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

Xiangyuan Zhu¹, Kehua Guo¹*, Hui Fang², Rui Ding¹, Zheng Wu¹, Gerald Schaefer²
¹School of Computer Science and Engineering, Central South University, China
²Department 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.

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...
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 perceptual 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
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 focuses 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 input image $I_{LR}$, we utilise a convolutional layer to enhance salient feature extraction. This feature representation is then fed into GWAM for feature fusion. For CCAM, we use the attention architecture from (Hu, Shen, and Sun 2018) to enhance features via cross-channel attention, yielding

$$F_{GWAM} = H_{GWAM}(conv(I_{LR})), \quad (1)$$

where $H_{GWAM}()$ denotes the GWAM module and $conv()$ 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

$$F_g = G(I_{LR}) = \| (I_{LR}(x+1), I_{LR}(x-1)) \|_2, \quad (2)$$

with

$$I_{LR}(x+1) = I_{LR}(x+1) - I_{LR}(x-1), \quad (3)$$

and

$$I_{LR}(x-1) = I_{LR}(x+1) - I_{LR}(x-1), \quad (4)$$

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

$$F_{GGAM} = H_{GGAM}(F_{GWAM}, F_g), \quad (5)$$

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

$$F_{CCAM} = H_{CCAM}(F_{GWAM}). \quad (6)$$

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

$$F_{cat} = ResBs(\text{conv}(H_{cat}(F_{CCAM}, F_{GGAM}))) \quad (7)$$

as the output of this feature fusion.

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

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

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 characteristic 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
After obtaining the updated adjacency matrix, we normalise each row using the softmax function
\[
S_{ij} = \frac{\exp(W_{ij})}{\sum_{j=1}^{N} \exp(W_{ij})},
\]
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
\[
p_i' = (p_i, S_{i,}).
\]

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 \(L_{rec}\) is obtained as
\[
L_{rec} = \|I_{SR} - I_{HR}\|^2_2,
\]
where \(I_{SR}\) and \(I_{HR}\) are the generated SR image and the HR image, respectively.

The character perceptual loss \(L_{cha}\) 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
\[
L_{cha} = \tau \sum_{i=1}^{H} \sum_{j=1}^{W} \left| M_{SR}^{ij} - M_{HR}^{ij} \right| + \psi \sum A \ln A,
\]
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
\[
\mathcal{L} = L_{rec} + \lambda_{cha} L_{cha},
\]
where \(\lambda_{cha}\) 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
super-resolution datasets, RealSR (Cai et al. 2019) and SRRAW (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), SRRNet (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 (Ni 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
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>easy</th>
<th>medium</th>
<th>hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>22.35/0.7884</td>
<td>18.98/0.6254</td>
<td>19.39/0.6592</td>
</tr>
<tr>
<td>SRCNN</td>
<td>23.48/0.8379</td>
<td>19.06/0.6323</td>
<td>19.34/0.6791</td>
</tr>
<tr>
<td>VDSR</td>
<td>24.62/0.8631</td>
<td>18.96/0.6166</td>
<td>19.79/0.6989</td>
</tr>
<tr>
<td>SRResNet</td>
<td>24.36/0.8681</td>
<td>18.88/0.6406</td>
<td>19.29/0.6911</td>
</tr>
<tr>
<td>LapSRN</td>
<td>24.58/0.8556</td>
<td>18.85/0.6480</td>
<td>19.77/0.7087</td>
</tr>
<tr>
<td>EDSR</td>
<td>24.26/0.8633</td>
<td>18.63/0.6440</td>
<td>19.14/0.7108</td>
</tr>
<tr>
<td>RDN</td>
<td>22.27/0.8249</td>
<td>18.95/0.6427</td>
<td>19.70/0.7113</td>
</tr>
<tr>
<td>RRDB</td>
<td>22.12/0.8351</td>
<td>18.35/0.6194</td>
<td>19.15/0.6856</td>
</tr>
<tr>
<td>RCAN</td>
<td>22.15/0.8525</td>
<td>18.81/0.6465</td>
<td>19.83/0.7227</td>
</tr>
<tr>
<td>CARN</td>
<td>22.70/0.8384</td>
<td>19.15/0.6412</td>
<td>20.02/0.7172</td>
</tr>
<tr>
<td>SAN</td>
<td>22.69/0.8597</td>
<td>18.77/0.6477</td>
<td>19.82/0.7280</td>
</tr>
<tr>
<td>SRFBN</td>
<td>22.95/0.8465</td>
<td>19.13/0.6469</td>
<td>20.09/0.7232</td>
</tr>
<tr>
<td>IMDN</td>
<td>23.31/0.8536</td>
<td>19.17/0.6457</td>
<td>20.03/0.7258</td>
</tr>
<tr>
<td>HAN</td>
<td>23.30/0.8691</td>
<td>19.02/0.6537</td>
<td>20.16/0.7387</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>GWAM</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
<th>✓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gradient map</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GGAM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>accuracy</td>
<td>61.8</td>
<td>62.5</td>
<td>62.7</td>
<td>62.9</td>
<td>63.0</td>
<td>63.5</td>
<td>63.7</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>no. of groups</th>
<th>easy</th>
<th>medium</th>
<th>hard</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = 2^3$</td>
<td>77.8</td>
<td>63.1</td>
<td>43.9</td>
<td>62.6</td>
</tr>
<tr>
<td>$M = 2^1$</td>
<td>78.7</td>
<td>63.5</td>
<td>44.6</td>
<td>63.3</td>
</tr>
<tr>
<td>$M = 2^2$</td>
<td>78.4</td>
<td>63.8</td>
<td>45.1</td>
<td>63.5</td>
</tr>
<tr>
<td>$M = 2^3$</td>
<td>79.6</td>
<td>64.1</td>
<td>45.2</td>
<td>64.1</td>
</tr>
<tr>
<td>$M = 2^4$</td>
<td>78.1</td>
<td>63.3</td>
<td>44.5</td>
<td>63.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>patch size</th>
<th>easy</th>
<th>medium</th>
<th>hard</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 1</td>
<td>76.5</td>
<td>62.1</td>
<td>41.8</td>
<td>61.2</td>
</tr>
<tr>
<td>3 × 3</td>
<td>78.9</td>
<td>63.5</td>
<td>42.9</td>
<td>62.9</td>
</tr>
<tr>
<td>5 × 5</td>
<td>79.8</td>
<td>63.1</td>
<td>43.9</td>
<td>63.4</td>
</tr>
<tr>
<td>7 × 7</td>
<td>79.6</td>
<td>64.1</td>
<td>45.2</td>
<td>64.1</td>
</tr>
<tr>
<td>9 × 9</td>
<td>78.2</td>
<td>63.1</td>
<td>44.1</td>
<td>62.9</td>
</tr>
<tr>
<td>11 × 11</td>
<td>78.1</td>
<td>62.2</td>
<td>44.1</td>
<td>62.5</td>
</tr>
</tbody>
</table>

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 $M = 2^3 = 8$.

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 $1 \times 1$ to $11 \times 11$ and report the obtained results in Table 5. We find that a patch size of $7 \times 7$ 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 20A488; National Science Foundation of Hunan Province 2021JJ30082.

References


