

A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution

Jianqi Ma¹ Zhetong Liang² Lei Zhang¹

¹The Hong Kong Polytechnic University; ²OPPO Research

{csjma, cslzhang}@comp.polyu.edu.hk, zhetongliang@163.com

Abstract

Scene text image super-resolution aims to increase the resolution and readability of the text in low-resolution images. Though significant improvement has been achieved by deep convolutional neural networks (CNNs), it remains difficult to reconstruct high-resolution images for spatially deformed texts, especially rotated and curve-shaped ones. This is because the current CNN-based methods adopt locality-based operations, which are not effective to deal with the variation caused by deformations. In this paper, we propose a CNN based Text ATTention network (TATT) to address this problem. The semantics of the text are firstly extracted by a text recognition module as text prior information. Then we design a novel transformer-based module, which leverages global attention mechanism, to exert the semantic guidance of text prior to the text reconstruction process. In addition, we propose a text structure consistency loss to refine the visual appearance by imposing structural consistency on the reconstructions of regular and deformed texts. Experiments on the benchmark TextZoom dataset show that the proposed TATT not only achieves state-of-the-art performance in terms of PSNR/SSIM metrics, but also significantly improves the recognition accuracy in the downstream text recognition task, particularly for text instances with multi-orientation and curved shapes. Code is available at <https://github.com/mjq11302010044/TATT>.

1. Introduction

The text in an image is an important source of information in our daily life, which can be extracted and interpreted for different purposes. However, scene text images often encounter various quality degradation during the imaging process, resulting in low resolution and blurry structures. This problem significantly impairs the performance of the downstream high-level recognition tasks, including scene text detection [23, 46], optical character recognition (OCR) and scene text recognition [21, 31, 32]. Thus, it is neces-



Figure 1. SR recovery of different models on rotated and curve-shaped text images. ‘R’, ‘P’ and ‘S’ stand for recognition, PSNR and SSIM results. Characters in red are missing or wrong.

sary to increase the resolution as well as enhance the visual quality of scene text images.

In the past few years, many scene text image super-resolution (STISR) methods have been developed to improve the image quality of text images, with notable progress obtained by deep-learning-based methods [4, 9, 35, 36, 41]. By using a dataset of degraded and original text image pairs, a deep convolutional neural network (CNN) can be trained to super-resolve the text image. With strong expressive capability, CNNs can learn various priors from data and demonstrate much strong performance. A recent advance is the TPGSR model [22], where the semantics of the text are firstly recognized as prior information and then used to guide the text reconstruction process. With the high-level prior information, TPGSR can restore the semantically correct text image with compelling visual quality.

Despite the great progress, many CNN-based methods still have difficulty in dealing with spatially-deformed text images, including those with rotation and curved shape. Two examples are shown in Fig. 1, where the text in the left image has rotation and the right one has a curved shape. One can see that the current representative methods, includ-

ing TSRN [35] and TPGSR [22], produce blurry texts with semantically incorrect characters. This is because the architectures in current works mainly employ locality-based operations like convolution, which are not effective in capturing the large position variation caused by the deformations. In particular, the TPGSR model adopts a simplistic approach to utilize the text prior: it merely merges text prior with image feature by convolutions. This arrangement can only let the text prior interact with the image feature within a small local range, which limits the effect of text prior on the text reconstruction process. Based on this observation, some globality-based operations (*e.g.*, attention) should be employed to capture long range correlation in the text image for better STISR performance.

In this paper, we propose a novel architecture, termed Text ATTention network (TATT), for spatial deformation robust text super resolution. Similar to TPGSR, we first employ a text recognition module to recognize the character semantics as text prior (TP). Then we design a transformer-based module termed TP Interpreter to enforce global interaction between the text prior and the image feature. Specifically, the TP Interpreter operates cross attention between the text prior and the image feature to capture long-range correlation between them. The image feature can then receive rich semantic guidance in spite of the spatial deformation, leading to improved text reconstruction. To further refine the text appearance under spatial deformation, we design a text structure consistency loss, which measures the structural distance between the regular and deformed texts. As can be seen in Fig. 1, the characters recovered by our method show better visual quality with correct semantics.

Overall, our contributions can be summarized as follows:

- We propose a novel method to align the text prior with the spatially-deformed text image for better SR recovery by using CNN and Transformer.
- We propose a text structure consistency loss to enhance the robustness of text structure recovery from spatially-deformed low-resolution text images.
- Our proposed model not only achieves state-of-the-art performance on the TextZoom dataset in various evaluation metrics, but also exhibits outstanding generalization performance in recovering orientation-distorted and curve-shaped low-resolution text images.

2. Related Works

2.1. Single Image Super Resolution

Single image super resolution (SISR) aims at recovering a high-resolution (HR) image from a given low-resolution (LR) input image. The traditional methods design hand-crafted image priors for this task, including statistical prior [11], self-similarity prior [24] and sparsity prior [40]. The recent deep-learning-based methods train

convolutional neural networks (CNNs) to address the SISR task and achieve leading performance. The seminal work SRCNN [8] adopts a three-layer CNN to learn the SR recovery. Later on, more complex CNN architectures have been developed to upgrade the SISR performance, *e.g.*, residual block [19], Laplacian pyramid [17], dense connections [44] and channel attention mechanism [43]. Recently, generative adversarial networks have been employed in SISR to achieve photo-realistic results [5, 18, 37].

2.2. Scene Text Image Super Resolution (STISR)

Different from the general purposed SISR that works on natural scene images, STISR focuses on scene text images. It aims to not only increase the resolution of text image, but also reconstruct semantically correct texts that can benefit the down-stream recognition task. The early methods directly adopt the CNN architectures from SISR for the task of STISR. In [9], Dong *et al.* extended SRCNN [8] to text images, and obtained the best performance in ICDAR 2015 competition [27]. PlugNet [25] adopts a pluggable super-resolution unit to deal with LR images in feature domain. TextSR [36] utilizes the text perceptual loss to generate the desired HR images to benefit the text recognition.

To address the problem of STISR on real-world scenes, Wang *et al.* [35] built a real-world STISR image dataset, namely the TextZoom, where the LR and HR text image pairs were cropped from real-world SISR datasets [2, 42]. They also proposed TSRN [35] to use the sequential residual block to exploit the semantic information in internal features. SCGAN [39] employs a multi-class GAN loss to supervise the STISR model for more perceptual-friendly face and text images. Further, Quan *et al.* [29] proposed a cascading model for recovering blurry text images in high-frequency domain and image domain collaboratively. Chen *et al.* [4] and Zhao *et al.* [45] enhanced the network block structures to improve the STISR performance by self-attending the image features and attending channels.

2.3. Scene Text Recognition

Scene text recognition aims to extract text content from the input images. Some early approaches tend to recognize each character first and then interpret the whole word [12, 14], while some others regard the text image as a whole and performing word-level classification [13]. Considering text recognition as an image-to-sequence problem, CRNN [31] extracts image features and uses the recurrent neural networks to model the semantic information. It is trained with CTC [10] loss to align the predicted sequence and the target sequence. Recently, attention-based methods achieve a great progress due to the robustness in extracting text against shape variance of text images [6, 7]. Despite the great performance achieved by the recent methods, it is still difficult to recognize the text in low-resolution images.

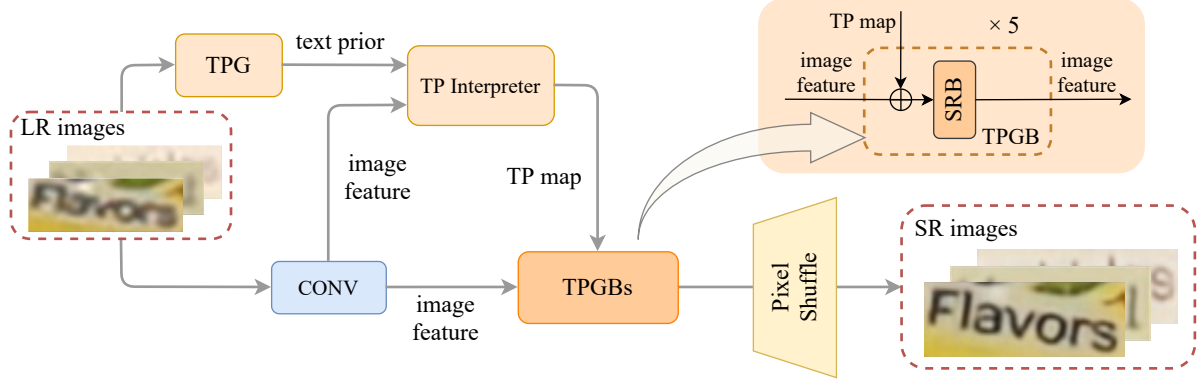


Figure 2. Architecture of our proposed TATT network for STISR. TPGB, TPG and SRB are short for text prior guided blocks, TP Generator and Sequential-Recurrent Blocks, respectively, while \oplus means the element-wise addition.

Therefore, we aim to solve the problem of high-resolution text image restoration for better recognition in this paper.

3. Methodology

3.1. Overall Architecture

The pipeline of our TATT network is shown in Fig. 2. It takes low-resolution (LR) text images $Y \in \mathbb{R}^{h \times w \times 3}$ as input, which is processed in the following two paths. In the first path, the input images are sent into a TP Generator (TPG) to predict the recognition probability sequence as text prior f_P (similar to [22]). This process can be denoted as $f_P = TPG(Y)$. $f_P \in \mathbb{R}^{l \times |\mathcal{A}|}$ is an l -length sequence composed of categorical probability vectors with size $|\mathcal{A}|$. \mathcal{A} denotes the character set which is composed of ‘0’ to ‘9’, ‘a’ to ‘z’ and a blank class (37 in total). The second path extracts image features $f_I \in \mathbb{R}^{h \times w \times c}$ from the input LR image Y by a 9×9 convolution layer (we denote this process as $f_I = Conv(Y)$).

Then, the text prior f_P and the image feature f_I are passed to the TP Interpreter $TPI(\cdot)$ to calculate a TP map $f_{TM} \in \mathbb{R}^{h \times w \times c}$, which is denoted as $f_{TM} = TPI(f_P, f_I)$. The TP Interpreter computes the correlation between the text prior f_P and image feature f_I , and assigns the semantic guidance in f_P to the corresponding location in the spatial domain to guide the final SR text recovery. The resultant TP map f_{TM} is a modulating map which can be used to enhance the semantics-specific part of the image feature.

Finally, the TP map f_{TM} and the image feature f_I are passed into a reconstruction module. This module includes 5 Text-Prior Guided Blocks (TPGBs) that progressively fuse f_{TM} and f_I , and a final Pixel-Shuffle layer to increase the resolution. Each of the 5 TPGBs firstly merges f_{TM} and f_I by element-wise addition, followed by a Sequential-Recurrent Block (SRB) [35] to reconstruct the high-resolution image feature. The output of this module is the super-resolved (SR) text image.

3.2. TP Interpreter

In the proposed architecture, the crucial part lies in the design of TP Interpreter (TPI). The TP Interpreter aims to interpret the text prior f_P to the image feature f_I so that the influence of the semantics guidance can be exerted to the correlated spatial position in the image feature domain. One intuitive idea is to enlarge f_P to the shape of f_I and then merge them by convolution. Since the convolution operation has a small effective range, the semantics of f_P cannot be assigned to the distant spatial location in f_I , especially in the case of spatially-deformed text. Thus, we turn to design a Transformer-based TP Interpreter with attention mechanism to enforce global correlation between text prior f_P and the image feature f_I .

As shown in Fig. 3, the proposed TP Interpreter consists of an Encoder part and a Decoder part. The Encoder encodes the text prior f_P by performing correlation between the semantics of each character in f_P and outputs the context-enhanced feature f_E . The decoder performs cross attention between f_E and f_I to interpret the semantic information to the image feature.

Encoder. The Encoder takes the text prior f_P as input and project it to C channels to match the image feature channel. Since the input text prior is processed in parallel in the encoder, the model is not aware of the semantic order in TP. We thus encode the position by adding the Fixed Positional Encoding (FPE) to f_P in an element-wise manner before feeding it into the encoder. Note that we adopt Sinusoidal Positional Encoding [34] as our FPE in this paper. After encoding the position, the text prior is passed into the encoder module. The encoder has a Multi-head Self Attention (MSA) layer and a FeedForward Network (FFN) layer [34]. Skip connection is deployed between the current layer and the previous layer to enable residual learning. The MSA layer performs global correlation between the semantic elements in text prior f_P , resulting in a contextually enhanced TP feature $f_E \in \mathbb{R}^{l \times c}$ for later computation. Due

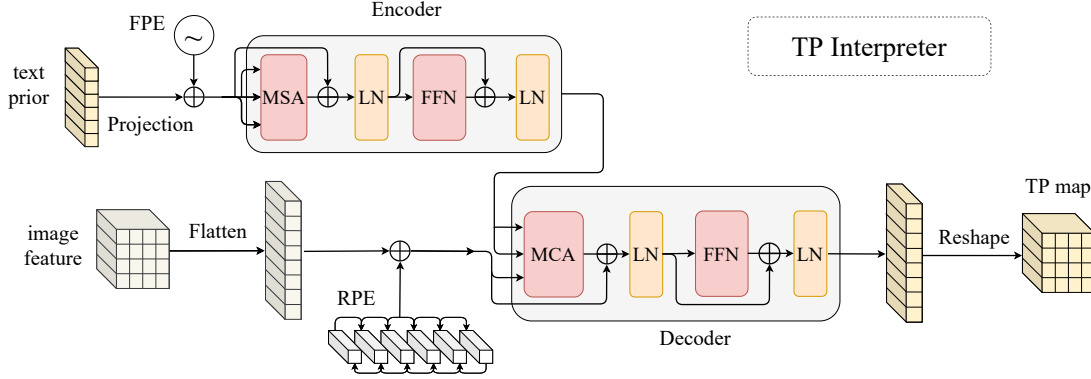


Figure 3. Architecture of TP Interpreter. ‘MSA’, ‘LN’, ‘MCA’ and ‘FFN’ namely mean the Multi-head Self-Attention, Layer-Norm, Multi-head Cross-Attention and Feed-Forward Network layers, while ‘FPE’ and ‘RPE’ refer to the Fixed Positional Encoding and Recurrent Positional Encoding. While \oplus means the element-wise addition.

to the space limit, the description of MSA and FFN is omitted. One can refer to [34] for details.

Decoder. The decoder module accepts the output from the encoder module f_E and image feature f_I to perform global cross attention. Similar to the setting in encoder, we firstly add a position encoding to f_I to incorporate position information. We design a recurrent positional encoding (RPE) to better encode the bias contained in sequential dependency of image feature in horizontal direction, and better help the model look up the text semantic features in the subsequent cross attention [20, 33]. In RPE, we maintain the learnable parameter with the same shape as image feature and encode the sequential dependency in horizontal direction to help the model better learn the neighboring context. See **supplementary file** for more details.

The position-encoded image feature, denoted by f'_I , and the encoder output f_E are then delivered to the decoder module for correlation computation. We process the two inputs with a Multi-head Cross Attention (MCA) layer, which performs cross attention operation between f_E and f'_I . Firstly, the features of f_E and f'_I are divided into n subgroups in the channel dimension. Then a cross attention operation CA_i is performed on the i -th group of f_E and f'_I :

$$CA_i(f_{Ei}, f'_{Ii}) = SM\left(\frac{(f'_{Ii}W_i^\alpha)(f_{Ei}W_i^\beta)^T}{\sqrt{d_k}}\right)(f_{Ei}W_i^\gamma) \quad (1)$$

where $f_{Ei} \in \mathbb{R}^{l \times \frac{c}{n}}$ and $f'_{Ii} \in \mathbb{R}^{hw \times \frac{c}{n}}$ denote the i -th group of f_E and f'_I , respectively. $W_i^\alpha \in \mathbb{R}^{\frac{c}{n} \times d_k}$, $W_i^\beta \in \mathbb{R}^{\frac{c}{n} \times d_k}$ and $W_i^\gamma \in \mathbb{R}^{\frac{c}{n} \times d_k}$ are the parameters of linear projections. SM refers to the Softmax operation. We process the results CA_i ($i \in \{0, 1, \dots, n-1\}$) with a channel-wise concatenation $\odot(\cdot)$ and a linear projection W^o , described as

$$MCA = \odot(CA_0, CA_1, \dots, CA_{n-1})W^o \quad (2)$$

The output of MCA is passed to a FFN for feature refinement, and then reshaped to obtain the TP map f_{TM} .

By using the MCA operation, the text prior f_E can effectively interact with the image feature f'_I by correlating every element in semantic domain to the position in spatial domain. Thus, the semantically meaningful regions in the spatial domain are strengthened in the TP map f_{TM} , which can be used to modulate the image feature for semantic-specific text reconstruction.

3.3. Text Structure Consistency Loss

While the proposed TATT network can attain a good performance, the reconstructed text image still needs some refinement to improve the visual appearance. This is because it is a bit difficult for a CNN model to represent the deformed text features as it does for regular text features, and the reconstructed text image has weaker character structures with relatively low contrast. As a remedy, we simulate deformed text images and design a text structure consistency (TSC) loss to train the proposed TATT network.

We consider minimizing the distance of three images, *i.e.*, the deformed version of the SR text image $\mathcal{D}\mathcal{F}(Y)$, the SR version of the deformed LR text image $\mathcal{F}(\mathcal{D}Y)$ and the deformed ground truth $\mathcal{D}(X)$, where \mathcal{D} denotes the random deformation¹. By increasing the similarity among the three items, we can encourage the CNN model to reduce the performance drop when encountering spatial deformations. The proposed TSC loss firstly measures the structural similarity between the above triplet. For this purpose, we extend the Structure-Similarity Index Measure (SSIM) [38] to a triplex SSIM (TSSIM), described as

$$TSSIM(x, y, z) = \frac{(\mu_x\mu_y + \mu_y\mu_z + \mu_x\mu_z + C_1)(\sigma_{xy} + \sigma_{yz} + \sigma_{xz} + C_2)}{(\mu_x^2 + \mu_y^2 + \mu_z^2 + C_1)(\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + C_2)} \quad (3)$$

where μ_x, μ_y, μ_z and $\sigma_x, \sigma_y, \sigma_z$ represent the mean and

¹We consider rotation, shearing and resizing in this paper.

standard deviation of the triplet x , y and z , respectively. σ_{xy} , σ_{yz} and σ_{xz} denote the correlation coefficients between (x, y) , (y, z) and (x, z) , respectively. C_1 and C_2 are small constants to avoid instability for dividing values close to zero. The derivation is in the **supplementary file**.

Lastly, TSC loss L_{TSC} is designed to measure the mutual structure difference among $\mathbf{D}\mathcal{F}(Y)$, $\mathcal{F}(\mathbf{D}Y)$ and $\mathbf{D}X$:

$$L_{TSC}(X, Y; \mathbf{D}) = 1 - TSSIM(\mathbf{D}\mathcal{F}(Y), \mathcal{F}(\mathbf{D}Y), \mathbf{D}X) \quad (4)$$

3.4. Overall Loss Function

In the training, the overall loss function includes a super resolution loss L_{SR} , a text prior loss L_{TP} and the proposed TSC loss L_{TSC} . The SR loss L_{SR} measures the difference between our SR output $\mathcal{F}(Y)$ and the ground-truth HR image X . We adopt L_2 norm for this computation. The TP loss measures the L_1 norm and KL Divergence between the text prior extracted from the LR image and those from the ground truth. Together with TSC loss L_{TSC} , the overall loss function is described as follows:

$$L = L_{SR} + \alpha L_{TP} + \beta L_{TSC} \quad (5)$$

where the α and β are the balancing parameters.

4. Experiments

4.1. Implementation Details

TATT is trained and tested on a single RTX 3090 GPU. We adopt Adam [16] optimizer to train the model with batch size 64. The training lasts for 500 epochs with learning rate 10^{-3} . The input image of our model is of width 64 and height 16, while the output is the $2 \times$ SR result. We set the α and β in (5) to 1 and 0.1, respectively (see **supplementary file** for ablations). The deformation operation \mathbf{D} in L_{TSC} is implemented by applying random rotation in a range of $[-10, 10]$ degree, shearing and aspect ratio in a range of $[0.5, 2.0]$. The head numbers of MSA and MCA layers are both set to 4 (following the best settings in [3]). The number of image feature channels c , d_k in MSA, MCA and FFN calculation are all set to 64. The model size of TATT is 14.9M in total. When training, the TPG is initialized with pretrained weights derived from [1], while other parts are randomly initialized. When testing, TATT will occupy 6.5GB of GPU memory with batch size 50.

4.2. Datasets

TextZoom. TextZoom [35] has 21,740 LR-HR text image pairs collected by changing the focal length of the camera in real-world scenarios, in which 17,367 samples are used for training. The rest samples are divided into three subsets, based on the camera focal length, for testing,

namely easy (1,619 samples), medium (1,411 samples) and hard (1,343 samples). Text label is provided in TextZoom.

Scene Text Recognition Datasets. Besides experiments conducted in TextZoom, we also adopt ICDAR2015 [15], CUTE80 [30] and SVTP [28] to evaluate the robustness of our model in recovering spatially-deformed LR text images. ICDAR2015 has 2,077 scene text images for testing. Most text images suffer from both low quality and perspective-distortion, making the recognition extremely challenging. CUTE80 is also collected in the wild. The test set has 288 samples in total. Samples in SVTP are mostly curve-shaped text. The total size of the test set is 649. Besides evaluating our model on the original samples, we further degrade the image quality to test the model generalization against unpredicted bad conditions.

4.3. Ablation Studies

In this section, we investigate the impact of TP Interpreter, the TSC loss function and the effectiveness of positional encoding. All evaluations in this section are performed on the real-world STISR dataset TextZoom [35]. The text recognition is performed by CRNN [31].

Impact of TP Interpreter in SR recovery. Since our TP Interpreter aims at providing better alignment between TP and the image feature and use text semantics to guide SR recovery, we compare it with other guiding strategies, *e.g.*, first upsampling the TP to match the image feature with deconvolution layers [22] or pixel shuffle to align text prior to image feature, and then fusing them to perform guidance with element-wise addition or SFT layers [37]². The results are shown in Tab. 1. One can see that the proposed TP interpreter obtains that highest PSNR/SSIM, which also indicates the best SR performance.

Referring to the SR text image recognition, one can see that using Pixel-Shuffle and deconvolution strategies provides inferior guidance (46.2% and 49.8%). There is no stable improvement by combining them with the SFT layers (47.9% and 48.6%). This is because none of the competing strategies performs global correlation between the text semantics and the image feature, resulting in inferior semantic guidance for SR recovery. In contrast, our TP Interpreter can obtain a good semantics context and accurate alignment to the text region. It thus strengthens the guidance in image feature and improves the text recognition result to 52.6%. This validates that using TP Interpreter is an effective way to utilize TP semantics for SR recovery. Some visual comparisons are shown in Fig. 4. One can see that the setting with TP interpreter can lead to the highest quality SR text image with correct semantics.

To demonstrate how the TP Interpreter provides global context, we visualize the attention heatmap provided by our

²The SFT layer merges the semantics of image feature with channel-wise affine transformation.

Strategy	avg	PSNR	SSIM
w/o TP	41.4%	21.42	0.7690
PS + A	46.2%	20.58	0.7683
PS + S [37]	47.9%	20.72	0.7560
D [23] + A	50.6%	21.10	0.7819
D [23] + S [37]	49.6%	20.87	0.7783
TPI	52.6%	21.52	0.7930

Table 1. Modules adopted in aligning and guiding the TP sequence to the image feature. D and PS refer to aligning operations Deconvolution and Pixel-Shuffle, respectively. A and S refer to guidance fusion operations by element-wise Addition and SFT Layers [37], respectively. TPI is the TP Interpreter.

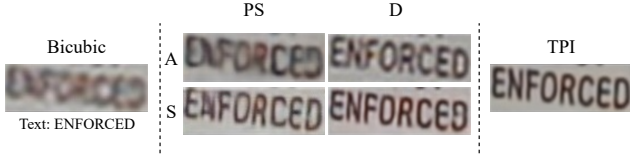


Figure 4. SR recovery by different guiding strategies.

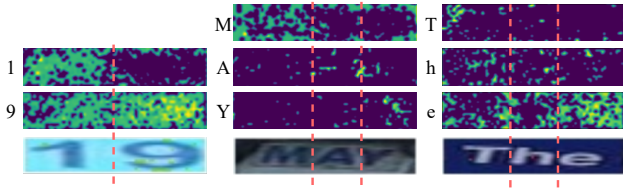


Figure 5. Attention heatmap of the foreground characters.

MCA (the outputs from the SM layer in (1)) in Fig. 5. One can see that the region of the corresponding foreground character has the highest weight (highlighted). It thus proves that the ability of TP Interpreter in finding semantics in image features. Some other highlighted regions in the neighborhood also demonstrate that the TP Interpreter can be aware of the neighboring context, which can provide better guidance for final SR recovery.

Impact of training with TSC loss. To validate the effectiveness of the TSC loss in refining text structure, we compare the results of 4 models trained with and without the TSC loss, including non-TP based TSRN [35], TBSRN [4], TP based TPGSR [22] and TATT. From the results in Tab. 2, one can see that all models lead to a performance gain (4.3% for TSRN, 1.3% for TBSRN, 0.8% for TPGSR, and 1.0% for TATT) in SR text recognition when adopting our TSC loss. Notably, though TBSRN [4] is claimed to be robust for multi-oriented text, it can still be improved with our TSC loss, indicating that training with the TSC loss can improve the robustness of reconstructing the character structure against various spatial deformations.

Effectiveness of the RPE. We evaluate the impact of recurrent positional encoding in learning text prior guidance. We deploy different combinations of fixed positional encoding (FPE), learnable positional encoding [3] and the pro-

Approach	L_{TSC}	easy	medium	hard	avg
TSRN [35]	×	52.5%	38.2%	31.4%	41.4%
	✓	58.0%	43.2%	33.4%	45.7%
TBSRN [4]	×	59.6%	47.1%	35.3%	48.1%
	✓	60.8%	49.6%	36.1%	49.4%
TPGSR [22]	×	61.0%	49.9%	36.7%	49.8%
	✓	62.0%	49.8%	37.4%	50.6%
ours	×	62.1%	52.1%	37.8%	51.6%
	✓	62.6%	53.4%	39.8%	52.6%

Table 2. TextZoom results of models with and without TSC loss.

Approach	Enc	Dec	avg
Ours	FPE	FPE	50.5%
	FPE	LPE	50.8%
	FPE	RPE	52.6%

Table 3. SR text image recognition results of different positional encoding ablations on TextZoom. The Enc and Dec refer to the encoder and decoder of the TP Interpreter.

posed recurrent positional encoding (RPE) in the encoder and decoder modules, and compare the corresponding text recognition results on the SR text images. From Tab. 3, we observe that using LPE or FPE in decoder shows limited performance because they are weak in learning the sequential information. By adopting RPE in the decoder, the SR recognition is increased by 1.8%, indicating that RPE is beneficial to text sequential semantics learning.

4.4. Comparison with State-of-the-Arts

Results on TextZoom. We conduct experiments on the real-world STISR dataset TextZoom [35] to compare the proposed TATT network with state-of-the-art SISR models, including SRCNN [8] and SRResNet [18] and HAN [26], and STISR models, including TSRN [35], TPGSR [22], PCAN [45] and TBSRN [4]. For TPGSR, we compare two models of it, *i.e.*, 1-stage and 3-stage (TPGSR-3). The evaluation metrics are SSIM/PSNR and text recognition accuracy. The comparison results are shown in Tab. 4 and Tab. 5.

One can see that our model trained with L_{TSC} achieves the best PSNR (21.52) and SSIM (0.7930) overall performance. This verifies the superiority of our method in improving the image quality. As for the SR text recognition, our method achieves new state-of-the-art accuracy under all settings by using the text recognition models of ASTER [32] and CRNN [31]. It even surpasses the 3-stage model TPGSR-3 by using only a single stage.

We also test the inference speed of the three most competitive STISR methods, *i.e.*, TBSRN (982 fps), TPGSR (1,085 fps) and our TATT model (960 fps). TATT has comparable speed with TPGSR and TBSRN, while surpasses them by 2.7% and 3.6% in SR image text recognition by using ASTER as the recognizer.

To further investigate the performance on spatially de-

Method	Loss	PSNR				SSIM			
		easy	medium	hard	avg	easy	medium	hard	avg
Bicubic	\times	22.35	18.98	19.39	20.35	0.7884	0.6254	0.6592	0.6961
SRCNN [8]	L_2	23.48	19.06	19.34	20.78	0.8379	0.6323	0.6791	0.7227
SRResNet [18]	$L_2+L_{IV}+L_P$	24.36	18.88	19.29	21.03	0.8681	0.6406	0.6911	0.7403
HAN [26]	L_2	23.30	19.02	20.16	20.95	0.8691	0.6537	0.7387	0.7596
TSRN [35]	L_2+L_{GP}	25.07	18.86	19.71	21.42	0.8897	0.6676	0.7302	0.7690
TBSRN [22]	$L_{POS}+L_{CON}$	23.46	19.17	19.68	20.91	0.8729	0.6455	0.7452	0.7603
PCAN [45]	L_2+L_{EG}	24.57	19.14	20.26	21.49	0.8830	0.6781	0.7475	0.7752
TPGSR [22]	L_2+L_{TP}	23.73	18.68	20.06	20.97	0.8805	0.6738	0.7440	0.7719
TPGSR-3 [22]	L_2+L_{TP}	24.35	18.73	19.93	21.18	0.8860	0.6784	0.7507	0.7774
TATT	$L_2+L_{TP}+L_{TSC}$	24.72	19.02	20.31	21.52	0.9006	0.6911	0.7703	0.7930

Table 4. PSNR/SSIM indices for competing SISR and STISR methods. ‘-3’ means multi-stage settings in [22].

Method	Loss	ASTER [32]				MORAN [21]				CRNN [31]			
		easy	medium	hard	avg	easy	medium	hard	avg	easy	medium	hard	avg
Bicubic	\times	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 [8]	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%
SRResNet [18]	$L_2+L_{IV}+L_P$	69.4%	47.3%	34.3%	51.3%	60.7%	42.9%	32.6%	46.3%	39.7%	27.6%	22.7%	30.6%
HAN [26]	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 [35]	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%
TBSRN [22]	$L_{POS}+L_{CON}$	75.7%	59.9%	41.6%	60.0%	74.1%	57.0%	40.8%	58.4%	59.6%	47.1%	35.3%	48.1%
PCAN [45]	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%
TPGSR [22]	L_2+L_{TP}	77.0%	60.9%	42.4%	60.9%	72.2%	57.8%	41.3%	57.8%	61.0%	49.9%	36.7%	49.8%
TPGSR-3 [22]	L_2+L_{TP}	78.9%	62.7%	44.5%	62.8%	74.9%	60.5%	44.1%	60.5%	63.1%	52.0%	38.6%	51.8%
TATT	$L_2+L_{TP}+L_{TSC}$	78.9%	63.4%	45.4%	63.6%	72.5%	60.2%	43.1%	59.5%	62.6%	53.4%	39.8%	52.6%
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%

Table 5. SR text recognition for competing SISR and STISR methods. ‘-3’ means multi-stage settings in [22].

Method	AS [32]	MO [21]	CR [31]	PSNR	SSIM
Bicubic [35]	36.1%	32.2%	19.5%	19.68	0.6658
TSRN [35]	46.6%	43.8%	35.2%	19.70	0.7157
TBSRN [4]	48.5%	45.1%	37.3%	19.10	0.7066
TPGSR [23]	46.6%	45.3%	40.2%	19.79	0.7293
Ours	51.7%	47.3%	43.8%	20.20	0.7535
HR	80.8%	75.7%	68.8%	-	-

Table 6. Evaluation of competitive STISR models on spatially-deformed samples picked in TextZoom in terms of recognition, PSNR and SSIM. ‘AS’, ‘MO’ and ‘CR’ refer to ASTER [32], MORAN [21] and CRNN [31], respectively.

formed text images, we manually pick 804 rotated and curve-shaped samples from TextZoom test set to evaluate the compared models. Results in Tab. 6 indicate that our TATT model obtains the best performance, and the average gap over models like TPGSR and TBSRN becomes larger when encountering spatially deformed text.

We also visualize the recovery results of both regular samples and spatially-deformed samples of TextZoom in Fig. 6. Without TP guidance, TSRN and TBSRN perform far from readable and they are visually unacceptable. With the TP guidance, TPGSR is still unstable in recovering spatially-deformed images. In contrast, our TATT network performs much better in recovering text semantics in samples of all cases compared to all the competitors. With TSC loss, our model further upgrades the visual quality of the

	super-resolver	AS [32]	MO [21]	CR [31]
O	Bicubic	38.1%	29.1%	18.1%
	TSRN [35]	41.5%	33.8%	26.6%
	TBSRN [4]	46.8%	45.3%	38.3%
	TPGSR [22]	53.1%	52.3%	42.5%
	Ours	53.4%	59.1%	47.2%
CO	Bicubic	33.2%	28.1%	23.6%
	TSRN [35]	46.4%	42.1%	29.1%
	TBSRN [4]	45.5%	44.7%	31.9%
	TPGSR [22]	48.3%	52.8%	38.3%
	Ours	54.7%	54.0%	45.1%
GN	Bicubic	29.4%	25.8%	7.5%
	TSRN [35]	31.3%	27.5%	11.5%
	TBSRN [4]	40.2%	33.4%	15.8%
	TPGSR [22]	35.7%	31.7%	18.1%
	Ours	43.0%	33.4%	21.1%
GB	Bicubic	27.0%	22.3%	5.5%
	TSRN [35]	39.2%	35.8%	20.4%
	TBSRN [4]	42.6%	42.8%	20.8%
	TPGSR [22]	45.9%	43.8%	29.6%
	Ours	47.4%	43.8%	35.7%

Table 7. Impact of using different STISR models as super-resolver against degradation. ‘O’, ‘CO’, ‘GB’ and ‘GN’ refer to original images and image degradation in terms of contrast, Gaussian blurring and Gaussian noise. ‘AS’, ‘MO’ and ‘CR’ refer to ASTER [32], MORAN [21] and CRNN [31], respectively.

samples with better-refined character structure.

Generalization to recognition dataset. We evaluate the generalization performance of our TATT network to other

	Regular		Spatially-Deformed				
Bicubic							
	fo s tiflash	co f erria	1 n 969	s d irs	po r vu l ail	wo s l	pu l andess
TSRN							
	fo b tiflash	cb i ifornia	18969	shap e r	formu l abl	wo s e	obligator (i) os
TBSRN							
	fo b yiflash	califo r aia	1996	shap e s	formu l ad l	ro s e	obligator i cs
TPGSR							
	fo b tiflash	california	1996 9	shape o	formu l as	po s e	obligator (i) os
TATT							
	fortiflash	california	1896 9	shap e s	formu l as	po s e	obligator i cs
TATT w TSC							
	fortiflash	california	1996	shap e s	formu l as	ro s e	obligator i os
HR							
	fortiflash	california	1996	shap e s	formu l as	ro s e	obligator i os

Figure 6. Visualization of regular and spatially-deformed samples from TextZoom recovered by state-of-the-art STISR models and the SR text recognition results. Characters in red are missing or wrong. ‘w TSC’ means that the model is trained with our TSC loss.



Figure 7. Visualization of the STISR and text recognition results on extremely compressed and blurred text samples.

real-world text image datasets, including ICDAR15 [15], CUTE80 [30] and SVTP [28]. These datasets are built for text recognition purpose and contain spatially deformed text image in natural scenes. Since some of the images in these datasets have good quality, we only pick the low-resolution images (*i.e.*, lower than 16×64) to form our test set with 533 samples (391 from ICDAR15, 3 from CUTE80 and 139 from SVTP). Since the degradation is relatively small, we manually add some degradation on them, including contrast variation, Gaussian noise and Gaussian blurring (see details in **supplementary file**). We compare with TSRN [35], TBSRN [4] and TPGSR [22] in this test and evaluate the recognition accuracy on the SR results. All models are trained on TextZoom and tested on the picked low-quality images.

The results are illustrated in Tab. 7, we can see that the proposed TATT network achieves the highest recognition accuracy across all types of degradations. This indicates that our TATT network, though trained on TextZoom, can be well generalized to images in other datasets. The recon-

structed high-quality text images by TATT can benefit the downstream tasks such as text recognition.

5. Conclusion and Discussions

In this paper, we proposed a Text ATTention network for single text image super-resolution. We leveraged a text prior, which is the semantic information extracted from the text image, to guide the text image reconstruction process. To tackle with the spatially-deformed text recovery, we developed a transformer-based module, called TP Interpreter, to globally correlate the text prior in the semantic domain to the character region in image feature domain. Moreover, we proposed a text structure consistency loss to refine the text structure by imposing structural consistency between the recovered regular and deformed texts. Our model achieved state-of-the-art performance in not only the text super resolution task but the downstream text recognition task.

Though recording state-of-the-art results, the proposed TATT network has limitation on recovering extremely blurry texts, as shown in Fig. 7. In such cases, the strokes of the characters in the text are mixed together, which are difficult to separate. In addition, the computational complexity of our TATT network grows exponentially with the length of the text in the image due to the global attention adopted in our model. It is expected to reduce the computational complexity and improve run-time efficiency of TATT, which will be our future work.

6. Acknowledgements

This work is supported by the Hong Kong RGC RIF grant (R5001-18). We thank Dr. Lida Li for the useful discussion on this project.

References

- [1] crnn.pytorch. <https://github.com/meijieru/crnn.pytorch>. 5
- [2] 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 *Int. Conf. Comput. Vis.*, pages 3086–3095, 2019. 2
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 5, 6
- [4] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12026–12035, 2021. 1, 2, 6, 7, 8
- [5] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2492–2501, 2018. 2
- [6] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Int. Conf. Comput. Vis.*, pages 5076–5084, 2017. 2
- [7] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. AON: Towards arbitrarily-oriented text recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5571–5579, 2018. 2
- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2015. 2, 6, 7
- [9] 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. 2
- [10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Int. Conf. Machine. Learn.*, pages 369–376, 2006. 2
- [11] Bahadır K. Gunturk, Yucel Altunbasak, and Russell M. Mersereau. Super-resolution reconstruction of compressed video using transform-domain statistics. *IEEE Trans. Image Process.*, 13(1):33–43, 2004. 2
- [12] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang. Reading scene text in deep convolutional sequences. In *AAAI*, 2016. 2
- [13] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.*, 116(1):1–20, 2016. 2
- [14] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In *Eur. Conf. Comput. Vis.*, pages 512–528. Springer, 2014. 2
- [15] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos 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 *Int. Conf. Doc. Anal. Recog.*, pages 1156–1160. IEEE, 2015. 5, 8
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5
- [17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 624–632, 2017. 2
- [18] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4681–4690, 2017. 2, 6, 7
- [19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 136–144, 2017. 2
- [20] Antoine Liutkus, Ondřej Cifka, Shih-Lun Wu, Umut Simsekli, Yi-Hsuan Yang, and Gael Richard. Relative positional encoding for transformers with linear complexity. In *Int. Conf. Machine. Learn.*, pages 7067–7079. PMLR, 2021. 4
- [21] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognit.*, 90:109–118, 2019. 1, 7
- [22] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *arXiv preprint arXiv:2106.15368*, 2021. 1, 2, 3, 5, 6, 7, 8
- [23] Jianqi Ma, Weiyan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimedia*, 20(11):3111–3122, 2018. 1, 6, 7
- [24] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 2272–2279. IEEE Computer Society, 2009. 2
- [25] 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 *Eur. Conf. Comput. Vis.*, 2020. 2
- [26] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Eur. Conf. Comput. Vis.*, pages 191–207. Springer, 2020. 6, 7
- [27] Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. ICDAR2015 competition on text image super-resolution. In *Int. Conf. Doc. Anal. Recog.*, pages 1201–1205. IEEE, 2015. 2
- [28] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Int. Conf. Comput. Vis.*, pages 569–576, 2013. 5, 8
- [29] Yuhui Quan, Jieting Yang, Yixin Chen, Yong Xu, and Hui Ji. Collaborative deep learning for super-resolving blurry text images. *IEEE Trans. Comput. Imaging*, 6:778–790, 2020. 2

- [30] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, 41(18):8027–8048, 2014. 5, 8
- [31] 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 Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2016. 1, 2, 5, 6, 7
- [32] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2035–2048, 2018. 1, 6, 7
- [33] Vighnesh Shiv and Chris Quirk. Novel positional encodings to enable tree-based transformers. *Adv. Neural Inform. Process. Syst.*, 32:12081–12091, 2019. 4
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 3, 4
- [35] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3, 5, 6, 7, 8
- [36] Wenjia Wang, Enze Xie, Peize Sun, Wenhai Wang, Lixun Tian, Chunhua Shen, and Ping Luo. TextSR: Content-aware text super-resolution guided by recognition. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2
- [37] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 606–615, 2018. 2, 5, 6
- [38] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 4
- [39] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *Int. Conf. Comput. Vis.*, pages 251–260, 2017. 2
- [40] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Process.*, 19(11):2861–2873, 2010. 2
- [41] Haochen Zhang, Dong Liu, and Zhiwei Xiong. CNN-based text image super-resolution tailored for OCR. In *Vis. Commun. Image Process.*, pages 1–4. IEEE, 2017. 1
- [42] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3762–3770, 2019. 2
- [43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pages 286–301, 2018. 2
- [44] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2472–2481, 2018. 2
- [45] Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. Scene text image super-resolution via parallelly contextual attention network. In *ACM Int. Conf. Multimedia*, pages 2908–2917, 2021. 2, 6, 7
- [46] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: An efficient and accurate scene text detector. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5551–5560, 2017. 1