Scene Text Telescope: Text-Focused Scene Image Super-Resolution

Jingye Chen, Bin Li*, Xiangyang Xue
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
{jingyechen19, libin, xyxue}@fudan.edu.cn

Abstract

Image super-resolution, which is often regarded as a pre-processing procedure of scene text recognition, aims to recover the realistic features from a low-resolution text image. It has always been challenging due to large variations in text shapes, fonts, backgrounds, etc. However, most existing methods employ generic super-resolution frameworks to handle scene text images while ignoring text-specific properties such as text-level layouts and character-level details. In this paper, we establish a text-focused super-resolution framework, called Scene Text Telescope (STT). In terms of text-level layouts, we propose a Transformer-Based Super-Resolution Network (TBSRN) containing a Self-Attention Module to extract sequential information, which is robust to tackle the texts in arbitrary orientations. In terms of character-level details, we propose a Position-Aware Module and a Content-Aware Module to highlight the position and the content of each character. By observing that some characters look indistinguishable in low-resolution conditions, we use a weighted cross-entropy loss to tackle this problem. We conduct extensive experiments, including text recognition with pre-trained recognizers and image quality evaluation, on TextZoom and several scene text recognition benchmarks to assess the super-resolution images. The experimental results show that our STT can indeed generate text-focused super-resolution images and outperform the existing methods in terms of recognition accuracy.

1. Introduction

Scene text recognition (STR) has drawn much research interest of the computer vision community due to its various applications such as license plate recognition and ID card recognition [17, 23, 39]. While STR has made a big step forward with the development of deep learning, recognition performance on low-resolution (LR) text images is still subpar [44]. LR text images exist in many situations, *e.g.*, a photo taken with a low-focal camera or a document image compressed to reduce disk usages. When handling LR text images, existing recognition or spotting methods usually employ interpolation methods, like bicubic and bilinear interpolations, to upsample the original images [4, 27, 37, 38]. As shown in Figure 1, the image upsampled by the interpolation method is still blurred, which indeed brings difficulties to existing recognition models.

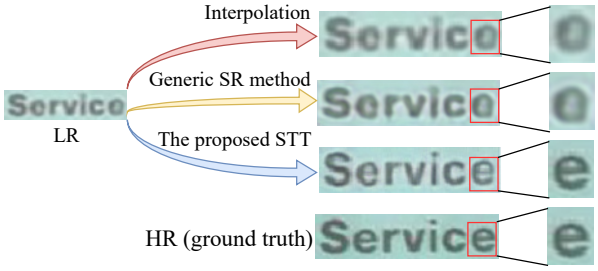


Figure 1. The proposed STT generates a relatively clearer text image and pays more attention to character details compared with interpolation methods and generic SR methods. "SR" and "HR" denote super-resolution and high-resolution, respectively.

In recent years, several works employ generic super-resolution methods for text image super-resolution. For example, in [8], SRCNN [7] with a shallow network is used as the backbone. In [42], a Laplacian-pyramid backbone is employed to combine features from several middle layers to upsample low-resolution images. However, these methods are not suitable for processing text images [44] since they see text images as general ones without taking text-specific properties (*e.g.* text-level layouts and character-level details) into consideration. In contrast, there are few methods that take a part of these properties into account. For example, PlugNet [30] designs a multi-task framework, aiming to recognize and upsample text images in one model. In [44], a Text Super-Resolution Network (TSRN) containing a horizontal and a vertical BLSTMs [11] is proposed to capture sequential information of text images. However, the BLSTMs are not suitable for capturing sequential information of inclined or curved text images.

*Corresponding author

In this paper, we propose a text-focused super-resolution framework, called Scene Text Telescope. To tackle texts in arbitrary orientations, we propose a novel backbone, namely Transformer-Based Super-Resolution Network (TBSRN) to capture sequential information. We notice that the previous methods usually employ loss functions that focus on every pixel of the image, which may suffer great disturbances from backgrounds. According to Inattentional Blindness [28, 29], when humans observe a text image, they will naturally pay more attention to text regions rather than backgrounds, *i.e.*, there is no need to improve the quality of the whole image in the super-resolution task. Based on this fact, we put forward a Position-Aware Module and a Content-Aware Module to focus on the position and the content of each character. By observing that there are some confusable characters in the low-resolution situation (*e.g.* in Figure 1, "c" and "e" look similar), we employ a weighted cross-entropy loss in the Content-Aware Module to address this problem. Since these two modules are only used as text-specific guidance when training, they will not bring additional time overhead in the test stage.

We mainly evaluate our method on TextZoom [44], which contains LR-HR pairs captured from digital cameras. Furthermore, we conduct several experiments on scene text recognition benchmarks to further verify the capabilities of our STT as a preprocessor. In this work, we employ some widely used recognition models (*e.g.* ASTER [38], MORAN [26], and CRNN [37]) and image quality metrics, to evaluate the generated SR images. The experimental results show that the proposed STT can indeed generate text-focused super-resolution images and outperform existing methods in terms of recognition accuracy. Contributions of the proposed STT can be concluded in three-fold:

- We propose TBSRN to capture sequential information, which is more robust on texts in arbitrary orientations.
- A Position-Aware Module and a Content-Aware Module with a weighted cross-entropy loss are proposed to highlight the position and content of characters without bringing additional time overhead when testing.
- The proposed STT generates text-focused SR images and achieves higher recognition accuracy on pre-trained recognizers than other existing methods.

2. Related Work

In this section, we first review the methods for scene text recognition, then go down to the literature of applying super-resolution to text images.

2.1. Scene Text Recognition

Scene text recognition (STR) has made great progress in recent years. CRNN [37] first combines CNN and RNN to obtain sequential features of text images, which are further fed into a CTC decoder [10] to maximize the probability of paths that can reach the ground truth. ASTER [38] employs a Spatial Transformer Network [14] to rectify text images and uses the attention mechanism to focus on a specific character at each time step. FAN [3] observes the attention drift problem and proposes a focusing network to rectify attention regions. Char-Net [24] rectifies text images at the character-level. However, these methods employ a 1-D feature map to encode text images, which is not suitable for tackling curved texts. Therefore, SAR [21] employs a 2-D attention map and achieves better performance on many STR benchmarks. In [22], a framework is proposed to solve STR in a two-dimensional perspective. Besides, several methods such as [36, 46] use Transformer in STR and show competitive performance with the existing methods. However, low-resolution text images still bring difficulties to these methods. Therefore, the research on text image super-resolution is of great significance.

2.2. Text Image Super-Resolution

Text image SR methods can be divided into two classes: traditional methods and deep learning-based methods. Traditional methods tend to employ traditional machine learning strategies. In [2], two estimators are proposed to enhance text images, including a Maximum *a posteriori* (MAP) estimator [9] based on a Huber prior and an estimator regularized using the Total Variation norm. In [6], a Bayesian framework is formulated to improve the visual quality of fax documents or low-resolution scans. In [31], a regularization method with bilateral Total Variation stabilizer and bimodal penalty function is used to perform SR. In recent years, plenty of text image super-resolution methods are conducted with deep learning-based models. In [8], SRCNN [7] is used as the backbone to perform super-resolution. In [32], three SR frameworks are proposed to perform SR on binary document images. In [42], a Laplacian-pyramid backbone is employed to upsample low-resolution images. These methods are not suitable for handling scene text images [44] because they directly use generic SR frameworks and ignore text-specific properties such as text-level layouts and character-level details. There are few methods that take a part of these properties into account. PlugNet [30] designs a multi-task framework by performing recognition and super-resolution simultaneously. In [44], a Text Super-Resolution Network (TSRN) containing a horizontal BLSTM and a vertical BLSTM [11] is proposed to capture sequential information in text images. However, these methods focus on every pixel in the image, which may suffer great disturbances from backgrounds, thereby affecting the performance of upsampling on text regions. Previous works show that the scene text image super-resolution task is far from being resolved.

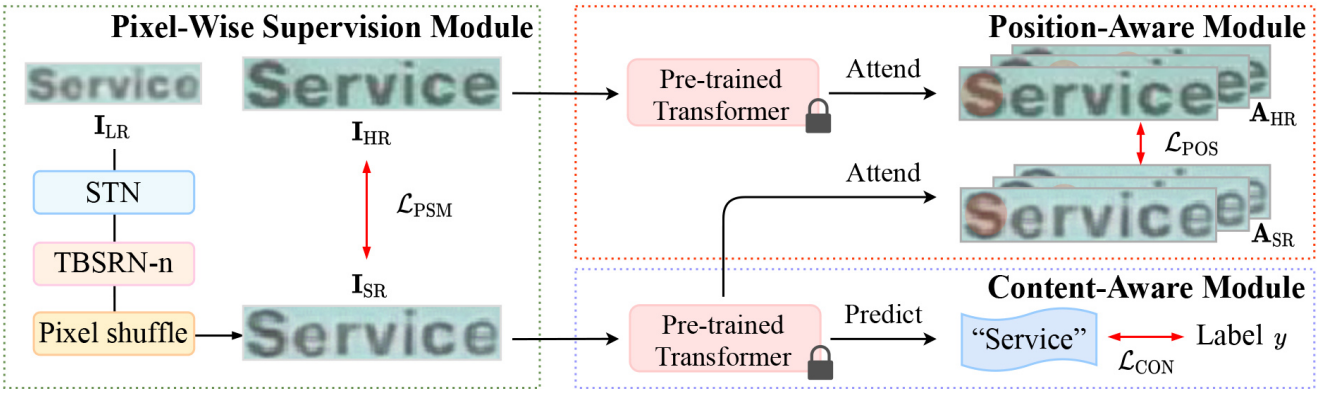


Figure 2. The overall architecture of our STT consists of three parts, including a Pixel-Wise Supervision Module, a Position-Aware Module, and a Content-Aware Module. "TBSRN-n" means n consecutive TBSRN blocks. Parameters in the pre-trained Transformer are frozen.

3. Methodology

The overall architecture of the proposed STT is shown in Figure 2. In the Pixel-Wise Supervision Module (green dotted frame), a low-resolution text image is first rectified by a Spatial Transformer Network (STN) [14] to tackle the misalignment problem, which is mentioned in [44]. Sequentially, the rectified image goes down to a series of Transformer-Based Super-Resolution Networks (TBSRN), then upsamples to a super-resolution text image by pixel shuffling. In the Position-Aware Module (red dotted frame), by taking the corresponding HR image as a reference, the attention maps of the HR image and the SR image are supervised by an L1 loss. The Content-Aware Module (blue dotted frame) provides clues about the content and employs a weighted cross-entropy loss to distinguish confusable characters. Details are introduced in the following.

3.1. Pixel-Wise Supervision Module

Most super-resolution methods often employ variants of Fully Convolutional Network [25] as backbones [7, 40, 41]. To capture sequential information in text images, TSRN [44] appends a horizontal BLSTM and a vertical BLSTM [11] in the backbone. However, texts in natural scenes can be inclined or curved, which brings difficulties to the BLSTMs. Inspired by the 2-D attention in STR [21, 22], we propose a Transformer-Based Super-Resolution Network (TBSRN), which mainly contains a Self-Attention Module and a Position-Wise Feed-Forward Module. As the Self-Attention Module can correlate any pixel pairs in feature maps, it is robust to handle text images in arbitrary orientations. Each TBSRN unit is demonstrated in Figure 3. After rectified by STN, the image is fed into two consecutive CNNs to extract a feature map, which is further sent to the Self-Attention Module to capture sequential information. Unlike those RNN-based models which can implicitly attain positional clues according to time steps, the Self-Attention Module is not aware of spatial positional information since

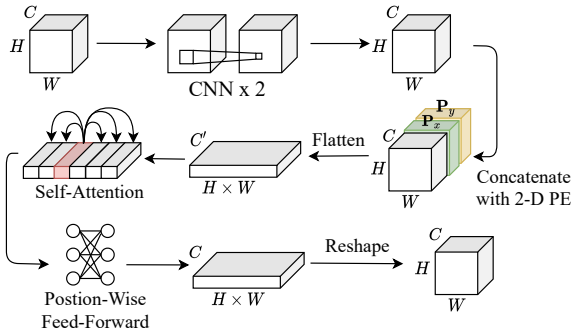


Figure 3. The illustration of the proposed TBSRN. The detailed configuration is shown in the supplementary material.

the input is processed in parallel [43]. Under this circumstance, we concatenate a 2-D positional encoding (PE) with the feature map. In practice, we set the dimension of PE for both x-axis (height) and y-axis (width) to d_{PE}. For a grid (m, n) in the image, we calculate its positional encoding for each axis as follows:

$$\begin{aligned}
\mathbf{P}_x[m, n, 2i] &= \sin(m/10000^{2i/d_{\text{PE}}}) \\
\mathbf{P}_x[m, n, 2i+1] &= \cos(m/10000^{2i/d_{\text{PE}}}) \\
\mathbf{P}_y[m, n, 2i] &= \sin(n/10000^{2i/d_{\text{PE}}}) \\
\mathbf{P}_y[m, n, 2i+1] &= \cos(n/10000^{2i/d_{\text{PE}}})
\end{aligned} \tag{1}$$

where $\mathbf{P}_x \in \mathbb{R}^{H \times W \times d_{\text{PE}}}$ and $\mathbf{P}_y \in \mathbb{R}^{H \times W \times d_{\text{PE}}}$. For example, $\mathbf{P}_x[m, n, 2i]$ means the $2i$-th element in the PE of x-axis at the grid (m, n). Then the feature map is concatenated with $\mathbf{P}_x$ and $\mathbf{P}_y$ and flattened to a 1-D sequence, sequentially sent to the Self-Attention Module and the Position-Wise Feed-Forward Module. Specifically, the size of the generated feature map is reshaped to the same size as the input image. Finally, the SR image is generated by pixel shuffling following [44]. In this module, we employ an L2 loss to constrain these two images: $\mathcal{L}_{\text{PSM}} = ||\mathbf{I}_{\text{HR}} - \mathbf{I}_{\text{SR}}||_2^2$, where $\mathbf{I}_{\text{HR}}$ and $\mathbf{I}_{\text{SR}}$ denote the high-resolution image and the super-resolution image, respectively.

3.2. Position-Aware Module

Generally, we mainly focus on the character regions in the text image super-resolution task while paying less attention to backgrounds. However, most existing methods simply employ an L1 loss or an L2 loss focusing on each pixel in the images [30, 44]. In fact, texts in natural scenes are often in company with complicated backgrounds, which are great disturbances to the super-resolution model. Therefore, we employ a Position-Aware Module to highlight character regions with the reference of high-resolution images.

To achieve this aim, we first pre-train a Transformer-based recognition model using synthetic text datasets including Syn90k [13] and SynthText [12], then leveraging its attending regions at each time step as positional clues. The configuration of the pre-trained Transformer is shown in the supplementary material. Given an HR text image, the Transformer outputs a list of attention maps $\mathbf{A}_{\text{HR}} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_l)$, where $\mathbf{a}_i$ denotes the attention map at the i-th time step and l is the length of its text label. Taking advantage of HR images that often have clear gaps between characters, we utilize their attention maps as the label of character regions. The generated SR image is also fed to the Transformer to obtain another list $\mathbf{A}_{\text{SR}}$, the length of which is the same as $\mathbf{A}_{\text{HR}}$. We employ an L1 loss to supervise two attention maps as follows:

$$\mathcal{L}_{\text{POS}} = ||\mathbf{A}_{\text{HR}} - \mathbf{A}_{\text{SR}}||_1 \tag{2}$$

3.3. Content-Aware Module

Clear super-resolution text images can be well identified by recognition models. Given the super-resolution images, we employ a pre-trained Transformer (the same as the one used in the Position-Aware Module) to predict a text sequence. Generally, we can simply leverage a cross-entropy loss to supervise text predictions following [3, 4, 37]. Since the parameters in the pre-trained Transformer are **frozen**, the Content-Aware Module will guide the super-resolution model to generate a more distinguishable text image by backpropagation.

However, we notice that there are some character pairs that look similar in low-resolution condition (*e.g.* "c" and "e" in Figure 1), which is hard for the super-resolution procedure. To tackle this problem, we first train a Variational Autoencoder (VAE) [19] using EMNIST [5] to obtain each character's 2-D latent representation. Details of VAE are introduced in the supplementary material. As demonstrated in Figure 4, positions of similar characters are usually close in the latent space. We denote the alphabet as $\mathcal{A}$, the length of the alphabet as $|\mathcal{A}|$, and the i-th character as $\mathcal{A}_i$. Given this 2-D latent space representation, the Euclidean distance between character $\mathcal{A}_i$ and $\mathcal{A}_j$ is denoted as d_{ij}. And we set their confusable coefficient c_{ij} as $\frac{1}{d_{ij}}$ provided that $i \neq j$. Otherwise, we set c_{ij} to 1 when $i = j$.

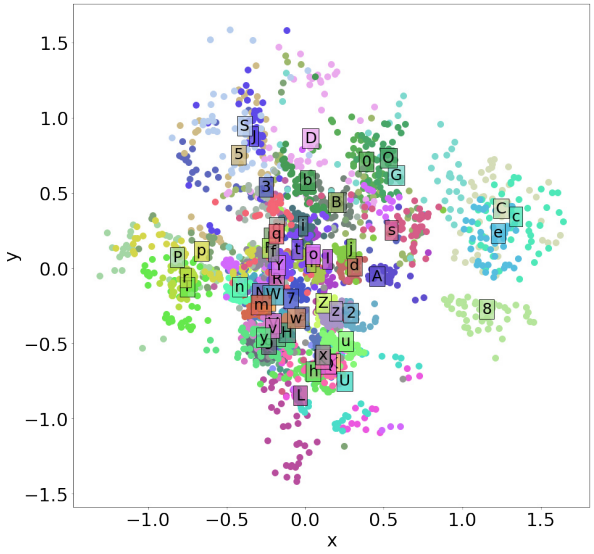


Figure 4. Latent variable clustering over all digits and letters. Some confusable characters are close in the latent space (*e.g.* {"C","c","e"} and {"2","z","Z"}). Each label is placed in the center of the corresponding class.

As shown in Figure 5, assume that at the time step t, the pre-trained Transformer generates an output vector $\mathbf{o} = \{o_1, o_2, ..., o_{|\mathcal{A}|}\}$. For each $o_i \in \mathbb{R}$, the larger the value, the more likely the recognition model is to predict the i-th character at the current time step. If the ground truth is $\mathcal{A}_j$, we calculate its weighted activation a_j as follows:

$$a_j = \frac{\mathrm{e}^{o_j}}{\sum_{i=1}^{|\mathcal{A}|} c_{ij}\mathrm{e}^{o_i}} \tag{3}$$

The content loss $\mathcal{L}_{\text{CON}}$ for all time steps is computed by: $\mathcal{L}_{\text{CON}} = -\sum_t \ln a_{y_t}$, where y_t denotes the ground truth at the t-th time step. Note that when each confusable coefficient is constrained to 1, $\mathcal{L}_{\text{CON}}$ is equal to the vanilla cross-entropy loss. In the following, we will prove its effectiveness. According to Figure 5, the last few layers of the recognition model can be concluded as three layers, including a hidden layer, an output layer, and a softmax layer. At time step t, if the prediction is $\mathcal{A}_i$ while the ground truth is $\mathcal{A}_j$, the gradient is computed as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{CON}}}{\partial w_{ki}} &= \frac{\partial \mathcal{L}_{\text{CON}}}{\partial a_j}\frac{\partial a_j}{\partial o_i}\frac{\partial o_i}{\partial w_{ki}} \\ &= -\frac{1}{a_j}\frac{-c_{ij}\mathrm{e}^{o_i}\mathrm{e}^{o_j}}{(\sum_{m=1}^{|\mathcal{A}|} c_{jm}\mathrm{e}^{o_m})^2}h_k \\ &= \frac{c_{ij}\mathrm{e}^{o_i}}{\sum_{m=1}^{|\mathcal{A}|} c_{jm}\mathrm{e}^{o_m}}h_k\end{aligned} \tag{4}$$

If $\mathcal{A}_i$ and $\mathcal{A}_j$ look similar, c_{ij} will be a high value (*i.e.* much greater than 1), which results in a numerically higher

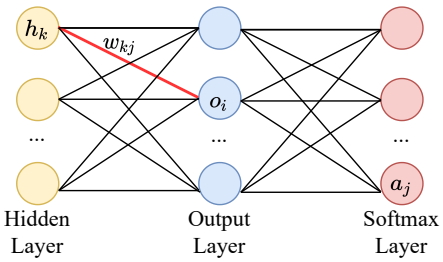


Figure 5. The last few layers of generic recognition models.

gradient for backpropagation (*i.e.* get more punishment). Compared with vanilla cross-entropy loss, the weighted cross-entropy loss will pay more attention to those confusable characters.

3.4. Overall Loss Function

The overall loss function $\mathcal{L}$ can be calculated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{PSM}} + \lambda_{\text{POS}}\mathcal{L}_{\text{POS}} + \lambda_{\text{CON}}\mathcal{L}_{\text{CON}} \tag{5}$$

where λ_{POS} and λ_{CON} are hyperparameters to balance three terms. See the supplementary material about the discussion on choices between the L1 and L2 loss for $\mathcal{L}_{\text{PSM}}$ and $\mathcal{L}_{\text{POS}}$.

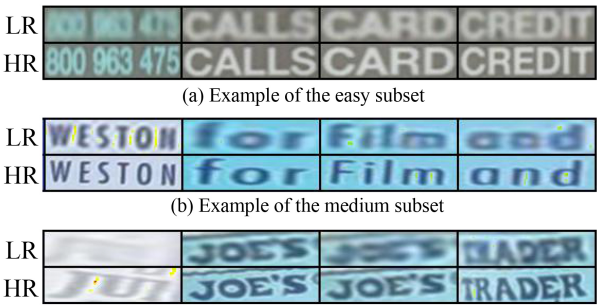
4. Experiments

In this section, we conduct experiments to verify: 1) The effectiveness of each component in STT. 2) STT's performance on TextZoom and its ability as a preprocessor on STR benchmarks. At last, we display some failure cases.

Introduction of TextZoom. The images in TextZoom [44] originate from two single image super-resolution datasets, including RealSR [1] and SR-RAW [48]. These datasets contain LR-HR pairs which are taken by digital cameras in real scenes. TextZoom contains $17,367$ LR-HR pairs for training and $4,373$ pairs for testing. According to different focal lengths of digital cameras, the test set is divided into three subsets, including $1,619$ pairs for the easy subset, $1,411$ pairs for the medium subset, and $1,343$ pairs for the hard subset. LR images are resized to 16×64 and HR images are resized to 32×128. The examples of three subsets are demonstrated in Figure 6.

Evaluation Metrics. To calculate the recognition accuracy, we remove all punctuations and convert uppercase letters to lowercase letters, which follows the same setting of [44] for a fair comparison. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [45] are used to evaluate the quality of SR images.

While the vanilla PSNR and SSIM consider all pixels in the image, we mainly focus on text regions in this text image super-resolution task. Therefore, we propose two novel metrics, including Text Region PSNR (TR-PSNR) and Text Region SSIM (TR-SSIM). When calculating these two metrics, we only take the pixels in the text region into consideration. Benefiting from the text-level bounding box labels in SynthText [12], we can obtain rough text regions (refer to white pixels in Figure 7) by training a segmentation model based on U-Net [35]. The details of the segmentation model are shown in the supplementary material.



(a) Example of the easy subset

(b) Example of the medium subset

(c) Example of the hard subset

Figure 6. Examples of the three subsets in TextZoom. LR images are upsampled to the same size as HR images using the bicubic interpolation. The text becomes more blurred when the difficulty increases, which are hard to recognize even for human eyes.

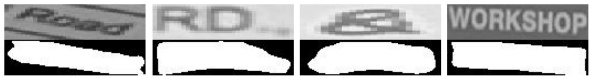


Figure 7. Examples of masks generated by the segmentation model based on U-Net. Masks can roughly represent text regions.

Implementation Details. Our method is implemented in PyTorch. All experiments are conducted on four NVIDIA TITAN Xp GPUs with 12GB memory. The model is trained using the Adam [18] optimizer. The batch size is set to 80. The learning rate is set to 0.0001. d_{PE} is set to 32. Through ablation studies, we set the trade-off weight of λ_{POS} as 10 and λ_{CON} as 0.0005. The parameters in the pre-trained Transformer are **frozen**. We use official PyTorch code and released pre-trained models of CRNN [37][1], ASTER [38][2], and MORAN [26][3].

4.1. Ablation Study

In this section, we will evaluate the effectiveness of each component, including the backbone, the Position-Aware Module, the Content-Aware Module, and $\mathcal{L}_{\text{PSM}}$. Ablation studies are conducted on TextZoom and recognition accuracy is computed by the pre-trained CRNN [37].

Ablation Study on Backbone. We compare the TBSRN with other super-resolution backbones, including SRCNN [7], SRResNet [20], as well as TSRN [44]. As shown in Table 1, benefiting from the 2-D attention map, TBSRN-5 outperforms TSRN [44] by 1.8% on average accuracy. More-

[1]https://github.com/meijieru/crnn.pytorch
[2]https://github.com/ayumiymk/aster.pytorch
[3]https://github.com/Canjie-Luo/MORAN_v2

Backbone	Easy	Medium	Hard	Average
BICUBIC	36.4%	21.1%	21.1%	26.8%
SRCNN [7]	41.1%	22.3%	22.0%	29.2%
SRResNet [20]	45.2%	32.6%	25.5%	35.1%
TSRN [44]	52.5%	38.2%	31.4%	41.4%
TBSRN-1	51.5%	35.5%	30.2%	39.8%
TBSRN-2	54.2%	37.4%	30.2%	41.4%
TBSRN-3	**55.0%**	37.9%	31.0%	42.1%
TBSRN-4	53.8%	37.8%	31.3%	41.7%
TBSRN-5	54.2%	**40.6%**	**32.7%**	**43.2%**
TBSRN-6	53.8%	39.4%	31.4%	42.3%
TBSRN-5 w/o PE	50.0%	33.4%	28.2%	37.9%

Table 1. Ablation study on the backbone. "BICUBIC" means LR images are directly upsampled by the bicubic interpolation.

λ_{POS}	Easy	Medium	Hard	Average
0	54.2%	40.6%	32.7%	43.2%
0.1	53.6%	39.3%	31.7%	42.3%
1	55.8%	41.0%	33.0%	44.0%
10	**57.7%**	**43.0%**	**33.6%**	**45.6%**
100	57.0%	42.3%	33.4%	45.0%

Table 2. Ablation study on the Position-Aware Module.

over, the performance degrades by 0.9% on average when using more blocks. We also observe that the model in the absence of a 2-D PE degrades severely (drop 5.3% on average accuracy), which reflects the significance of positional clues in the backbone. Therefore, we utilize TBSRN-5 to conduct the following ablation studies.

Ablation Study on Position-Aware Module. To verify the effectiveness of this module, we explore λ_{POS} from $\{0, 0.1, 1, 10, 100\}$. Specifically, when $\lambda_{\text{POS}} = 0$, there does not exist any supervision on attention maps. The results are shown in Table 2. Compared with the baseline (λ_{POS} = 0), the average accuracy on CRNN increases by 2.4% when λ_{POS} = 10, which shows its superiority as position-level guidance. We set λ_{POS} to 10 in the following experiments. More visual results of this modules are shown in the supplementary material.

Ablation Study on Content-Aware Module. We explore λ_{CON} from $\{0, 0.0001, 0.0005, 0.001, 0.01\}$ and the results are shown in Table 3. When $\lambda_{\text{CON}} = 0.0005$, the model outperforms all its counterparts. Compared with the baseline ($\lambda_{\text{CON}} = 0$), the average accuracy increases by 1.5%. Furthermore, we observe that the model with a weighted cross-entropy loss boosts the accuracy by 1.0%, indicating the weights can indeed help pave the way for distinguishing confusable characters. We set λ_{CON} to 0.0005 in the following experiments. More visual results of this modules are shown in the supplementary material.

Ablation Study on $\mathcal{L}_{\text{PSM}}$. Since we have employed $\mathcal{L}_{\text{POS}}$ and $\mathcal{L}_{\text{CON}}$ to help the model focus on character regions, we also investigate whether $\mathcal{L}_{\text{PSM}}$ is necessary for the super-resolution task. We conduct an ablation study on $\mathcal{L}_{\text{PSM}}$ and the results are shown in Table 4. Interestingly, the model without $\mathcal{L}_{\text{PSM}}$ degrades severely on recognition accuracy and quality metrics. Several examples are shown in the supplementary material. Through the visualization, we observe that $\mathcal{L}_{\text{PSM}}$ provides the supervision about color and pixel-level character outlines. Moreover, the model tends to generate green-toned text images in the absence of $\mathcal{L}_{\text{PSM}}$, which are not friendly to human eyes. Although the model generates some images that can be successfully identified by CRNN, there is a huge gap between SR images and HR images in terms of appearance. Therefore, $\mathcal{L}_{\text{PSM}}$ plays an important part in the super-resolution task.

λ_{CON}	WCE	Easy	Medium	Hard	Average
0	-	57.7%	43.0%	33.6%	45.6%
0.0001	-	58.3%	44.2%	35.0%	46.6%
	✓	58.6%	46.6%	35.2%	47.5%
0.0005	-	59.1%	44.7%	35.0%	47.1%
	✓	**59.6%**	**47.1%**	**35.3%**	**48.1%**
0.001	-	58.3%	44.4%	33.4%	46.2%
	✓	59.0%	45.6%	34.7%	47.2%
0.01	-	58.1%	44.4%	33.4%	46.1%
	✓	58.5%	44.5%	33.4%	46.3%

Table 3. Ablation study on the Content-Aware Module. "WCE" denotes the weighted cross-entropy loss.

$\mathcal{L}_{\text{PSM}}$	$\mathcal{L}_{\text{POS}}$	$\mathcal{L}_{\text{CON}}$	Easy	Medium	Hard	Average
✓	✓	✓	**59.6%**	**47.1%**	**35.3%**	**48.1%**
-			44.7%	35.2%	28.3%	36.6%

$\mathcal{L}_{\text{PSM}}$	$\mathcal{L}_{\text{POS}}$	$\mathcal{L}_{\text{CON}}$	PSNR	TR-PSNR	SSIM	TR-SSIM
✓	✓	✓	**23.82**	**24.60**	**0.8660**	**0.9003**
-			4.97	6.31	0.1968	0.3399

Table 4. Ablation study on $\mathcal{L}_{\text{PSM}}$ with respect to recognition accuracy (top) and image quality (below).

4.2. Experimental Results

In this section, we assess STT's performance on TextZoom and conduct experiments on scene text recognition benchmarks to verify STT's ability as a preprocessor.

Results on TextZoom. We compare our model with other existing super-resolution models on three recognition models, including CRNN [37], ASTER [38], and MORAN [26] (see Table 5(a)). Our model outperforms its counterparts on every recognizer. Compared with vanilla TSRN (without $\mathcal{L}_{\text{POS}}$ and $\mathcal{L}_{\text{CON}}$) [44], our model boosts average accuracy by 1.8% on ASTER, 3.0% on MORAN, and 6.7% on CRNN. Furthermore, we add the Position-Aware Module and the Content-Aware Module to other backbones and observe that the two newly added modules also boost the performance. For example, SRResNet [20] armed with these two modules boosts average accuracy by 6% on CRNN. As shown

(a) Comparison of the recognition accuracy.

Backbone	$\mathcal{L}_{POS}$	$\mathcal{L}_{CON}$	ASTER [38]				MORAN [26]				CRNN [37]			
			Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
BICUBIC	-	-	64.7%	42.4%	31.2%	47.2%	60.6%	37.9%	30.8%	44.1%	36.4%	21.1%	21.1%	26.8%
SRCNN [7]	-	-	70.6%	44.0%	31.5%	50.0%	63.9%	40.0%	29.4%	45.6%	41.1%	22.3%	22.0%	29.2%
	✓	-	69.0%	46.9%	32.8%	50.8%	64.5%	42.7%	31.1%	47.2%	40.6%	24.7%	22.6%	29.9%
	✓	✓	70.2%	49.4%	32.5%	51.9%	66.2%	44.4%	31.3%	48.4%	41.7%	25.4%	23.1%	30.7%
SRResNet [20]	-	-	69.4%	50.5%	35.7%	53.0%	66.0%	47.1%	33.4%	49.9%	45.2%	32.6%	25.5%	35.1%
	✓	-	74.8%	56.2%	38.6%	57.7%	70.2%	53.4%	37.0%	54.6%	47.7%	35.6%	27.1%	37.5%
	✓	✓	74.2%	57.3%	38.5%	57.8%	71.1%	54.4%	37.0%	55.2%	52.3%	39.6%	29.3%	41.1%
TSRN [44]	-	-	75.1%	56.3%	40.1%	58.3%	70.1%	55.3%	37.9%	55.4%	52.5%	38.2%	31.4%	41.4%
	✓	-	74.2%	57.7%	40.0%	58.4%	71.5%	55.5%	38.4%	56.2%	53.9%	39.6%	31.7%	42.5%
	✓	✓	74.3%	59.7%	39.6%	58.9%	72.3%	55.6%	39.8%	56.9%	54.3%	40.4%	31.7%	42.9%
TBSRN	-	-	75.2%	56.7%	40.2%	58.5%	71.1%	55.2%	39.5%	56.3%	54.2%	40.6%	32.7%	43.2%
	✓	-	**76.1%**	58.9%	**41.6%**	60.0%	73.8%	56.2%	**40.9%**	58.0%	57.7%	43.0%	33.6%	45.6%
	✓	✓	75.7%	**59.9%**	**41.6%**	**60.1%**	**74.1%**	**57.0%**	40.8%	**58.4%**	**59.6%**	**47.1%**	**35.3%**	**48.1%**

(b) Comparison of the image quality.

Backbone	$\mathcal{L}_{POS}$	$\mathcal{L}_{CON}$	PSNR			TR-PSNR			SSIM			TR-SSIM		
			Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
BICUBIC	-	-	22.35	18.98	19.39	23.58	20.25	19.62	0.7884	0.6254	0.6592	0.8474	0.7139	0.7433
SRCNN [7]	-	-	23.13	19.57	19.56	23.83	20.26	20.44	0.8152	0.6425	0.6833	0.8613	0.7208	0.7554
	✓	-	22.49	19.62	19.55	23.24	20.40	20.43	0.7962	0.6231	0.6668	0.8480	0.7099	0.7451
	✓	✓	22.51	**19.66**	19.44	23.25	**20.41**	20.34	0.7952	0.6288	0.6694	0.8475	0.7131	0.7463
SRResNet [20]	-	-	20.65	18.90	19.53	21.62	19.58	20.36	0.8176	0.6324	0.7060	0.8597	0.7089	0.7710
	✓	-	23.03	19.36	19.71	23.89	20.06	20.60	0.8188	0.6248	0.6873	0.8663	0.7090	0.7613
	✓	✓	23.16	19.25	19.94	24.05	19.93	20.80	0.8311	0.6314	0.6999	0.8758	0.7122	0.7700
TSRN [44]	-	-	22.95	19.26	19.76	23.80	19.88	20.62	0.8562	0.6596	0.7285	0.8895	0.7274	0.7854
	✓	-	23.26	19.01	19.75	24.00	19.89	20.77	0.8441	0.6578	0.7103	0.8875	0.7196	0.7860
	✓	✓	23.34	19.16	19.81	24.22	19.84	20.70	0.8466	0.6379	0.7023	0.8867	0.7157	0.7710
TBSRN	-	-	**24.13**	19.08	**20.09**	**24.99**	19.72	**20.90**	**0.8729**	0.6455	0.7452	**0.9031**	0.7113	**0.7995**
	✓	-	23.46	18.86	19.85	24.31	19.55	20.72	0.8612	**0.6639**	0.7252	0.8953	**0.7299**	0.7870
	✓	✓	23.82	19.17	19.68	24.60	19.73	20.65	0.8660	0.6533	**0.7490**	0.9003	0.7234	**0.7995**

Table 5. Comparison with the existing methods in terms of the recognition accuracy and the image quality on TextZoom.

in Table 5(b), our model shows comparable performance on four image quality metrics with other existing methods. Although the performance does not reach the best, it is not so important compared to accuracy in this task. We visualize several examples in Figure 8. Compared with other methods, our method pays more attention to the character-level details. Furthermore, the model is robust to the inclined or curved text images (see the last three columns). The experimental results on more recognizers, including NRTR [36], SEED [33] as well as AutoSTR [47], are shown in the supplementary material. The computational cost is also discussed in the supplementary material.

Results on Scene Text Recognition Benchmarks. In this section, we validate the ability of our model as a preprocessor on scene text recognition benchmarks. Specifically, we prepare LR images in two settings: 1) Manually downsample and degrade all original images. 2) Pick low-resolution images from existing benchmarks.

For the first setting, we choose all images in IC13 [16], IC15 [15], as well as CT80 [34] to conduct experiments. Since image styles and distribution of labels in these benchmarks are quite different from those in TextZoom, there exists a domain shift problem that is challenging for the proposed STT. The introduction of these benchmarks is shown in the supplementary material. We first resized the images to 16×64 and sequentially employ Gaussian blur kernels with different radii to blur these images (see Figure 9), and use official CRNN [37] to evaluate recognition accuracy. The results are shown in Table 6. If preprocessing the LR images, we employ STT pre-trained on TextZoom to upsample it to 32×128. Otherwise, images are directly upsampled with the bilinear interpolation. The effect of STT becomes more powerful when we manually blur the input images. For example, the accuracy increases by 7.8% on IC13, 3.2% on IC15, and 7.3% on CT80 when the radius reaches 3, which shows its superiority as a preprocessor.

For the second setting, we choose two robust text recognizers including NRTR [36][4] and SEED [33][5]. Compared with IC15, other benchmarks have relatively higher resolution, so we only test LR text images in IC15. To validate the ability of the proposed STT, we first pick all text images with low resolution (smaller than 16×64) in IC15 and resize them to 16×64, resulting in 352 samples. Sequentially, we utilize the proposed STT pre-trained on TextZoom to upsample them to 32×128. After the pre-processing, the accuracy on the 352 text images increases by 5.7% (65.6%

[4]https://github.com/Belval/NRTR
[5]https://github.com/Pay20Y/SEED

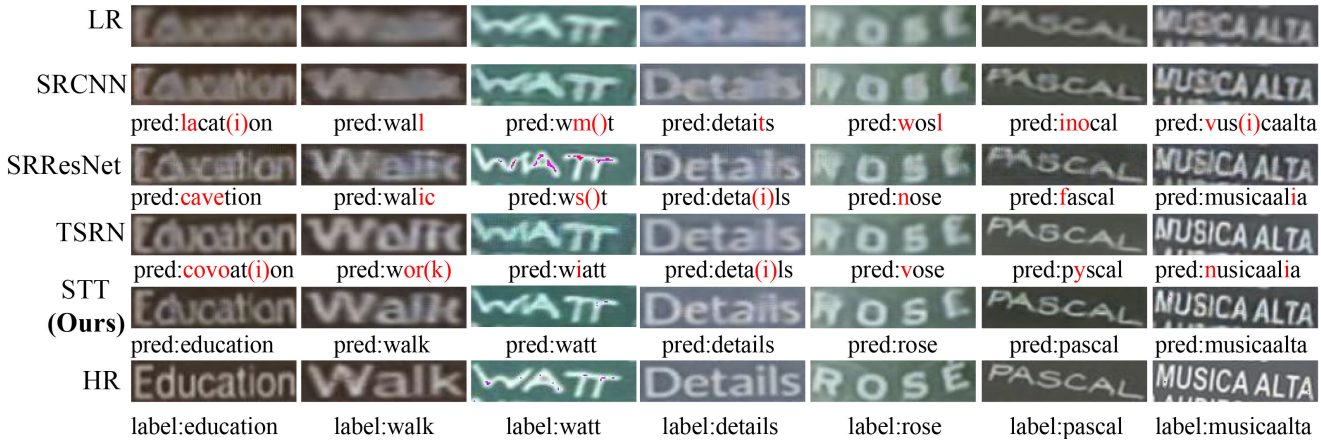


Figure 8. Examples of the generated images in TextZoom. The SR images generated from STT are clearer than other methods in terms of character-level details. Moreover, our model is robust on inclined or curved text images (refer to the last three columns).

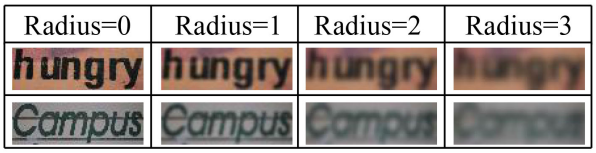


Figure 9. Images processed by Gaussian blur kernels.

R	P	IC13 [16]	IC15 [15]	CT80 [34]	Average
0	-	**76.4%**	53.7%	47.0%	60.9%
	✓	75.8%	**56.8%**	**47.2%**	**62.4%**
1	-	71.2%	38.5%	46.5%	50.4%
	✓	**72.4%**	**49.4%**	**48.0%**	**57.2%**
2	-	56.7%	13.4%	36.5%	30.3%
	✓	**64.0%**	**21.6%**	**41.0%**	**37.9%**
3	-	43.0%	5.2%	26.4%	20.1%
	✓	**50.8%**	**8.4%**	**33.7%**	**25.2%**

Table 6. Results on scene text recognition benchmarks. "R" denotes the radius. "P" means whether to preprocess with STT.

to 71.3%) for NRTR and 4.8% (71.0% to 75.8%) for SEED. For the full set of IC15, the accuracy on the 1,811 text images increases by 1.1% (76.5% to 77.6%) for NRTR and 0.9% (79.7% to 80.6%) for SEED. So the proposed STT is capable of boosting performance on robust text recognizers.

4.3. Failure Cases

As shown in Figure 10, we visualize several failure cases from TextZoom. We notice that the model is hard to handle the LR images with long and small texts. When the text image has a complicated background or occlusion, the performance of our model is subpar. Furthermore, artistic fonts and handwriting texts bring difficulties to the model. We also observe that our model is hard to tackle those images whose labels have not appeared in the training set.

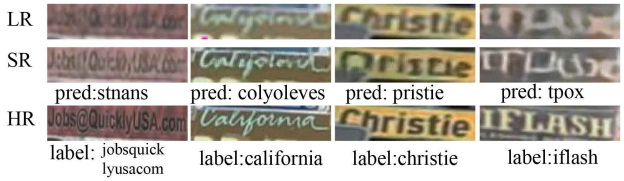


Figure 10. Failure cases in TextZoom. Long texts , artistic fonts, occlusion, and unseen texts indeed bring difficulties for our STT.

5. Conclusion

In this paper, we put forward a text-focused super-resolution model, called Scene Text Telescope, aiming to excavate text-specific properties. The proposed backbone termed TBSRN utilizes the self-attention mechanism to tackle irregular text images. The Position-Aware Module and Content-Aware Module help the model pay more attention to the position and the content of each character without bringing additional time overhead. Furthermore, the weighted cross-entropy loss alleviates the difficulty caused by confusable characters. With these components, the generated images are more distinguishable for recognition models. Hence, the proposed method shows its superiority in upsampling low-resolution scene text images.

6. Acknowledgements

This research was supported in part by STCSM Projects (20511100400, 20511102702), Shanghai Municipal Science and Technology Major Projects (2017SHZDZX01, 2018SHZDZX01), Shanghai Research and Innovation Functional Program (17DZ2260900), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, and ZJLab. We also thank Jianqi Ma for discussion and the proofreading provided by the teammates of FudanOCR.

References

[1] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, 2019.

[2] David Capel and Andrew Zisserman. Super-resolution enhancement of text image sequences. In *ICPR*, volume 1, pages 600–605, 2000.

[3] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *ICCV*, pages 5076–5084, 2017.

[4] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *CVPR*, pages 5571–5579, 2018.

[5] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *IJCNN*, pages 2921–2926, 2017.

[6] Gerald Dalley, Bill Freeman, and Joe Marks. Single-frame text super-resolution: A bayesian approach. In *ICIP*, volume 5, pages 3295–3298, 2004.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014.

[8] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211*, 2015.

[9] J-L Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing*, 2(2):291–298, 1994.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006.

[11] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016.

[13] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116(1):1–20, 2016.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *ICDAR*, pages 1156–1160, 2015.

[16] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493, 2013.

[17] Vijeta Khare, Palaiahnakote Shivakumara, Chee Seng Chan, Tong Lu, Liang Kim Meng, Hon Hock Woon, and Michael Blumenstein. A novel character segmentation-reconstruction approach for license plate recognition. *Expert Systems with Applications*, 131:219–239, 2019.

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

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[21] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *AAAI*, volume 33, pages 8610–8617, 2019.

[22] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *AAAI*, volume 33, pages 8714–8721, 2019.

[23] Hoang Danh Liem, Nguyen Duc Minh, Nguyen Bao Trung, Hoang Tien Duc, Pham Hoang Hiep, Doan Viet Dung, and Dang Hoang Vu. Fvi: An end-to-end vietnamese identification card detection and recognition in images. In *NICS*, pages 338–340, 2018.

[24] Wei Liu, Chaofeng Chen, and Kwan-Yee K Wong. Char-net: A character-aware neural network for distorted scene text recognition. In *AAAI*, volume 1, page 4, 2018.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

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

[27] Jianqi Ma. Rrpn++: Guidance towards more accurate scene text detection. *arXiv preprint arXiv:2009.13118*, 2020.

[28] Arien Mack. Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science*, 12(5):180–184, 2003.

[29] Arien Mack, Irvin Rock, et al. *Inattentional blindness*. MIT press, 1998.

[30] Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *ECCV*, 2020.

[31] Andrey V Nasonov and Andrey S Krylov. Text images superresolution and enhancement. In *ICISP*, pages 617–620, 2012.

[32] Ram Krishna Pandey, K Vignesh, AG Ramakrishnan, et al. Binary document image super resolution for improved readability and ocr performance. *arXiv preprint arXiv:1812.02475*, 2018.

[33] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *CVPR*, pages 13528–13537, 2020.

[34] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[36] Fenfen Sheng, Zhineng Chen, and Bo Xu. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *ICDAR*, pages 781–786, 2019.

[37] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2016.

[38] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048, 2018.

[39] Sergio Montazzolli Silva and Claudio Rosito Jung. Real-time license plate detection and recognition using deep convolutional neural networks. *Journal of Visual Communication and Image Representation*, page 102773, 2020.

[40] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, pages 1920–1927, 2013.

[41] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, pages 111–126, 2014.

[42] Hanh TM Tran and Tien Ho-Phuoc. Deep laplacian pyramid network for text images super-resolution. In *RIVF*, pages 1–6, 2019.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[44] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *ECCV*, 2020.

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[46] Lu Yang, Peng Wang, Hui Li, Zhen Li, and Yanning Zhang. A holistic representation guided attention network for scene text recognition. *Neurocomputing*, 414:67–75, 2020.

[47] Hui Zhang, Quanming Yao, Mingkun Yang, Yongchao Xu, and Xiang Bai. Efficient backbone search for scene text recognition. In *ECCV*, 2020.

[48] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, pages 3762–3770, 2019.