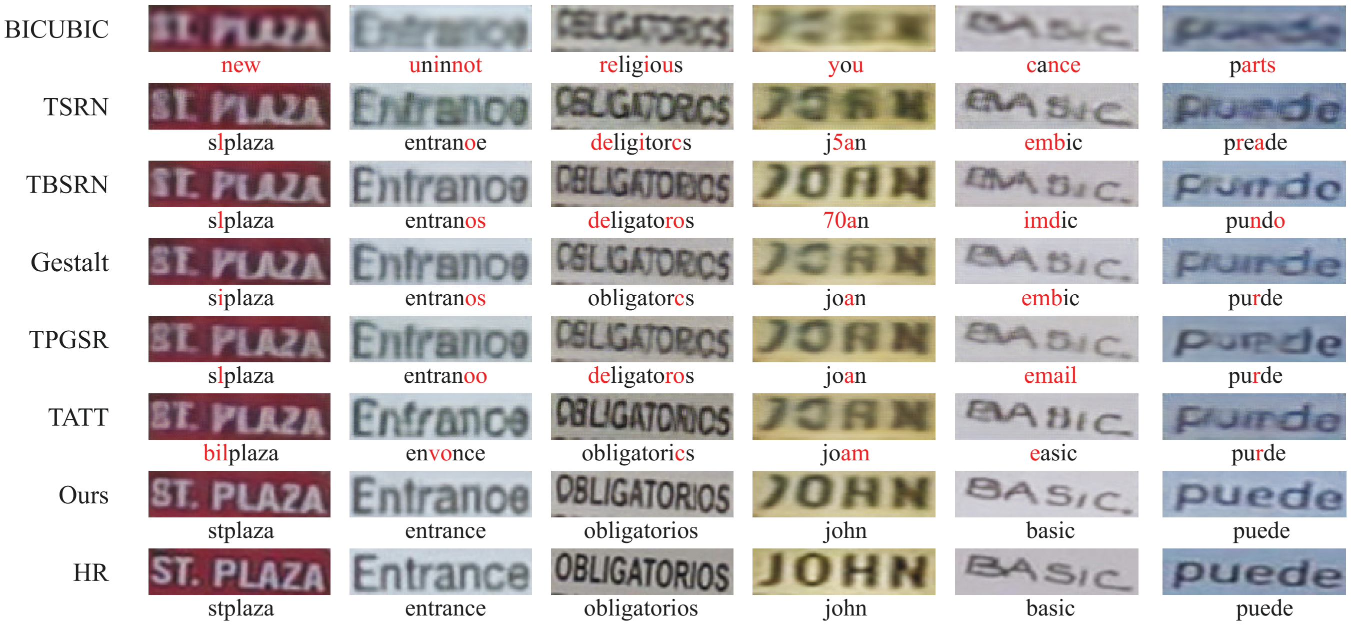


Figure 3: Examples of super-resolution images obtained by all STISR methods and the resulting text recognition results based on ASTER. Incorrectly recognised characters are indicated in red.

yielding better recognition performance than the best SISR approach. The best performance here is achieved by TATT with an average recognition accuracy of 52.6%, 59.5% and 63.6% using CRNN, MORAN and ASTER, respectively.

Our proposed scene text image super-resolution approach is able to even better the performance of TATT, yielding impressive recognition performance. In particular, we obtain recognition accuracies of 53.0%, 62.7% and 64.1% based on CRNN, MORAN and ASTER, respectively, conclusively outperforming all other models and thus the current state-of-the-art in STISR while also giving the best results for the majority of the sub-groups of the TextZoom test dataset.

Some examples of this performance are given in Figure 3, where we show both the visual outputs of the various STISR methods and the consequent recognition results obtained by ASTER (more results are given in the supplementary material). As we can see, our TSAN approach yields high-quality super-resolution images that exhibit clearer text compared to the other methods and in turn allow perfect text recognition on these examples while all other approaches lead to incorrectly recognised characters.

Finally, in Table 2 we list the PSNR and SSIM results obtained on the test datasets by all methods. As is evident from there, our approach yields SOTA SR image quality.

Ablation Study

We perform a series of ablation experiments to verify the effectiveness of each component of our model. In the following, we present the ablations of each attention module and two important parameters, the number of groups in GWAM and the patch size in GGAM module, while further ablation results are given in the supplementary material. Due to space restrictions we use ASTER as the recognition model here,

	easy	medium	hard
Bicubic	22.35/0.7884	18.98/0.6254	19.39/0.6592
SRCNN	23.48/0.8379	19.06/0.6323	19.34/0.6791
VDSR	24.62/0.8631	18.96/0.6166	19.79/0.6989
SRResNet	24.36/0.8681	18.88/0.6406	19.29/0.6911
LapSRN	24.58/0.8556	18.85/0.6480	19.77/0.7087
EDSR	24.26/0.8633	18.63/0.6440	19.14/0.7108
RDN	22.27/0.8249	18.95/0.6427	19.70/0.7113
RRDB	22.12/0.8351	18.35/0.6194	19.15/0.6856
RCAN	22.15/0.8525	18.81/0.6465	19.83/0.7227
CARN	22.70/0.8384	19.15/0.6412	20.02/0.7172
SAN	22.69/0.8597	18.77/0.6477	19.82/0.7280
SRFBN	22.95/0.8465	19.13/0.6469	20.09/0.7232
IMDN	23.31/0.8536	19.17/0.6457	20.03/0.7258
HAN	23.30/0.8691	19.02/0.6537	20.16/0.7387
TSRN	22.95/0.8562	19.26/0.6596	19.76/0.7285
TBSRN	24.13/0.8729	19.08/0.6455	20.09/0.7452
PCAN	24.57/0.8830	19.14/0.6781	20.26/0.7475
TPGSR	24.35/0.8860	18.73/0.6784	19.93/0.7507
Gestalt	23.95/0.8611	18.58/0.6621	19.74/0.7520
TATT	24.72/**0.9006**	19.02/**0.6911**	20.31/**0.7703**
Ours	**25.34**/0.8920	**19.72**/0.6809	**20.69**/0.7606

Table 2: Image quality results on TextZoom for all methods, reported in terms of PSNR/SSIM. The best result for each group/measure is bolded.

GWAM		✓		✓	✓		✓	✓
CCAM	✓	✓	✓		✓	✓		✓
Gradient map						✓	✓	✓
GGAM			✓	✓	✓	✓	✓	✓
accuracy	61.8	62.5	62.7	62.9	63.0	63.5	63.7	64.1

Table 3: Ablation study results, reporting ASTER recognition accuracies (in %) of the proposed full model in comparison to results obtained when turning off the various components of the model.

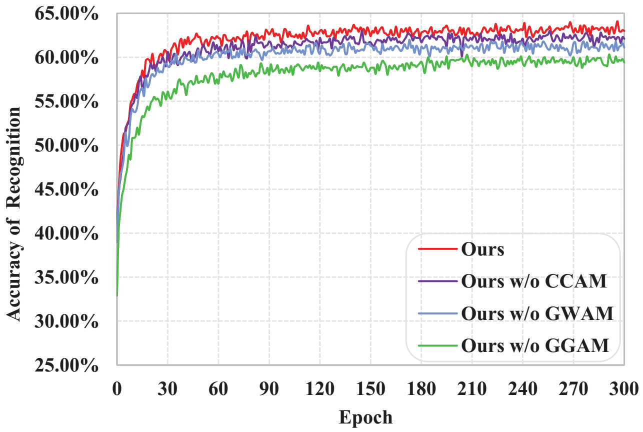


Figure 4: Training convergence curves (using ASTER) for different variants of our model.

while ablation results for CRNN and MORAN are provided in the supplementary material.

Table 3 reports the recognition accuracies obtained when turning on/off the various components that we introduce in our model, while Figure 4 shows the obtained convergence curves of the full model in comparison to turning off the various attention components.

GWAM We introduce the GWAM to embed global salient context information so as to yield enhanced texture representations. As we can see from Table 3, this does indeed lead to improved recognition performance. Removing GWAM from the full model results in a 0.6% performance drop, while similar differences are also observed for the other variants with/without GWAM.

We also conduct an experiment with regards to the grouping strategy, evaluation the recognition performance for different values of M, the number of groups used. When $M = 1$, GWAM reverts to a non-local attention module (Wang et al. 2018a), while we evaluate group sizes of up to 16. The results, given in Table 4, show that the best results are obtained for $M = 2^3 = 8$.

no. of groups	easy	medium	hard	all
$M = 2^0$	77.8	63.1	43.9	62.6
$M = 2^1$	78.7	63.5	44.6	63.3
$M = 2^2$	78.4	63.8	45.1	63.5
$M = 2^3$	**79.6**	**64.1**	**45.2**	**64.1**
$M = 2^4$	78.1	63.3	44.5	63.0

Table 4: Ablation results (in %) for different values of M, the number of groups in GWAM.

patch size	easy	medium	hard	all
1 × 1	76.5	62.1	41.8	61.2
3 × 3	78.9	63.5	42.9	62.9
5 × 5	**79.8**	63.1	43.9	63.4
7 × 7	79.6	**64.1**	**45.2**	**64.1**
9 × 9	78.2	63.1	44.1	62.9
11 × 11	78.1	62.2	44.1	62.5

Table 5: Ablation results (in %) for different patch sizes in GGAM.

CCAM The CCAM is used to further exploit attentions across channels for feature enhancement. As we can conclude from Table 3, this leads to an effective improvement in the quality of the reconstructed image and consequently improved recognition accuracy, boosting the latter by 0.4% for the full model compared to one without CCAM.

GGAM Our proposed GGAM has the most significant impact on the performance, making it the most important component in our TSAN model. As is apparent from Table 3, introducing the module leads to an impressive performance gain of 1.6%. Furthermore, we can see that the use of the gradient map in GGAM is important to reach this performance.

Since the patch size of the graph representations decides how the local regional information is embedded via the graph attention mechanism, we also evaluate different patch sizes, ranging from 1 × 1 to 11 × 11 and report the obtained results in Table 5. We find that a patch size of 7 × 7 gives the highest recognition accuracy and thus performs best to capture the local stroke-level patterns.

Conclusions

In this paper, we have proposed a novel two-stage text structure attention embedding method exploiting the well-structured character layout and repetitive local patterns for scene text image super-resolution in the wild. By leveraging salient and local regional text structure into the feature representations, we achieve significantly improved high resolution text image reconstruction and consequently improved performance for the downstream text recognition task. In our future work, we will apply our model to real-world applications with a focus on improving inference efficiency.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 62076255; Open Research Projects of Zhejiang Lab (No. 2022RC0AB07); Hunan Provincial Science and Technology Plan Project 2020SK2059; Key projects of Hunan Education Department 20A88; National Science Foundation of Hunan Province 2021JJ30082.

References

Ahn, N.; Kang, B.; and Sohn, K.-A. 2018. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision*, 252–268.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 3086–3095.

Chen, J.; Li, B.; and Xue, X. 2021. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Chen, J.; Yu, H.; Ma, J.; Li, B.; and Xue, X. 2022. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 285–293.

Cheng, Z.; Xu, Y.; Bai, F.; Niu, Y.; Pu, S.; and Zhou, S. 2018. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5571–5579.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015a. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 38(2): 295–307.

Dong, C.; Zhu, X.; Deng, Y.; Loy, C. C.; and Qiao, Y. 2015b. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211*.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2315–2324.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight Image Super-Resolution with Information Multi-distillation Network. In *Proceedings of the ACM International Conference on Multimedia*, 2024–2032.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1): 1–20.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1646–1654.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitaev, N.; Kaiser, L.; and Levskaya, A. 2019. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 624–632.

Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; and Wu, W. 2019. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3867–3876.

Liao, M.; Zhang, J.; Wan, Z.; Xie, F.; Liang, J.; Lyu, P.; Yao, C.; and Bai, X. 2019. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8714–8721.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144.

Liu, W.; Chen, C.; and Wong, K.-Y. 2018. Char-net: A character-aware neural network for distorted scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Luo, C.; Jin, L.; and Sun, Z. 2019. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90: 109–118.

Ma, J.; Guo, S.; and Zhang, L. 2021. Text prior guided scene text image super-resolution. *arXiv preprint arXiv:2106.15368*.

Ma, J.; Liang, Z.; and Zhang, L. 2022. A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5911–5920.

Mou, Y.; Tan, L.; Yang, H.; Chen, J.; Liu, L.; Yan, R.; and Huang, Y. 2020. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *Proceedings of the European Conference on Computer Vision*, 158–174.

Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; and Shen, H. 2020. Single image super-resolution via a holistic attention network. In *Proceedings of the European Conference on Computer Vision*, 191–207.

Park, S.; Yoo, J.; Cho, D.; Kim, J.; and Kim, T. H. 2020. Fast adaptation to super-resolution networks via meta-learning. In *Proceedings of the European Conference on Computer Vision*, 754–769.

Shi, B.; Bai, X.; and Yao, C. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11): 2298–2304.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2019. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2035–2048.

Wang, W.; Xie, E.; Liu, X.; Wang, W.; Liang, D.; Shen, C.; and Bai, X. 2020. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision*, 650–666.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018a. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018b. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*.

Wu, Y.-C.; Yin, F.; and Liu, C.-L. 2017. Improving handwritten Chinese text recognition using neural network language models and convolutional neural network shape models. *Pattern Recognition*, 65: 251–264.

Zhang, X.; Chen, Q.; Ng, R.; and Koltun, V. 2019. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3762–3770.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018a. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018b. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2472–2481.

Zhao, C.; Feng, S.; Zhao, B. N.; Ding, Z.; Wu, J.; Shen, F.; and Shen, H. T. 2021. Scene text image super-resolution via parallelly contextual attention network. In *Proceedings of the ACM International Conference on Multimedia*, 2908–2917.