Text-Enhanced Scene Image Super-Resolution via Stroke Mask and Orthogonal Attention

Rui Shu, Cairong Zhao, Shuyang Feng, Liang Zhu, and Duoqian Miao

***Abstract*— Low-resolution text images are very commonplace in real life and their information is hard to be extracted by using existing text recognition methods only. Although this problem can be solved by introducing super-resolution (SR) techniques, most existing SR methods fail to process stroke regions and background regions of input text images distinctively. In this paper, we propose a text-specific super-resolution network named Text Enhanced Attention Network (TEAN) to solve this problem. First of all, we compensate for disadvantages of traditional thresholding mask operation proposed in Text Super-Resolution Network (TSRN) by utilizing deep-learning based semantic segmentation method to get correct masks as prior semantic information and propose a Text-Segmented-Contextual-Attention (TSCA) branch on the basis of them. Besides, we design an Orthogonal Contextual Attention Module (OCAM) working with TSCA to implicitly enhance stroke regions of LR images. Secondly, to effectively fuse shallow features and deep features of SR model, we propose a convolutional structure named Weight Balanced Fusion Module (WBFM) to improve traditional feature fusion methods of SR network. Finally, extensive experiments on TextZoom dataset demonstrate that the proposed network can improve the recognition accuracy of text images on existing text recognition models. Using TEAN to process low-resolution text images improves the recognition accuracy by 25.4% on CRNN, by 17.4% on ASTER, by 17.3% on MORAN, by 20.7% on NRTR, by 17.3% on SAR and by 15.9% on MASTER compared with directly recognizing them, which attains competitive performances against state-of-the-art methods. Furthermore, cross-dataset experiments on IC15_2077 demonstrate that TEAN is helpful for scene text recognition task, especially for low-resolution images even with the cross-domain issue.**

***Index Terms*— Scene text super-resolution, stroke regions, attention, scene text recognition.**

I. Introduction

TRADITIONAL text recognition methods such as [1] and [2] use predefined formulas to deal with sequence recognition issues so that they don't generalize and perform well. In recent years, with the rapid development of deep learning techniques, deep-learning based text recognition models such as CTC-based models [3] and attention-based models [4], [5], [6] replace traditional text recognition methods and achieve excellent performance. Deep-learning based text recognition are also widely used in many other fields such as license plate recognition, documents retrieval and auto-driving. However, in real-life scenarios, many blurred text images always get in the way of correct text recognition [7]. These text images are often caused by optical degradation and noise, which is unavoidable. Existing text recognition models [3], [4], [5], [8], [9], [10] have difficulty in processing them correctly, which remains a challenge to be solved.

Recently, the issue of optical degradation has attracted a great deal of attention so that many deep-learning based super-resolution methods have been developed to alleviate this issue. [11] proposed SRCNN and adapted it into text SR in ICDAR 2015 TextSR competition [12]. However, these models are not text-oriented and trained on synthetic datasets [12] obtained by BICUBIC, Gaussian blurring, etc, which means that they are not feasible to be used directly in real-life scenarios. In other words, these networks learn the inverse mapping algorithm of interpolation rather than how to super-resolve LR images. To solve this problem, [13] first proposed a super-resolution dataset based on real-life scenarios named TextZoom and a network named Text Super-Resolution Network (TSRN) trained on this dataset. TextZoom includes cropped text images pairs of low-resolution and high-resolution. In real-world applications, we firstly localize texts in images by means of text detection methods and crop these text images according to the detected bounding boxes. Based on this work, [14] proposed a Parallelly Contextual Attention Network (PCAN) to effectively learn sequence-dependent features and model high-frequency information of the reconstruction in text images.

Most existing text SR methods [13], [14] improve their model by utilizing contextual information, which has been proven to be effective. However, majority of these methods have a problem that they fail to process stroke regions and background regions of text images distinctively but rather treat them equally, which introduces some background information that is not relevant to the texts into these networks. Although [13], [15] make use of segmentation information [16] to process stroke regions and background regions distinctively, the masks generated by their traditional thresholding methods are not always precise. Actually, deep-learning based

Manuscript received 31 October 2022; revised 12 January 2023 and 16 February 2023; accepted 11 April 2023. Date of publication 14 April 2023; date of current version 30 October 2023. This work was supported in part by the National Natural Science Fund of China under Grant 62076184, Grant 61673299, Grant 61976160, and Grant 62076182; in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; in part by the Shanghai Natural Science Foundation of Shanghai under Grant 22ZR1466700; in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100; and in part by the Fundamental Research Funds for the Central Universities. This article was recommended by Associate Editor C. Yang. *(Rui Shu and Cairong Zhao contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

The authors are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: 2133056@tongji.edu.cn; zhaocairong@tongji.edu.cn; fengshuyang@tongji.edu.cn; 1941778@tongji.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3267133.

Digital Object Identifier 10.1109/TCSVT.2023.3267133

segmentation methods [17] can generate precise masks and compensate for disadvantages of traditional methods. These correct masks can be regarded as prior semantic information by channel-wise concatenation with original RGB images and used to enhance stroke regions by the proposed TSCA branch, which will be detailed in Section III. Besides, due to the lack of attentional guidance [18], [19], most existing text SR methods also focus on some low-frequency information incorrectly, i.e. the background regions, which results in the texts of some super-resolved images generated by these methods are not clear.

To deal with this problem, we propose a text-specific SR method, named Text Enhanced Attention Network (TEAN). First of all, inspired by [13], based on LapSRN [20], we construct a strong baseline by adding Thin-Plate-Spline (TPS) Align module and orthogonal Recurrent Neural Networks (RNNs) into it. The above modules are essential for text SR and it is also vital to conduct experiments based on a strong baseline to verify the effectiveness of the proposed methods. After that, to explicitly enhance the text regions, we not only make some calibration to the mask operation proposed in [13] and [15], but also further exploit the information of calibrated masks to assist the learning of the super-resolution network by designing a Text-Segmented-Contextual-Attention (TSCA) branch. This structure can explicitly assist the model to distinguish stroke regions and background regions. Besides, to suppress background regions and implicitly enhance stroke regions of text images, an Orthogonal Contextual Attention Module (OCAM) is designed to model high-frequency information horizontally and vertically. Finally, considering the complementarity and correlation between deep residual features and shallow upsampled features of LapSRN [20], a convolutional structure named Weight Balanced Fusion Module (WBFM) is proposed to balance the benefits of these two complementary information flows.

The contributions of this paper can be summarized as follows:

- To treat stroke regions and background regions distinctively, we design a Text-Segmented-Contextual-Attention branch to explicitly enhance stroke regions. Besides, we also design an Orthogonal Contextual Attention Module (OCAM) to model high-frequency information of two orthogonal directions, which can implicitly promote the model to distinguish stroke and background regions of text images. The above two structures can significantly improve the overall text perception ability of our model.
- To maximize the use of deep residual features and shallow upsampled features to construct final fused features, we design a convolutional structure named Weight Balanced Fusion Module (WBFM) to adaptively fuse them.
- Based on the above two points, we propose a text-specific network named Text Enhanced Attention Network (TEAN). The proposed TEAN focuses on enhancing stroke regions of low-resolution images and attains competitive performance against state-of-the-art methods on TextZoom dataset. Moreover, the proposed method achieves better cross-dataset performance than previous text-specific super-resolution methods, which verifies the improvement to scene text recognition brought by TEAN.

II. Related Work

A. Super-Resolution

Super-resolution [21], [22] [23] is a method of generating high-resolution images given low-resolution images, which has attracted a lot of interest in recent years. Traditional interpolation-based algorithms, such as bilinear and bicubic, their outputs are generated from the values of adjacent pixels according to the interpolation formulas. The quality of the images obtained by these methods have certain drawbacks, and with the development of deep learning and convolutional neural networks, super-resolution techniques [20], [24] [25] have also made great progress and are far superior in performance compared with traditional interpolation-based methods. In the era of deep learning, super-resolution networks are often trained based on LR-HR image pairs, which aims to make the super-resolved image as close as possible to the high-resolution image. However, these datasets used for training and evaluating are mostly generated by downsampling operations such as bicubic and bilinear, which do not fit in realistic scenarios. Therefore, previous networks trained on these datasets can not achieve competitive performance when applied to realistic scenarios. This also suggests that the network should be trained with different datasets for different application scenarios, which is an urgent problem that super-resolution methods need to address.

B. Text Recognition

Traditional text recognition methods adopt two manners, bottom-up manner, and top-down manner. Bottom-up manner is a method that detects and recognizes characters individually and then integrates them, while top-down manner is a method that regards text recognition as a classification problem. These methods [1], [2] used predefined formulas so they do not generalize and perform as well as deep-learning based methods.

With the development of deep learning and convolutional neural networks, existing text recognition methods [3], [4], [6], [26], [27], follow the top-down strategy. Based on how the loss function is calculated, these methods can be divided into two categories: CTC (Connectionist temporal classification)-based and attention-based. CRNN [3] is a representative kind of CTC-based method, which first introduced Recurrent Neural Network (RNN) and used CTC loss [28] to calculate the conditional probability between the predicted sequence and the label. Later, attention-based text recognition models such as ASTER [4] and MORAN [5] have made significant progress. They also introduced their respective text rectified modules to deal with distorted text images. Compared with CRNN [3], their performance have made a substantial increase. Referring to TSRN [13] and PCAN [14], we also select above three state-of-the-art text recognition algorithms, ASTER [4], MORAN [5] and CRNN [3] to evaluate the quality of our text SR method.

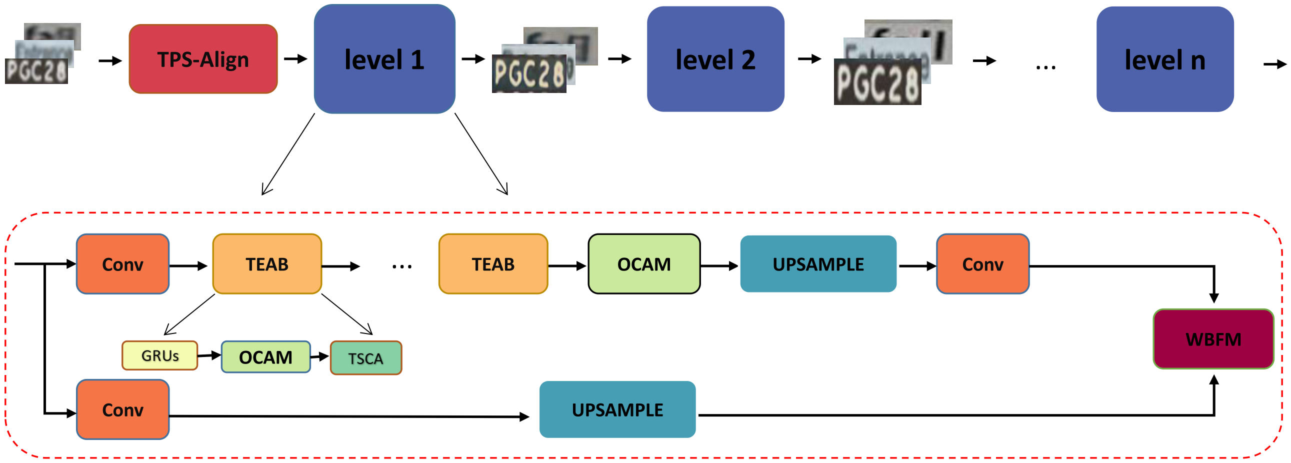

Fig. 1. Overall structure of TEAN and its detailed structure for each level. We obtain super-resolved images level by level and set five basic blocks (i.e. Text Enhanced Attention Block(TEAB) which is depicted in Fig. 5.) for each level.

C. Scene Text Image Super-Resolution

Scene text image super-resolution aims to improve the sharpness of the low-resolution text images so that the information can be extracted more easily. Mancas [29] listed several different artificial methods on text image SR and compared their performance. Pandey et al. [30] adopted a convolution-transposed convolution structure to cope with binary document SR and achieved good performance. Dong [11] proposed SRCNN and adapted it to text image SR in the ICDAR 2015 TextSR competition [12]. However, this model is very simple and not text-oriented. Moreover, as the network is trained on synthetic datasets, it will not perform well when applied to realistic scenarios.

Recently, [13] proposed a text SR dataset aiming at realistic scenarios named TextZoom. Besides, they also proposed an evaluation metric (the accuracy of existing text recognition methods on SR images) and designed a text-specific SR network TSRN [13] to improve the performance of text recognition methods on LR images. This work has played a crucial role in scene text image super-resolution and provided ideas and directions for subsequent researchers. On the basis of this work, [14] proposed a novel network named PCAN to effectively learn sequence-dependent features and focus more on high-frequency information of the reconstruction in text images. At the same time, [31] significantly improved performance by introducing information from the pre-trained transformer-based recognition model into TSRN [13]. As a result, performances of the above two works has improved considerably. Subsequently, [32] utilized text prior provided by CRNN [3] to guide the learning of super-resolution model and [33] embedded the text prior into super-resolution model by transformer-based structure [34], which enables better integration of image features and text prior.

III. The Proposed Model

In this section, we will present our proposed Text Enhanced Attention Network (TEAN) in detail. Firstly, the overall structure is described in Section A. Then we focus on overall text region perception methods of TEAN in Section B. The proposed Text-Segmented-Contextual-Attention (TSCA) and Orthogonal Contextual Attention Module (OCAM) will be detailed in this section. Then, the Weight Balanced Fusion Module (WBFM) is discussed in Section C. Finally, we focus on network optimization in Section D.

A. Overall Architecture

Our baseline is LapSRN [20], which means that we regard the image super-resolution as a progressive reconstruction process so that super-resolved images of the best quality can be gradually obtained level by level. Actually, this idea of progressive refinement is previously used in object detection field [35], [36]. Five basic blocks are also set for each level as the main structure referring to TSRN [13], PCAN [14] and SRResNet [24]. This is one of the structural differences between TEAN and baseline. The overall structure of TEAN and detailed structure for each level are shown in Fig.1.

First of all, LR inputs are rectified by TPS-align module, which was first introduced into scene text SR by TSRN [13]. This module was applied in text recognition models like ASTER [4] earlier and also used in subsequent scene text SR model [13], [14], [32] as an essential part. After that, stroke regions of feature maps are enhanced through stacked Text Enhanced Attention Blocks and Orthogonal Contextual Attention Module in each level. The Text Enhanced Attention Block consists of normal convolutional operations, two orthogonal RNNs, Orthogonal Contextual Attention module and Text-Segmented-Contextual Attention branch. In Fig.1 we omit normal convolution and its detailed structure is depicted in subsection B. Besides, residual features and upsampled features are adaptively fused by Weight Balanced Fusion Module (WBFM) to make full use of the similarity and complementarity between these two information flows. During the training stage, we use corresponding high-resolution images to supervise the output generated by each level to calculate the loss and assign different coefficients to these loss terms, which is detailed in section D.

B. Overall Text Region Perception

The purpose of text super-resolution is to make texts of images clearer, which requires the model to improve ability of

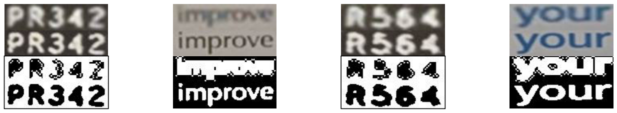

Fig. 2. Masks generated by calculating the average gray scale don't have a unified division of the stroke regions and the background regions.

Algorithm 1 Detailed Steps of the Proposed Mask Calibration Algorithm

```
Input: LR-HR pairs, pre-trained text segmentation model S(·);
Output: Calibrated masks.
  $binary\ mask \leftarrow S(HR)$
  if $mean\ grey\ scale\ of\ obtained\ mask > 35$ then
    Text of this mask is visible;
    Caculate the MIoU between original HR mask and mask generated by S(·).
    if $MIoU < 0.35$ then
      Original LR-HR masks pair is incorrect;
      Reverse LR-HR masks pair to get correct one as output.
    else
      Original LR-HR masks pair is correct which can be used as output directly.
    end if
  else
    Text of this mask is invisible, so we don't do any pre-process to this masks pair.
  end if
```

text region perception. In TEAN, we deal with this problem by TSCA and OCAM, which are illustrated in Fig. 4 (a) and Fig. 4 (b) respectively.

Mask Calibration and Text-Segmented-Contextual-Attention branch. On the one hand, previous works [11], [13], [14], [20], [24] fail to utilize segmentation information [37], [38], [39] for text SR effectively while there are many segmentation-based approaches widely used in scene text detection field [40], [41]. Although [13] proposed to generate binary masks by calculating the average gray scale of the RGB images, which are used to concatenate with text images as input. However, masks generated by this method do not have a unified division of stroke regions and background regions (some masks make the stroke regions render 1 and some masks make the stroke regions render 0) as shown in Fig. 2. If these inconsistent masks are concatenated with text images as input, ambiguous information will be introduced into the text SR model. By consulting previous works, we find masks generated by deep-learning-based segmentation methods do not suffer this problem. So inspired by [17], their pre-trained model texrnet-hrnet, which is the state-of-the-art method in text segmentation is used to process LR-HR pairs of TextZoom. As shown in Fig. 3 (a), masks generated by texrnet-hrnet handle stroke regions and background regions consistently (makes the stroke regions render 1 and the background regions render 0), although it performs even worse especially on LR images, which is reflected by the fact that no text is visible in the results. But this is due to cross-domain [42], [43] issues rather than the method itself.

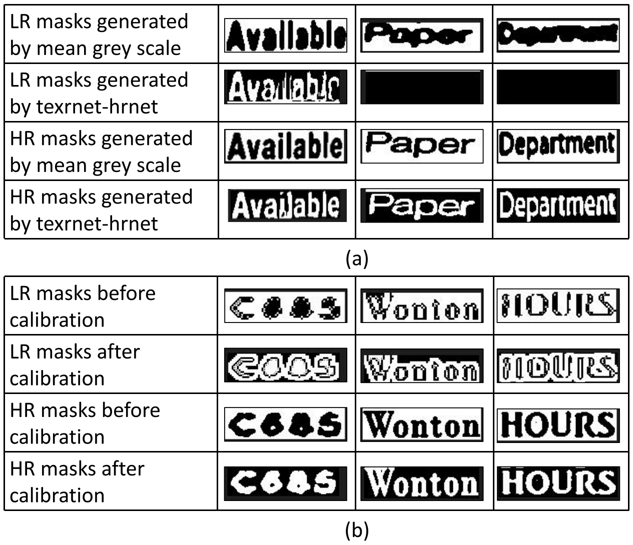
LR masks generated by mean grey scale
LR masks generated by texrnet-hrnet
HR masks generated by mean grey scale
HR masks generated by texrnet-hrnet
(a)
LR masks before calibration
LR masks after calibration
HR masks before calibration
HR masks after calibration
(b)

Fig. 3. In Fig. 3(a), we compare masks generated by traditional thresholding methods and masks generated by deep-learning based methods (texrnet-hrnet). In Fig. 3(b), some masks before and after calibration are visualized.

The most straightforward way to solve this problem is finetuning this model on TextZoom [13], but TextZoom [13] is a dataset aimed at text SR while texrnet-hrnet [17] is previously trained on TextSeg [17] dataset, which is aimed at semantic segmentation [37], [38], [39]. However, from another point of view, texrnet-hrnet [17] can be used to compensate for disadvantages of traditional thresholding segmentation methods. The pre-trained texrnet-hrnet model has a large number of parameters and heavy computational cost, due to the consideration of total training time and model size of text SR model, we obtain correct masks by pre-processing the train and test datasets of TextZoom [13]. The steps for pre-processing datasets are depicted in Alg. 1. Specifically, pre-trained segmentation model is only used to generated calibrated masks and not inclueded in the proposed text super-resolution model. The parameters of segmentation model are frozen during this process and no additional training is imposed on it. Under this algorithm, our statistics show that more than 6400 masks pairs of train datasets are reversed. In fact, there are still some masks that render the stroke regions as 0 and render the background regions as 1, and the outermost pixel points mostly have a value close to 255. So we can calculate the average of a circle of pixel points around these masks and select the masks whose pixel values are close to 255 for reversal. It can be said that with the above pre-processing steps we are able to get mostly correct masks. Some masks before and after calibration are visualized in Fig. 3 (b). Actually, the proposed Algorithm 1 can be used to generate calibrated masks of any cropped text images, so it can also be applied to other text recognition datasets for correct masks. In this paper, the algorithm is used to process text super-resolution dataset TextZoom [13] so that the training of super-resolution model can be guided.

Correct segmentation information not only can be used for text localization [40], but also it is helpful to suppress false positives [41]. Although utilizing text segmentation masks is common in scene text detection field, the semantic information provided by masks is not effectively exploited in text

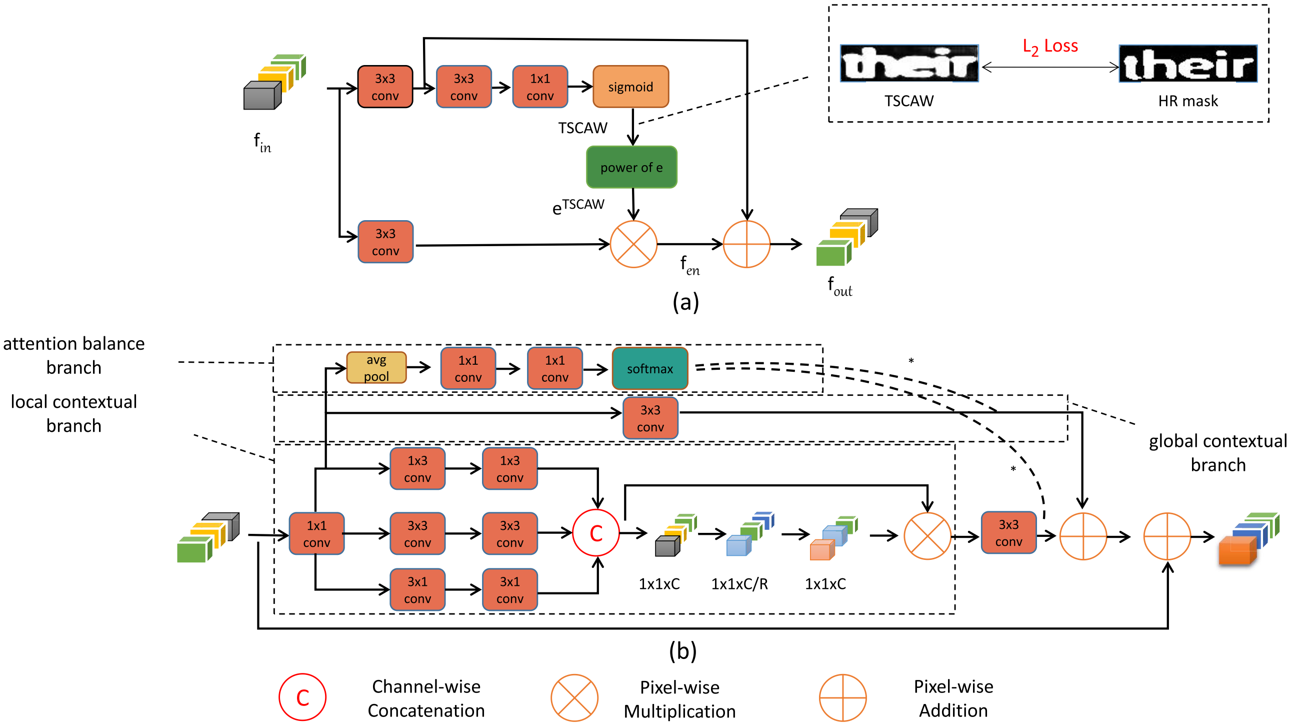

Fig. 4. Overall text region perception methods of TEAN. (a)The illustration of TSCA. (b)The illustration of OCAM.

super-resolution [13], [14], [24], [31], [32], [33], [44], [45]. Out of this consideration, we attempt to introduce information of text masks into text SR model and thus design a Text-Segmented-Contextual-Attention (TSCA) branch as shown in Fig. 4 (a) to explicitly improve the text region perception ability of our model. Specifically, three sequential convolutional operations are firstly used to extract the inner prior semantic information of input feature maps where the first two convolutional operations have a kernel size of 3 and the third convolutional operation has a kernel size of 1 and use Sigmoid function to get a Text-Segmented-Contextual-Attention-Weight (TSCAW), which can enhance the stroke regions of feature maps. The TSCAW with a value range 0 to 1 is supervised by the above calibrated binary masks of HR images during the training stage and e^{TSCAW} generated by power of e with a value range 1 to e is used to enhance the stroke regions of feature maps. In detail, with the supervision of calibrated HR masks, in the learned TSCAW, the value of stroke region tends to be larger, close to 1, and the value of background region tends to be smaller, close to 0. Therefore, calculated e^{TSCAW} is closer to e for stroke region and to 1 for background region. Utilizing it to multiply with the feature map can enhance the feature of background region less while significantly enhancing the feature of stroke region, thus prompting the model to focus more on stroke region. The output of the first convolutional operation is then added with the enhanced feature maps. Because output of the first convolutional operation captures much global contextual information [46], [47], it can be used to complement missing global contextual information of enhanced feature maps. The mathematical formulas for TSCA are expressed as follows:

$$TSCAW = s(h(g((f_{in})))) \tag{1}$$

$$f_{en} = e^{TSCAW} * g(f_{in}) \tag{2}$$

$$f_{out} = g(f_{in}) + f_{en} \tag{3}$$

where $g(\cdot)$ denotes convolutional operation whose kernel size is 3×3, $h(\cdot)$ denotes convolutional operation whose kernel size is 1×1, $s(\cdot)$ denotes Sigmoid function, f_{in} denotes input feature maps, f_{en} denotes enhanced feature maps and f_{out} denotes final output feature maps.

Orthogonal Contextual Attention Module. On the other hand, in the study of computer vision, attention [18], [48] has played an important role. The addition of attention mechanism implicitly promotes the model to focus on high-frequency information, i.e. text and to suppress unimportant information, i.e. background. In TEAN, stroke regions of feature maps are enhanced by TSCA branch, but TSCA can not suppress background regions because the e^{TSCAW} for TSCA takes value from 1 to e, so an Orthogonal Contextual Attention Module (OCAM) is designed to suppress unimportant background regions and refine these mixed features, whose structure is stated in Fig. 4 (b). In [49], orthogonal convolutional kernels are utilized to obtain horizontal and vertical heatmap for text detection, then final results are obtained by combining orthogonal heat maps to suppress false positives of two directions. This method achieves superior performances especially on texts of arbitrary shape. Similar to this work, 1 × 3 and 3 × 1 convolutional operations are used to model the vertical and horizontal high-frequency information respectively so that promotes the perception of arbitrarily shaped text features.

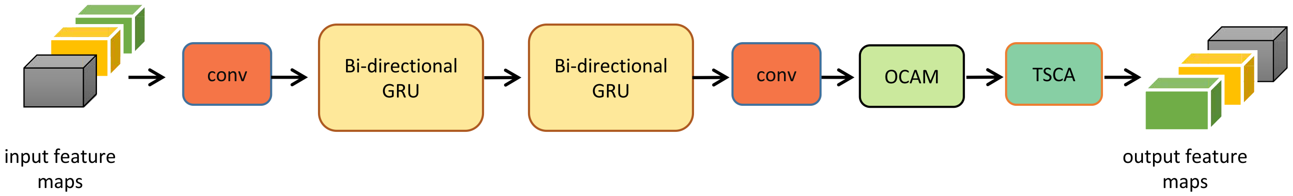

Fig. 5. The illustration of our proposed Text Enhanced Attention Block.

The specific settings of convolutional kernel size allow the model to suppress background regions of text images horizontally and vertically. However, this aim can only be achieved partly by these two operations, due to the limitations of their own receptive fields, a lot of deep local information [50] is ignored, so two normal 3 × 3 convolutions are also used to extract deep local information and features of these three flows are concatenated, which then be refined by channel-attention mechanism [18]. This forms the orthogonal local contextual branch of OCAM. But there is still some redundant and missing information in these features, inspired by [51], two branches are added, one global contextual branch which consists of a 3 × 3 convolutional operation to extract global information and the other attention balance branch which can calculate a pair of sum-to-one weights to balance the proportions [51], [52] between global and local information. The mathematical formula for the attention balance branch is expressed as follows:

$$[in\_we : 1 - in\_we] = s(g(g(avg\_pool(f)))) \tag{4}$$

where f denotes feature maps after the first 1×1 convolutional operation, $g(\cdot)$ denotes 1×1 convolutional operation, $s(\cdot)$ denotes Softmax function and $[in\_we : 1 - in\_we]$ denotes sum-to-one weights. This pair of weights is multiplied with output feature maps of global and local contextual branches and then sum them together. Finally, input feature maps are added to get the last output features. This three-branch structure forms the proposed Orthogonal Contextual Attention Module (OCAM).

The addition of the above two structures forms Text Enhanced Attention Block (TEAB), the basic block of TEAN as shown in Fig. 5. Similar to TSRN [13], we firstly introduce two sequential orthogonal RNNs and select GRU unit [53] as the basis of RNN. Then the proposed OCAM and TSCA are used to obtain stroke-enhanced feature maps, which have an overall perception of text regions.

C. Enhanced Feature Fusion

Feature fusion [54], [55] is an essential step in neural networks, which is usually achieved via direct channel-wise concatenation [56] or element-wise addition/multiplication [57] in previous works. In text SR field [13], [14], [24], shallow features and deep features are fused by element-wise addition. Although adding the above two features eases the training of overall network [58], it fails to focus on the core of text SR, which aims to fill LR images with information while minimizing the loss of their original information. Actually, deep features output from the last basic block usually capture much high-dimensional semantic information while missing some original low-dimensional information. [59] adaptively fused RGB saliency maps and depth saliency maps by switch map learnt by a convolutional saliency fusion model. The switch map can weigh the proportions of two complementary modalities. In order to balance the gains of shallow features and deep features, which are two similar and complementary information flows, we propose a Weight Balanced Fusion Module (WBFM) to learn balanced fusion weight and adaptively fuse them, which is depicted in Fig. 6. Firstly deep residual features and shallow upsampled features are concatenated together, after two convolutional operations and Sigmoid function, a balanced fusion weight that takes value from 0 to 1 can be obtained, and we use the balanced fusion weight to multiply the residual feature maps and 1-balanced fusion weight to multiply the upsampled feature maps. The mathematical formulas are expressed as follows:

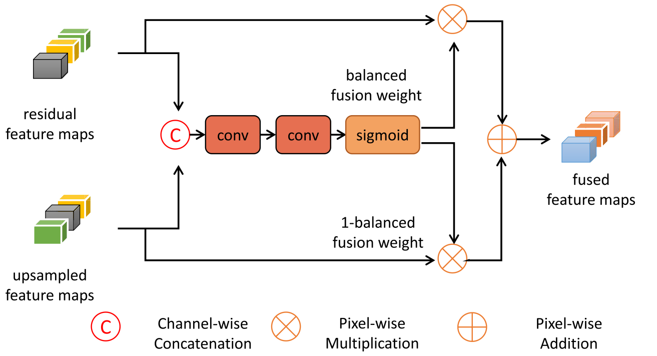

Fig. 6. The illustration of Weight Balanced Fusion Module. The addition of this module can adaptively fuse two complementary information flows.

$$\begin{aligned} fw &= s(h(g([f_r;\ f_u]))) \\ f_f &= fw * f_r + (1 - fw) * f_u \end{aligned} \tag{5}$$

where fw denotes the balanced fusion weight, $g(\cdot)$ denotes convolutional operation whose kernel size is 3×3, $h(\cdot)$ denotes convolutional operation whose kernel size is 1×1, $s(\cdot)$ denotes Sigmoid function, [;] denotes channel-wise concatenation, f_r denotes residual feature maps, f_u denotes upsampled feature maps and f_f denotes fused feature maps.

D. End-to-End Optimization

Based on previous works, MSELoss and Gradient Profile Prior loss proposed in TSRN [13] are applied as the main part of total loss function. Besides, because TEAN is a hierarchical structure for image reconstruction, so the bicubic downsampling is used to resize the super-resolved images to the same size as corresponding high-resolution text images at each level to calculate loss. The mathematical formula of main loss function is expressed as follows:

$$L_{main} = \sum_{i=1}^{L} (\rho_i * L_2(b(I_i^{SR}), I^{HR}) + \mu_i * L_{GP}(b(I_i^{SR}), I^{HR})) \tag{6}$$

where L denotes the number of total levels, $b()$ denotes bicubic [60] operation, I_i^{SR} denotes super-resolved images at i_{th} level, ρ_i and μ_i denote a set of hyper-parameters that adjust the important relationships between these loss terms ($i = 1, 2, 3 \ldots L$). In our model, we set the maximum weight for the image generated by the last level and the same smaller weight for images generated by remaining levels, ensuring the sum of ρ_i is 1. For example, as for the one-level network, we set ρ_1 as 1 and μ_1 as 10^{-4}, which is similar to [13]. The ratio of ρ and μ is constant in multi-level settings. When L=2, ρ_1 is set to $\frac{1}{3}$ and ρ_2 is set to $\frac{2}{3}$. In three-level network, ρ_1 and ρ_2 are equally set to $\frac{1}{4}$ and ρ_3 is set to $\frac{1}{2}$ following the above rules. Then, not only can calibrated masks [17], [61] be correct prior semantic information, but also masks of HR text images can be used to supervise the Text-Segmented-Contextual-Attention-Weight generated by Text-Segmented-Contextual-Attention(TSCA) branch. Therefore, a MSELoss is added to calculate a segmentation mask loss to narrow the gap between the Text-Segmented-Contextual-Attention-Weight and the segmentation masks of HR text images so that the prior semantic information of HR text images can be introduced into model to enhance the stroke regions. This method forms the proposed Text Segmented Mask Loss L_{tsm}, which can be formulated as follows:

$$L_{tsm} = L_2(TSCAW, HR\ masks) \tag{7}$$

Besides, [24] utilizes perception loss to generate super-resolved images which are photo-realistic and friendly for human eyes, to generate super-resolved images that can be easily recognized, a Text Semantic Loss L_{ts} is adpoted to introduce knowledge of pre-trained OCR model into our text SR model to construct a strong baseline on the basis of LapSRN [20]. Specifically, a KL loss is set to supervise this process. If there is no semantic information loss in the SR images compared to their corresponding HR images, semantic features of them will be the same and the KL loss will be infinitely close to zero. So, a pre-trained ASTER-ENCODER [4] is utilized to extract features of super-resolved images for each level and HR images and KL loss is used to calculate a semantic loss which can narrow the distribution between these features. Different coefficients are assigned to these loss terms. The Text Semantic Loss L_{ts} can be formulated as follows:

$$D_{KL}(p\|q) = \sum p(x) \log \frac{p(x)}{q(x)} \tag{8}$$

$$L_{ts} = \sum_{i=1}^{L} \rho_i * D_{KL}(AE(I^{HR}) \| AE(b(I_i^{SR}))) \tag{9}$$

where D_{KL} denotes the consistency of data distribution over two column vectors, AE denotes pre-trained ASTER-ENCODER. So the total optimization function can be formalized as follows:

$$L_{total} = L_{main} + L_{ts} + L_{tsm} \tag{10}$$

here, L_{main} focuses on pixel-wise reconstruction and edge contour details, L_{ts} is used to introduce semantic information of pre-trained OCR model into text SR model as an auxiliary supervision and L_{tsm} aims to make use of segmentation information of HR images to enhance the stroke regions.

E. Differences Between Image Deblurring, Image Super-Resolution and Image Enhancement

In this subsection we will discuss the differences between image deblurring, super-resolution and enhancement. Firstly, all three of them belong to the category of image processing. Secondly, image super-resolution and image deblurring can both be considered as subtasks of image restoration, which learn the causes of image quality degradation by using given training image pairs to recover the quality of the input image as much as possible. The goal of this task is to generate images close to their corresponding GroundTruth from the input low-quality images. The metrics for evaluating image restoration are PSNR and SSIM. However, image enhancement does not focus on the causes of image degradation as image restoration and it rather generates higher-quality images by equalising, balancing, sharpening etc. The goal of image enhancement is to make the image more visually appealing by means of specific filters, such as improving brightness, contrast and etc. The metrics for evaluating image enhancement are often based on subjective visual judgements or task-driven. Thirdly, for image deblurring tasks, resolution of input and output image are the same, while for image super-resolution tasks, the resolution of output image is usually several times that of the input image, which is achieved by the pixel-shuffle module [24]. In other words, image super-resolution increases the number of pixels of the input image while image deblurring does not. Actually, the above two are technically related, in [62], they removed the upscale module and added a global residual learning to transform the origin super-resolution network to a deblurring network and achieved state-of-the-art performance at that time. It also illustrates the differences between image super-resolution and image deblurring, despite their technical relevance.

IV. Experiments

A. Datasets and Evaluation

1) TextZoom Dataset: It was proposed by [13] as the first dataset which aimed at scene text super-resolution. As for training dataset, TextZoom [13] contains 17367 LR-HR image pairs and its magnification factor is ×2. It comes from two previous state-of-the-art SISR datasets: RealSR [63] and SRRAW [64] which consist of LR-HR image pairs captured by digital cameras. Based on these two datasets, TextZoom [13] includes additional text annotation. Out of different shooting distances and focal lengths of the images in the above two datasets, TextZoom [13] divides its test datasets into three subsets, easy, medium, and hard.

2) IC15 Dataset: To verify the generalizability of TEAN and its improvement to the performance of scene text recognition, we adopt IC15 [65] as the other test dataset. It contains 2077 text images and the reason for choosing it is that it has relatively lower resolution than other benchmarks, which is suitable to validate the effectiveness of text super-resolution.

3) Evaluation: The key point of text SR task is to improve the recognition accuracy of state-of-the-art text recognition models on real LR images. Similar to [13] and [14], the recognition accuracy of the three pre-trained state-of-the-art text recognition models, ASTER [4], CRNN [3] and MORAN [5] are utilized as the most important metric for evaluating the proposed text SR approach. And we also test the scene text recognition performance of super-resolved text images generated by TEAN on other three newly developed text recognition models [8], [9], [10]. Besides, the metrics PSNR and SSIM are also used as auxiliary references.

B. Implementation Details

As illustrated in section III, we firstly process TextZoom [13] datasets to get correct binary masks and concatenate them with corresponding text images. All the LR and HR images are respectively upsampled to 64×16, 128×32 for avoiding extra downsample degradation. Similar to [13] and [14], the learning rate for training is set to 0.0001 and Adam optimizer with momentum term 0.9 is used. When evaluating text recognition accuracy, the state-of-the-art text recognition model CRNN [3], ASTER [4], MORAN [5], NRTR [8], SAR [9] and MASTER [10] are used. The proposed model is trained by 500 epochs with two NVIDIA RTX 1080ti GPUs.

C. Ablation Study

1) The Effectiveness of Mask Calibration: As illustrated in section III, we make some calibration to the mask operation proposed in [13]. Correct masks not only can be positive prior semantic information but also can be used to supervise the TSCAW generated by TSCA. Actually, if information of inconsistent masks is introduced, the model won't enhance the stroke regions as expected. So in this subsection, we conduct a experiment to validate the effectiveness of mask calibration on the basis of a strong baseline. In detail, the constructed strong baseline is one-level LapSRN [20] with TPS-align module and basic blocks with orthogonal Grus, besides, MSELoss, Gradient Prior Loss [13] and Text Semantic Loss L_{ts} as shown in section III are used to supervise the network. After that, TSCA is added into each basic block and calibrated masks and uncalibrated masks are utilized to supervise the TSCAW respectively. As shown in Table I, performance of these two models and the constructed strong baseline is listed. We find that adding TSCA without mask calibration leads to even worse performance than baseline because it introduces some wrong and ambiguous prior semantic information of HR images so that suppresses text regions of super-resolved images. Compare with it, adding TSCA with mask calibration can outperform baseline by nearly 2%. So mask calibration is vital for adding TSCA, in subsequent experiments, TSCAW is supervised by calibrated masks by default.

TABLE I

ABLATION STUDY FOR EFFECTIVENESS OF MASK CALIBRATION WITH TSCA

Configuration		Accuracy of ASTER [4]			
Method	Loss	easy	medium	hard	average
baseline	L_2+L_{GP}+L_{ts}	75.4%	58.8%	40.6%	59.4%
baseline+TSCA	L_2+L_{GP}+L_{ts}+L_{tsm}	73.9%	58.6%	40.1%	58.6%
baseline+TSCA+mask calibration	L_2+L_{GP}+L_{ts}+L_{tsm}	**77.2%**	**60.3%**	**43.1%**	**61.3%**

TABLE II

ABLATION STUDY FOR DIFFERENT SETTINGS OF OUR METHOD

Configuration		Accuracy of ASTER [4]			
Method	Loss	easy	medium	hard	average
LapSRN [20]	*Charbonnier*	70.5%	49.8%	35.2%	53.0%
LapSRN [20]+tps+Gru [53]	L_2+L_{GP}+L_{ts}	75.4%	58.8%	40.6%	59.4%
LapSRN [20]+tps+Gru [53] +OCAM	L_2+L_{GP}+L_{ts}	77.8%	60.1%	42.8%	61.3%
LapSRN [20]+tps+Gru [53] +TSCA	L_2+L_{GP}+L_{ts}+L_{tsm}	77.2%	60.3%	43.1%	61.3%
LapSRN [20]+tps+Gru [53] +WBFM	L_2+L_{GP}+L_{ts}	76.1%	59.3%	41.3%	60.0%
Full model without TSCA	L_2+L_{GP}+L_{ts}	78.1%	60.6%	43.4%	61.7%
Full model without OCAM	L_2+L_{GP}+L_{ts}+L_{tsm}	77.7%	60.9%	43.6%	61.8%
Full model without WBFM	L_2+L_{GP}+L_{ts}+L_{tsm}	78.3%	61.2%	43.8%	62.2%
Full model (TEAN-one-level)	L_2+L_{GP}+L_{ts}+L_{tsm}	78.6%	61.8%	44.2%	62.6%
Full model (TEAN-two-level)	L_2+L_{GP}+L_{ts}+L_{tsm}	**80.4%**	**64.5%**	**45.6%**	**64.6%**

2) Overall Analysis: A series of experiments are conducted to validate effectiveness of the proposed modules in TEAN. With this in mind, several different experimental settings are listed: 1) original LapSRN [20]; 2) the constructed strong baseline: TPS-LapSRN [20] with GRu [53] and Text Semantic Loss(L_{ts}); 3) the model which only adds the proposed Orthogonal Contextual Attention Module(OCAM) on the basis of 2); 4) the model which only adds Text-Segmented-Contextual-Attention branch (TSCA) on the basis of 2); 5) the model which only adds Weight Balanced Fusion Module (WBFM) on the basis of 2); 6) the full model without TSCA; 7) the full model without OCAM; 8) the full model without WBFM; 9) the one-level full text SR model; 10) the two-level full text SR model. For the fair comparison, we conduct these ablation experiments on the basis of one-level structure. As shown in Table II, the performance of these models is enumerated.

3) Comparison of the Number of Total Levels: Hierarchical structure of LapSRN [20] can reconstruct LR images progressively and upscale them to different factors. In previous experiments, the one-level structure is adopted unanimously. Intuitively, performance of TEAN can be improved by adding the number of total levels. In fact, hierarchical structure inevitably have a lot of redundancy, which leads to a complex model. So to weigh the accuracy and complexity of the total model, we test different levels to compare their performance. For a fair comparison, we only add TSCA into basic blocks of the last level. As shown in Table III, TEAN with 1,2,3 level are compared. We can find that a deeper network does not boosts performance but even decreases. This is because a deeper structure can make optimization more difficult and the performance gained from such structure is not endless. So the number of total levels is set to 2 to get the best performance. We also use this setting in subsequent experimental phase.

4) Analysis About the Specific Architecture of OCAM: Strip convolutional operations are utilized in local contextual branch of OCAM. 1 × 3 and 3 × 1 convolution are essential for modelling orthogonal features, which has been validated in [49]. So in this subsection three different settings are taken into consideration as shown in Table IV. It can be found that adding either 1 × 3 or 3 × 1 convolution improves

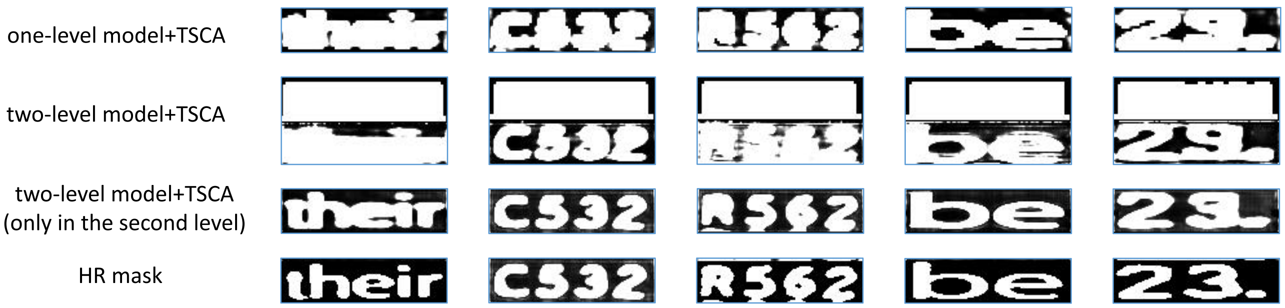

Fig. 7. Some visualization results of three different settings and corresponding HR masks. For the setting where TSCA is added into both levels, we visualize the Text-Segmented-Contextual-Attention-Weight generated in the first level and the second level. For the sake of fair comparison, we visualize the Text-Segmented-Contextual-Attention-Weight generated by the last TEAB at each level.

TABLE III

ABLATION STUDY FOR THE NUMBER OF TOTAL LEVELS

Configuration Levels	Accuracy of ASTER [4] easy	medium	hard	average
1	78.6%	61.8%	44.2%	62.6%
2	**80.4%**	**64.5%**	**45.6%**	**64.6%**
3	78.3%	58.8%	40.4%	60.4%

TABLE IV

ABLATION STUDY ABOUT OCAM

Configuration	Accuracy of ASTER [4] easy	medium	hard	average
only using channel attention in OCAM	77.4%	60.0%	42.5%	61.1%
channel attention without 3x1 convolution	79.7%	63.0%	44.6%	63.5%
channel attention without 1x3 convolution	79.1%	61.8%	44.7%	63.0%
full model	**80.4%**	**64.5%**	**45.6%**	**64.6%**

TABLE V

DIFFERENT SETTINGS FOR TSCA AND THEIR OWN EXPERIMENTAL RESULTS

Configuration	Accuracy of ASTER [4] easy	medium	hard	average
one-level model+TSCA	78.6%	61.8%	44.2%	62.6%
two-level model+TSCA	78.0%	61.4%	43.3%	62.0%
two-level model+TSCA (only in the second level)	**80.4%**	**64.5%**	**45.6%**	**64.6%**

performance on the basis of channel attention [18], and adding both 1×3 and 3×1 convolution achieves the best performance, which shows modelling horizontal and vertical high-frequency information are equally vital and complementary for text SR.

5) TSCA Branch in Two-Level Model: In this subsection, we conduct experiments to find out how to use TSCA to achieve the best results in two-level model. As shown in Table V, we compare three different settings:1) only use one-level model and add TSCA; 2) use two-level model and add TSCA into basic blocks of both levels; 3) use two-level model and only add TSCA into basic blocks of the second level. Besides, some visual results of Text-Segmented-Contextual-Weight generated by the above three models and corresponding HR masks are provided in Fig. 7. Qualitative and quantitative results both show that using a two-level model with TSCA only added in the second level can achieve the best performance. As for the one-level model, since the size of the Text-Segmented-Contextual-Weight generated by TSCA is half the size of HR masks, interpolating them to calculate the Text Segmented Mask Loss L_{tsm} would result in a loss of information, and comparing the interpolated weight with the HR mask would increase the learning difficulty of total network, this is part of the reason why the performance of three-level structure is also worse than two-level structure. Besides, the one-level model generates Text-Segmented-Contextual-Weight based on LR text images, which would also increase the learning difficulty of this model. This also shows the superiority of hierarchical structure compared to single-level structure and the proposed TSCA is more effective in two-level model than in one-level model. And as for two-level model, after the first level, our model can obtain sub-optimal images and during the second level our model can obtain better segmented masks easily through these sub-optimal images, which also reduces the learning difficulty. As for the two-level model which adds TSCA into basic blocks of both levels, the model will focus on the learning the second level so that the first level will obtain incorrect weights, and these incorrect weights will in turn affect the learning of the second level, this is why the weights generated by this model do not correspond well to the HR masks. And interpolating the TSCAW generated in the first level and then calculating the Text Segmented Mask Loss L_{tsm} with corresponding HR masks also lead to poor results as mentioned above. Combining these above points, we choose to add TSCA into basic blocks of the second level only to construct the final model.

D. Experiments on TextZoom Dataset

We compare the performance of fifteen different text SR methods on TextZoom [13] dataset as shown in Table VI, which includes SRCNN [11], VDSR [67], SRResNet [24], RRDB [68], EDSR [69], RDN [70], LapSRN [20], RCAN [71], SAN [72], HAN [73], TSRN [13], TSRGAN [44], TBSRN [45], Text Gestalt [31], PCAN [14]. All of them are trained on TextZoom [13] dataset and tested on three subsets. We directly list the reported experimental results of the above methods rather than re-implementing them. Compared with recognizing LR images of TextZoom [13] without super-resolution, the proposed TEAN can outperform by 17.4% on ASTER [4], 17.3% on MORAN [5] and 25.4% on CRNN [3] separately. Besides, compared with fifteen earlier proposed text SR methods, TEAN can increase by 3-15% on ASTER [4]. Compared with the SOTA text SR method [31], TEAN outperforms it by 3.3%, 2.0%, 3.3% on ASTER [4], MORAN [5]

TABLE VI

Performance of Mainstream SISR Algorithms on the Three Subsets in TextZoom [13]. In the First Two Rows We Show the Accuracy of the Three Recognition Models for Low and High Resolution Images Respectively. L_{tv} Denotes Total Variation Loss. L_p Denotes Perceptual Loss Proposed in [66]. $L_{Charbonnier}$ Denotes the Charbonnier Loss Proposed in LapSRN [20]. L_{GP} Denotes the Gradient Prior Loss Proposed in [13]. L_d Denotes the Adversarial Loss Used in [44]. L_{wav} Denotes the Wavelet Loss Used in [44]. L_{POS} Denotes the Position-Aware Loss Proposed in [45]. L_{CON} Denotes the Content-Aware Loss Proposed in [45]. L_{EG} Denotes Edge-Guidance Loss Proposed in [14]. L_{SFM} Denotes the Stoke-Focused Module Loss Proposed in [31]. L_{ts} Denotes Our Proposed Text Semantic Loss. L_{tsm} Denotes Our Proposed Text Segmented Mask Loss

Method	Loss	Accuracy of ASTER [4]				Accuracy of MORAN [5]				Accuracy of CRNN [3]			
		easy	medium	hard	average	easy	medium	hard	average	easy	medium	hard	average
LR	–	64.7%	42.4%	31.2%	47.2%	60.6%	37.9%	30.8%	44.1%	36.4%	21.1%	21.1%	26.8%
HR	–	94.2%	87.7%	76.2%	86.6%	91.2%	85.3%	74.2%	84.1%	76.4%	75.1%	64.6%	72.4%
BICUBIC	–	64.7%	42.4%	31.2%	47.2%	60.6%	37.9%	30.8%	44.1%	36.4%	21.1%	21.1%	26.8%
SRCNN [11]	L_2	69.4%	43.4%	32.2%	49.5%	63.2%	39.0%	30.2%	45.3%	38.7%	21.6%	20.9%	27.7%
VDSR [67]	L_2	71.7%	43.5%	34.0%	51.0%	62.3%	42.5%	30.5%	46.1%	41.2%	25.6%	23.3%	30.7%
SRResNet [24]	L_2+L_{tv}+L_p	69.4%	47.3%	34.3%	51.3%	60.7%	42.9%	32.6%	46.3%	39.7%	27.6%	22.7%	30.6%
RRDB [68]	L_1	70.9%	44.4%	32.5%	50.6%	63.9%	41.0%	30.8%	46.3%	40.6%	22.1%	21.9%	28.9%
EDSR [69]	L_1	72.3%	48.6%	34.3%	53.0%	63.6%	45.4%	32.2%	48.1%	42.7%	29.3%	24.1%	32.7%
RDN [70]	L_1	70.0%	47.0%	34.0%	51.5%	61.7%	42.0%	31.6%	46.1%	41.6%	24.4%	23.5%	30.5%
LapSRN [20]	$L_{Charbonnier}$	71.5%	48.6%	35.2%	53.0%	64.6%	44.9%	32.2%	48.3%	46.1%	27.9%	23.6%	33.3%
RCAN [71]	L_1	67.3%	46.6%	35.1%	50.7%	63.1%	42.9%	33.6%	47.5%	46.8%	27.9%	26.5%	34.5%
SAN [72]	L_1	68.1%	48.7%	36.2%	52.0%	65.6%	44.4%	35.2%	49.4%	50.1%	31.2%	28.1%	37.2%
HAN [73]	L_2	71.1%	52.8%	39.0%	55.3%	67.4%	48.5%	35.4%	51.5%	51.6%	35.8%	29.0%	39.6%
TSRN [13]	L_2+L_{GP}	75.1%	56.3%	40.1%	58.3%	70.1%	53.3%	37.9%	54.8%	52.5%	38.2%	31.4%	41.4%
TSRGAN [44]	L_2+L_d+L_{wav}	75.7%	57.3%	40.9%	59.1%	72.0%	54.6%	39.3%	56.3%	56.2%	42.5%	32.8%	44.6%
TBSRN [45]	L_2+L_{POS}+L_{CON}	75.7%	59.9%	41.6%	60.1%	74.1%	57.0%	40.8%	58.4%	59.6%	47.1%	35.3%	48.1%
PCAN [14]	L_2+L_{EG}	77.5%	60.7%	43.1%	61.5%	73.7%	57.6%	41.0%	58.5%	59.6%	45.4%	34.8%	47.4%
Text Gestalt [31]	L_2+L_{SFM}	77.9%	60.2%	42.4%	61.3%	75.8%	57.8%	41.4%	59.4%	61.2%	47.6%	35.5%	48.9%
TEAN (ours)	L_2+L_{GP}+L_{ts}+L_{tsm}	**80.4%**	**64.5%**	**45.6%**	**64.6%**	**76.8%**	**60.8%**	**43.4%**	**61.4%**	**63.7%**	**52.5%**	**38.1%**	**52.2%**

TABLE VII

Performance of Three Newly Developed Scene Text Recognition Models on the Three Subsets in TextZoom [13]. As for These Three Models, We Use the Pytorch Code and Released Model in [74]

Method	Loss	Accuracy of NRTR [8]				Accuracy of SAR [9]				Accuracy of MASTER [10]			
		easy	medium	hard	average	easy	medium	hard	average	easy	medium	hard	average
LR	–	56.5%	36.9%	28.8%	41.7%	54.7%	38.1%	29.7%	41.7%	64.8%	41.6%	30.6%	46.8%
HR	–	90.6%	86.2%	75.4%	84.5%	90.0%	86.9%	74.5%	84.2%	91.9%	88.2%	77.1%	86.2%
TSRN [13]	L_2+L_{GP}	68.8%	51.1%	37.0%	53.3%	65.1%	46.4%	34.0%	49.5%	70.8%	53.9%	37.5%	55.1%
PCAN [14]	L_2+L_{EG}	71.6%	55.7%	40.8%	57.0%	70.4%	51.3%	36.8%	53.9%	73.9%	58.6%	41.3%	59.0%
Text Gestalt [31]	L_2+L_{SFM}	75.3%	60.5%	42.3%	60.4%	71.8%	56.6%	38.8%	56.8%	75.5%	62.0%	43.1%	61.2%
TEAN (ours)	L_2+L_{GP}+L_{ts}+L_{tsm}	**77.6%**	**62.8%**	**43.8%**	**62.4%**	**74.2%**	**58.6%**	**41.2%**	**59.0%**	**77.0%**	**62.7%**	**45.6%**	**62.7%**

TABLE VIII

PSNR and SSIM Results of Different SR Methods on TextZoom [13]

Method	PSNR			SSIM		
	easy	medium	hard	easy	medium	hard
BICUBIC	22.35	18.98	19.39	0.7884	0.6254	0.6592
SRCNN [11]	23.48	19.06	19.34	0.8379	0.6323	0.6791
VDSR [67]	24.62	18.96	19.79	0.8631	0.6166	0.6989
SRResNet [24]	24.36	18.88	19.29	0.8681	0.6406	0.6911
RRDB [68]	22.12	18.35	19.15	0.8351	0.6194	0.6856
EDSR [69]	24.26	18.63	19.14	0.8633	0.6440	0.7108
RDN [70]	22.27	18.95	19.70	0.8249	0.6427	0.7113
LapSRN [20]	24.58	18.85	19.77	0.8556	0.6480	0.7087
RCAN [71]	22.15	18.81	19.83	0.8525	0.6465	0.7227
SAN [72]	22.69	18.77	19.82	0.8597	0.6477	0.7280
HAN [73]	23.30	19.02	20.16	0.8691	0.6537	0.7387
TSRN [13]	**25.07**	18.86	19.71	0.8897	0.6676	0.7302
PCAN [14]	24.57	19.14	20.26	0.8830	0.6781	0.7475
Text Gestalt [31]	23.34	19.06	19.90	0.8369	0.6499	0.6986
TEAN (ours)	24.91	**19.26**	**20.35**	**0.8945**	**0.6834**	**0.7581**

and CRNN [3] respectively. At the same time, to further validate the improvement brought by TEAN for scene text recognition, the performance on three newly developed text recognition models [8], [9], [10] is also measured. It can be seen that the recognition performance of the three newly developed recognition models for low-resolution images is not improved compared to ASTER [4] while these performance can be significantly improved by using text super-resolution methods. This shows that super-resolution technology is vital for text recognition especially in low-resolution scenes and TEAN can be adopted as a better pre-processor to improve the performance of existing text recognition models. Then, PSNR and SSIM are also provided as auxiliary references as shown in Table VIII, although they do not necessarily reflect the sharpness of the texts in super-resolved images [24], the proposed TEAN still remains ahead of existing methods in the vast majority of cases. This is because TEAN works well at enhancing text regions of LR images by introducing the segmentation information of HR images, which results in a high structure similarity between super-resolved images and their corresponding HR images. Finally, visual results of TEAN for super-resolving some low-resolution images are shown in Fig. 8, it is clear that TEAN has certain advantages over other state-of-the-art text SR methods.

E. Experiments on IC15 Dataset

We also conduct experiments on IC15 dataset to evaluate the generalization performance and improvement to scene text recognition of TEAN. As there is currently no

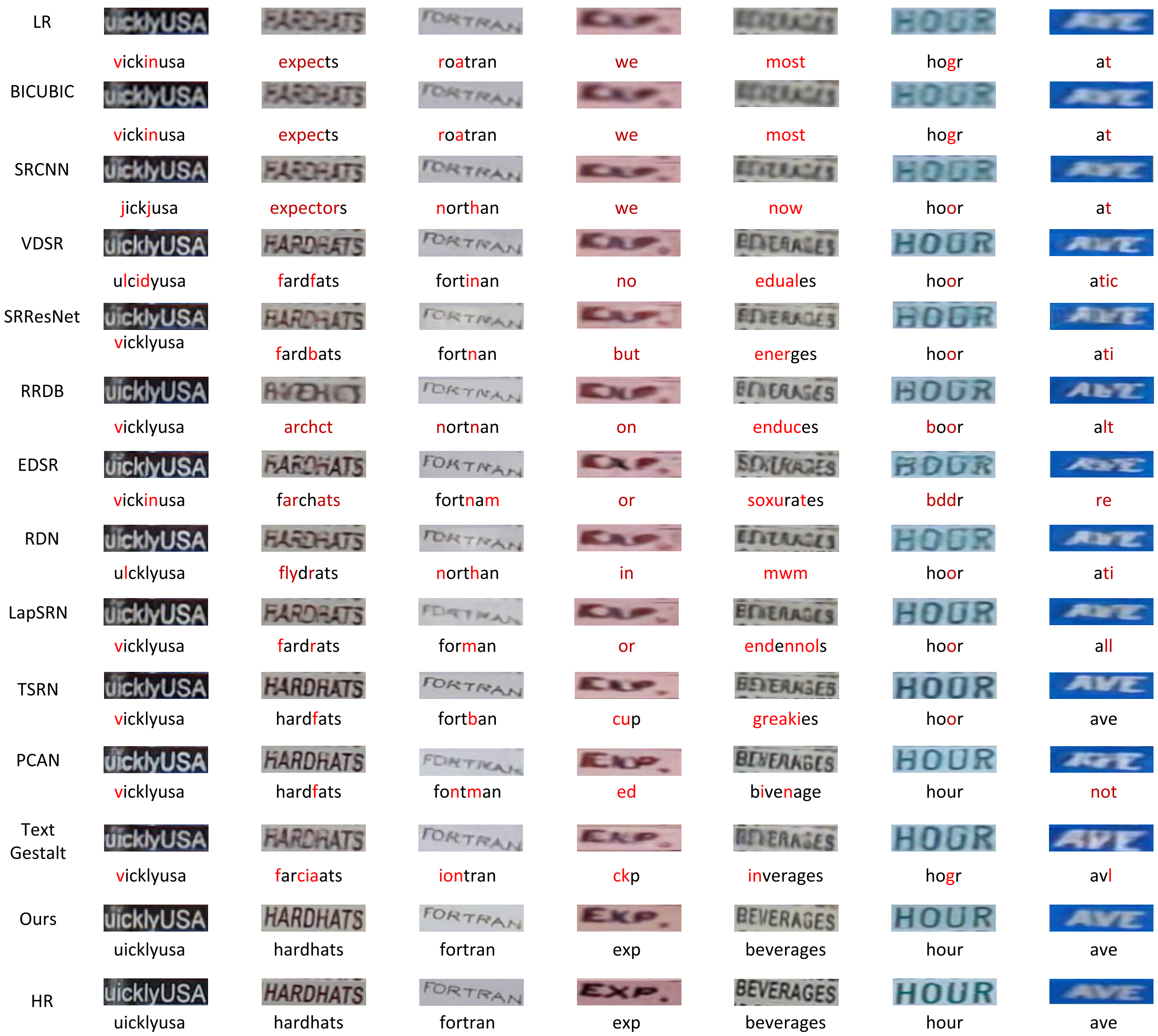

Fig. 8. Some visualization results of existing text SR methods on TextZoom [13] dataset, where the recognition results are provided by ASTER [4].

TABLE IX

THE EXPERIMENTAL RESULTS ON IC15_2077 [65] DATASET. '–' DENOTES RECOGNIZING DIRECTLY WITHOUT SUPER-RESOLUTION

Method	CRNN [3]	MORAN [5]	ASTER [4]	NRTR [8]	SAR [9]	MASTER [10]
–	56.6%	72.0%	**74.2%**	73.6%	78.3%	**78.7%**
TSRN [13]	55.1%	69.5%	70.2%	70.8%	73.3%	75.3%
PCAN [14]	57.8%	70.7%	71.9%	71.3%	73.8%	76.4%
Text Gestalt [31]	59.6%	71.6%	73.7%	73.2%	75.6%	77.3%
TEAN (ours)	**60.5%**	**73.3%**	73.5%	**74.1%**	**78.5%**	77.8%

TABLE X

THE EXPERIMENTAL RESULTS ON TEXT IMAGES INCORRECTLY PROCESSED BY EXISTING RECOGNITION MODELS OF IC15_2077 [65] DATASET

Method	CRNN [3]	MORAN [5]	ASTER [4]	NRTR [8]	SAR [9]	MASTER [10]
–	56.6%	72.0%	74.2%	73.6%	78.3%	78.7%
TEAN (ours)	**65.1%**	**75.8%**	**77.3%**	**76.9%**	**81.5%**	**82.0%**

TABLE XI

THE STATISTICS OF MODEL SIZE (INCLUDING THE SIZE OF TEXT SR MODEL AND ASTER [4] RECOGNITION MODEL) AND TOTAL PROCESSING SPEED (INCLUDING SR+ASTER [4] RECOGNITION MODEL). '–' DENOTES RECOGNIZING WITHOUT TEXT SUPER-RESOLUTION

Method	Model Size /MB	Processing Speed /fps
–	**80.443**	**40.14**
TSRN [13]	10.339+80.443	37.18
PCAN [14]	15.755+80.443	36.10
Text Gestalt [31]	10.339+80.443	37.18
TEAN (one-level)	15.260+80.443	35.24
TEAN (two-level)	18.551+80.443	34.53

dataset for scene text super-resolution except TextZoom [13], so IC15_2077 [65], which is a challenging benchmark for scene text recognition, is choosed to validate the proposed method. The experimental results are shown in Table IX. Since label distribution and image style of IC15 are different from those in TextZoom, there is a cross-domain issue which results in the model pre-trained on TextZoom performs less well on IC15_2077. From Table IX we can see that

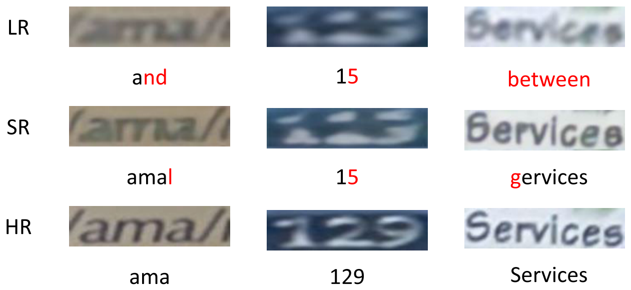

Fig. 9. Visualization of some failure cases and their corresponding recognition results, where the results are provided by ASTER [4].

TSRN [13], Text Gestalt [31] and PCAN [14] inevitably decrease in performances on all recognition models except CRNN [3] while TEAN achieves performance gains on most recognition models, even if some of them are slight. From the experimental results we find that even introducing information from transformer-based recognizer doesn't avoid performance decrease in [31]. This problem is common for scene text super-resolution models trained on TextZoom [13] and the proposed TEAN alleviates it to a certain extent by enhancing stroke regions of input images. Actually, many images in IC15 are clear enough for text recognition and super-resolving them by cross-domain models leads to wrong results instead. So TEAN is used to super-resolve images that are initially incorrectly recognized by text recognition models and experimental results are listed in Table X. Similar experiments have been conducted in [13], [31], [32], and [33], proving that text super-resolution is aimed at low-resolution images and choosing whether to use text SR or not by the resolution of input text images can achieve better overall performance. From the results in Table X we find that TEAN works well with misidentified images and attains more than 3% performance gains on all of text recognition models, which demonstrates the improvement to scene text recognition brought by TEAN even though with cross-domain issue.

F. Discussions of Failure Cases

Although TEAN has achieved comparable performance against state-of-the-art text SR methods, it fails to super-resolve extremely blur text images as shown in Fig. 9. Because this kind of images can't provide enough information for text SR, which results in TEAN fails to learn correct TSCAW to enhance the text regions. How to super-resolve such images poses great challenges to the development of text SR.

G. Discussions of Model Size and Processing Speed

In this subsection, the size of TEAN and its processing speed are discussed as shown in Table XI. For comprehensive comparison, statistics of one-level TEAN and two-level TEAN are both listed. As super-resolution is usually adopted as a pre-processing step in scene text recognition, we calculate the total model size and processing speed of scene text super-resolution and scene text recognition. Since [31] is supervised during the training stage by using transformer-based recognizer on the basis of TSRN [13], the number of its parameters and processing speed during the testing stage is the same as TSRN [13]. Actually, its computational overhead is concentrated in the training stage. Compared with the above methods, although TEAN has larger number of parameters and slower speed, it outperforms the above methods in accuracy and the sacrifices in terms of number of parameters and processing speed can be considered negligible in contrast to text recognition model.

V. Conclusion

In this paper, to process stroke regions and background regions of text images distinctively, segmentation information and attentional guidance are adopted for enhancement of text regions in the proposed model, which achieves excellent performance and should be focused in subsequent study. Meanwhile, deep features and shallow features are also integrated adaptively by the proposed WBFM to avoid loss of original information in text SR process. The proposed TEAN has played a key role in improving the accuracy of text recognition and the quality of text images, which achieves state-of-the-art performance against existing methods. It also achieves better cross-dataset performance compared with previous text-specific methods, which verifies its effectiveness to scene text recognition. Actually, the proposed model has difficulty in processing extremely blurry text images by making use of the limited prior information, which will be focused in our future work.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[2] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 35–48.

[3] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2016.

[4] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[5] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[6] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4168–4176.

[7] J. Baek et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4715–4723.

[8] F. Sheng, Z. Chen, and B. Xu, "NRTR: A no-recurrence sequence-to-sequence model for scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 781–786.

[9] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8610–8617.

[10] N. Lu et al., "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107980.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[12] C. Peyrard, M. Baccouche, F. Mamalet, and C. Garcia, "ICDAR2015 competition on text image super-resolution," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1201–1205.

[13] W. Wang et al., "Scene text image super-resolution in the wild," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 650–666.

[14] C. Zhao et al., "Scene text image super-resolution via parallelly contextual attention network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2908–2917.

[15] Y. Wang, F. Su, and Y. Qian, "Text-attentional conditional generative adversarial network for super-resolution of text images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1024–1029.

[16] J. Jiang, J. Liu, J. Fu, W. Wang, and H. Lu, "Super-resolution semantic segmentation with relation calibrating network," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108501.

[17] X. Xu, Z. Zhang, Z. Wang, B. Price, Z. Wang, and H. Shi, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12045–12055.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[20] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[21] J. Lei et al., "Deep stereoscopic image super-resolution via interaction module," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3051–3061, Aug. 2021.

[22] A. Niu et al., "MS2Net: Multi-scale and multi-stage feature fusion for blurred image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5137–5150, Feb. 2022.

[23] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–7.

[24] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[25] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 661–669.

[26] K. S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Multi-script-oriented text detection and recognition in video/scene/born digital images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1145–1162, Apr. 2019.

[27] B. Li, X. Tang, X. Qi, Y. Chen, C.-G. Li, and R. Xiao, "EMU: Effective multi-hot encoding net for lightweight scene text recognition with a large character set," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5374–5385, Aug. 2022.

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[29] C. Mancas-Thillou and M. Mirmehdi, "An introduction to super-resolution text," in *Digital Document Processing*. Berlin, Germany: Springer, 2007, pp. 305–327.

[30] R. K. Pandey, K. Vignesh, and A. G. Ramakrishnan, "Binary document image super resolution for improved readability and OCR performance," 2018, *arXiv:1812.02475*.

[31] J. Chen, H. Yu, J. Ma, B. Li, and X. Xue, "Text Gestalt: Stroke-aware scene text image super-resolution," 2021, *arXiv:2112.08171*.

[32] J. Ma, S. Guo, and L. Zhang, "Text prior guided scene text image super-resolution," 2021, *arXiv:2106.15368*.

[33] J. Ma, Z. Liang, and L. Zhang, "A text attention network for spatial deformation robust scene text image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5911–5920.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[35] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[36] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[37] P. Cheng, Y. Cai, and W. Wang, "A direct regression scene text detector with position-sensitive segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4171–4181, Nov. 2020.

[38] X. Weng, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Stage-aware feature alignment network for real-time semantic segmentation of street scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4444–4459, Oct. 2021.

[39] W. Shi et al., "RGB-D semantic segmentation and label-oriented voxelgrid fusion for accurate 3D semantic mapping," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 183–197, Jan. 2022.

[40] W. Wang et al., "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9336–9345.

[41] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proc. Conf. Artif. Intell.*, vol. 33, Aug. 2019, pp. 9038–9045.

[42] F. Lv, T. Liang, X. Chen, and G. Lin, "Cross-domain semantic segmentation via domain-invariant interactive relation transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4334–4343.

[43] F. Lv, G. Lin, P. Liu, G. Yang, S. J. Pan, and L. Duan, "Weakly-supervised cross-domain road scene segmentation via multi-level curriculum adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3493–3503, Sep. 2020.

[44] C. Fang, Y. Zhu, L. Liao, and X. Ling, "TSRGAN: Real-world text image super-resolution based on adversarial learning and triplet attention," *Neurocomputing*, vol. 455, pp. 88–96, Sep. 2021.

[45] J. Chen, B. Li, and X. Xue, "Scene text telescope: Text-focused scene image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12026–12035.

[46] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 726–743.

[47] W. Li et al., "Joint local correlation and global contextual information for unsupervised 3D model retrieval and classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3265–3278, May 2022.

[48] J. Chen, L. Yang, L. Tan, and R. Xu, "Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108753.

[49] Y. Wang, H. Xie, Z.-J. Zha, M. Xing, Z. Fu, and Y. Zhang, "ContourNet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11753–11762.

[50] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[51] H. Chen, J. Gu, and Z. Zhang, "Attention in attention network for image super-resolution," 2021, *arXiv:2104.09497*.

[52] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[53] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[54] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 269–284.

[55] P. Zhang, W. Liu, Y. Lei, and H. Lu, "Hyperfusion-Net: Hyper-densely reflective feature fusion for salient object detection," *Pattern Recognit.*, vol. 93, pp. 521–533, Sep. 2019.

[56] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 92–109.

[57] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[60] R. E. Carlson and F. N. Fritsch, "Monotone piecewise bicubic interpolation," *SIAM J. Numer. Anal.*, vol. 22, no. 2, pp. 386–400, Apr. 1985.

[61] S. Bonechi, M. Bianchini, F. Scarselli, and P. Andreini, "Weak supervision for generating pixel-level annotations in scene text segmentation," *Pattern Recognit. Lett.*, vol. 138, pp. 1–7, Oct. 2020.

[62] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image restoration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2020.

[63] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3086–3095.

[64] X. Zhang, Q. Chen, R. Ng, and V. Koltun, “Zoom to learn, learn to zoom,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3762–3770.

[65] D. Karatzas et al., “ICDAR 2015 competition on robust reading,” in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[66] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[67] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[68] X. Wang et al., “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Sep. 2018.

[69] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[70] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[71] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[72] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.

[73] B. Niu et al., “Single image super-resolution via a holistic attention network,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 191–207.

[74] Z. Kuang et al., “MMOCR: A comprehensive toolbox for text detection, recognition and understanding,” 2021, *arXiv:2108.06543*.

Rui Shu is currently pursuing the master’s degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, scene text recognition, and scene text super-resolution.

Cairong Zhao received the B.Sc. degree from Jilin University in 2003, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology in 2011. He is currently a Professor with Tongji University. He is the author of more than 30 scientific papers in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Shuyang Feng is currently pursuing the master’s degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, scene text-related research, and generative adversarial networks.

Liang Zhu is currently pursuing the Graduate degree majoring in computer science with Tongji University. His research interests include deep learning and computer vision, specifically on object detection in X-ray images.

Duoqian Miao was born in 1964. He is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University. He serves as the Vice President for the International Rough Set Society (IRSS), an Executive Manager for the Chinese Association for Artificial Intelligence (CAAI), the Chair for the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member for Chinese Computer Federation (CCF), the Vice President for the Shanghai Computer Federation, and the Vice President for the Shanghai Association for Artificial Intelligence. His research interests include machine learning, data mining, big data analysis, granular computing, artificial intelligence, and text image processing. He has published more than 200 papers in IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA MINING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Pattern Recognition*, *Information Sciences*, *Knowledge-Based Systems*, *Chinese Journal of Computers*, *Journal of Software* (in Chinese), *Journal of Computer Research and Development* (in Chinese), *Automatica Sinica* (in Chinese), and *ACTA Electronica Sinica* (in Chinese). He won the second prize at Wuwenjun AI Science and Technology in 2018. He serves as an Associate Editor for the *International Journal of Approximate Reasoning* and an Editor for the *Journal of Computer Research and Development* (in Chinese).