StyleSRN: Scene Text Image Super-Resolution with Text Style Embedding

Shengrong Yuan[1], Runmin Wang[1,*], Ke Hao[1], Xuqi Ma[1], Changxin Gao[2], Li Liu[3], Nong Sang[2]
[1]School of Information Science and Engineering, Hunan Normal University
[2]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[3]School of Electronic Science, National University of Defense Technology
{shengrongyuan, runminwang, ke, xqma}@hunnu.edu.cn, {cgao, nsang}@hust.edu.cn, liliu@oulu.fi

Abstract

Scene text image super-resolution (STISR) focuses on enhancing the clarity and readability of low-resolution text images. Existing methods often rely on text probability distribution priors derived from text recognizers to guide the super-resolution process. While effective in capturing general structural information of text, these priors lack the ability to preserve specific text style details, such as font, stereoscopic effect and spatial transformation, leading to a loss of visual quality and stylistic consistency. To address these limitations, we propose a Style embedding-based scene text image Super-Resolution Network (StyleSRN), which introduces a text style embedding mechanism to preserve and enhance text style features during the super-resolution process. The proposed architecture includes Style Enhancement Blocks for capturing multi-scale cross-channel dependencies, and Style Content Fusion Blocks that effectively integrates text content with style information, ensuring that the structure and style of the restored text are not distorted. Furthermore, we introduce a Text Style Loss based on the Gram matrix to supervise the reconstruction process at the style level, thereby maintaining the stylistic consistency of the restored text images. Extensive experiments on the TextZoom dataset and five scene text recognition benchmarks demonstrate the superiority of our method. Code is available at: https://github.com/Yuanssr/StyleSRN.

1. Introduction

Scene text images contain both visual and linguistic modalities, which can be extracted and analyzed to enhance various applications, including autonomous driving [32], visual question answering [1], and card recognition [13], etc. However, due to the limitations of imaging conditions of sensing devices, such as shooting jitter and low-focus cameras, scene text images often suffer from varying degradation, including blur, distortion, etc. Unlike general object images, scene text images possess more complex character structures and fine-grained stroke-level details, making even slight degradation challenging to recognize. Such low-resolution scene text images significantly hinder the performance of downstream tasks such as scene text detection, optical character recognition, and scene text recognition. Thus, restoring both the visual style quality and character structures of these degraded images is crucial.

Figure 1. Comparison of super-resolution results between previous methods and ours. Methods without text priors such as TSRN [30] retains less style information, while text priors-guided methods such as TATT [17] and LEMMA [7] almost completely lose the style information.

In recent years, many super-resolution methods specifically designed for scene text images have used the unique properties of text to improve restoration. Early approaches primarily focused on extracting sequential information from text images. Wang *et al*. [30] proposed a sequential residual block to capture the inherent sequential nature of text images. Chen *et al*. [2] introduced a text-focus loss function that emphasizes character positions and content, thereby improving recognition accuracy. More recently, methods that incorporate deep character classification information obtained from text recognizers have become the norm. Ma *et al*. [18] highlighted the benefits of integrating text

*Corresponding author

priors into the super-resolution process. Moreover, Ma *et al*. [17] developed an attention-based text priors interpreter to interactively fuse text probability distributions with image features. Guo *et al*. [7] demonstrated the effectiveness of using a bidirectional alignment strategy for text and images to obtain high-level guidance information. Most of these methods rely on text recognizers to generate text probability distribution priors, which are then encoded into the super-resolution network to guide the recovery of character structures.

However, the text probability distribution priors derived from text recognizers mainly capture the general character structures but are less sensitive to the variations in text styles across different images, which hampers the super-resolution process. This loss of style details not only affects subsequent scene text recognition tasks but also disrupts the visual consistency of the images. The problem is particularly pronounced in scene text images with distinctive text styles, such as artistic characters or logos, where the loss of stylistic features can lead to more severe distortion in the restored images. As shown in Figure 1, text priors-guided methods such as TATT [17] and LEMMA [7] almost completely lose the style information of the degraded text image, while methods without text priors such as TSRN [30] retain less style information. To address these challenges, we draw inspiration from Style Transfer [6, 8, 12] to explore the potential of preserving image style information to enhance the restoration of low-resolution text images. Unlike content information, which is sensitive to position, style information in an image can be viewed as the correlation between different feature channels and is largely position-insensitive. By effectively encoding and embedding this style information during the super-resolution process, our proposed method aims to preserve both the structural integrity and the stylistic consistency of the restored text images.

In this paper, we propose a Style embedding-based scene text image Super-Resolution Network (StyleSRN) for STISR, which not only restores the character structure but also preserves the stylistic consistency of the text. Unlike conventional methods that rely solely on text probability priors to guide the super-resolution process, our approach introduces a Text Style Embedding Branch that extracts and enhances style features. Additionally, we design a Style Enhancement Block (SEB), which captures multi-scale cross-channel dependencies, thereby enhancing the representation of style information. To further ensure the integration of style and content, we introduce a Style Content Fuse Block (SCFB) that effectively combines style embeddings with text priors. Finally, we propose a Text Style Loss based on the Gram matrix to supervise the reconstructed images at the style level. As illustrated in Figure 1, our method achieves superior visual quality and stylistic consistency in the restored images. Overall, our contributions can be summarized as follows:

- We, for the first time to our best knowledge, successfully embed text style information to guide super-resolution process for STISR task, and validate its effectiveness to ensure both structural integrity and stylistic consistency of the super-resolution process.
- We propose a Text Style Loss based on the Gram Matrix which leverages a text recognizer to extract features to offer robust supervision at style level for scene text images, which effectively improve both image quality and text recognition accuracy.
- Extensive experiment results demonstrate that the proposed method not only achieves state-of-the-art performance on the TextZoom dataset but also validates exceptional generalization across five scene text recognition datasets.

2. Related Work

2.1. Scene Text Image Super-resolution

Scene text image super-resolution aims to enhance the visual quality and readability of scene text images. Early studies in this field usually adopt general image super-resolution methods [4, 5, 14, 21, 28] to reconstruct scene text images. However, due to the complex character structure of text images, these methods often fail to produce sufficiently recognizable text images. Therefore, some unique attributes of scene text have been exploited to improve super-resolution performance[2, 3, 7, 17, 18, 30]. For example, some methods focus on the sequential information of scene text. Wang *et al*. [30] introduced sequential residual block to capture the sequential information in text images. Chen *et al*. [2] proposed a Transformer-Based Super-Resolution Network containing a Self-Attention Module to extract sequential information. Recently, the method of using text recognizers to obtain text probability distribution priors has become mainstream. For example, Ma *et al*. [18] incorporates categorical text priors into the STISR model training process, while Ma *et al*. [17] proposed a text prior interpreter to better transform text priors. Guo *et al*. [7] proposed a bimodal alignment mechanism to better utilize text priors. Additionally, some works use text priors to guide diffusion models to restore low-resolution images [20, 35]. Most of these methods use a pre-trained text recognizer to obtain deep character classification information to guide the super-resolution network. While this approach enables the model to effectively learn the general character structure, it results in the loss of specific style information in each image.

2.2. Style Transfer

Style transfer is a technique that involves altering the stylistic attributes of an image while maintaining its core con-

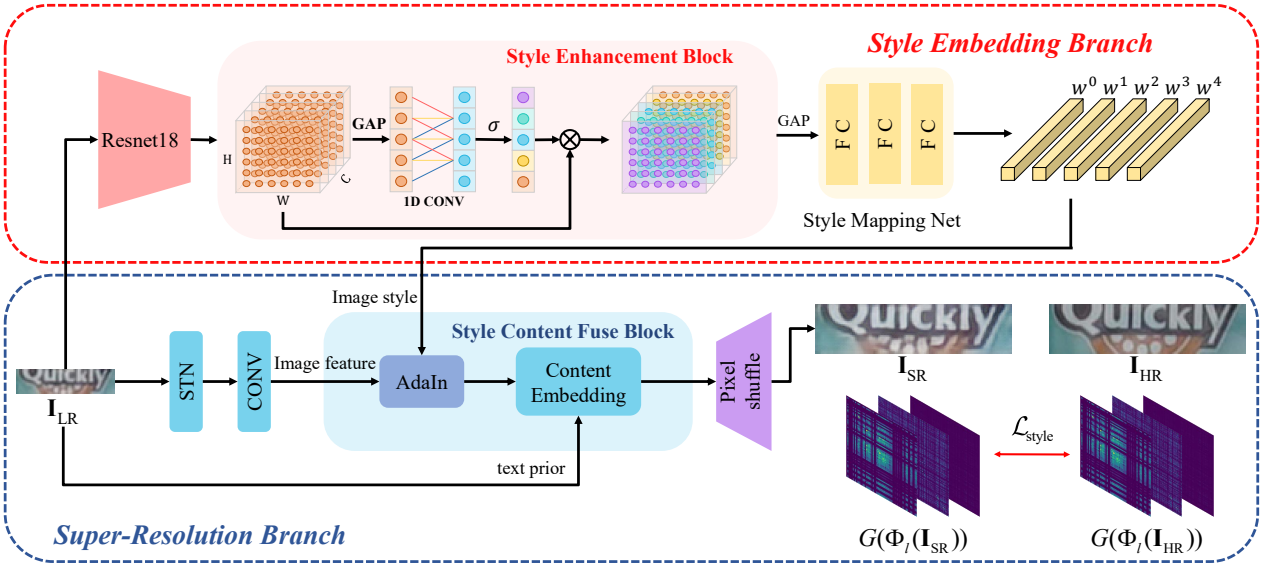

Figure 2. Overview architecture of the proposed StyleSRN. GAP denotes the Global Average Pooling. $G(\cdot)$ denotes the Gram matrix, $\Phi_l(\cdot)$ represents the feature map obtained from the l-th layer of the text recognizer. $\mathbf{I}_{\text{LR}}$, $\mathbf{I}_{\text{SR}}$ and $\mathbf{I}_{\text{HR}}$ represent low-resolution images, super-resolution images and high-resolution images, respectively.

tent. This concept was pioneered by Gatys *et al*. [6], who utilized convolutional neural networks (CNNs) to transfer artistic styles between images by matching the Gram matrices of feature maps. Subsequent work has expanded on this foundation, exploring various methods to improve both the quality and efficiency of style transfer. One significant advancement is the introduction of adaptive instance normalization (AdaIN) [8], which allows for real-time arbitrary style transfer by aligning the mean and variance of content features to those of style features. Some works employ style transfer technology specifically for scene text style transfer [23, 27]. In our work, we draw inspiration from these style transfer methods, particularly in how they encode and embed style features. By adopting these principles, we aim to restore text style information, which is crucial for maintaining the visual consistency of the super-resolved text images.

3. Methodology

3.1. Overall Architecture

The overall architecture of the proposed StyleSRN is illustrated in Figure 2. Our method consists of a text style embedding branch for extracting and enhancing style features and a super-resolution branch for restoring low-resolution images $\mathbf{I}_{\text{LR}}$. For the text style embedding branch, the low-resolution image will first pass through ResNet18 to obtain the feature map $\mathbf{X}_f \in \mathbb{R}^{C\times W\times H}$, where C represents the number of channels, and H and W represent the height and width of the feature map, respectively. Then several stacked Style Enhancement Blocks will extract the correlation information between the feature map channels. After the style enhancement, we perform global average pooling on it to obtain the latent vector $\mathbf{z}$. Given a latent vector $\mathbf{z}$ in the input latent space $\mathcal{Z}$, a non-linear mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$ first produces $\mathbf{w} \in \mathcal{W}$. Learned affine transformations then specialize $\mathbf{w}$ to styles $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$, where $\mathbf{y}_s$ and $\mathbf{y}_b$ denote the learned normalization scaling and bias coefficients, respectively. For the Super-Resolution Branch, the low-resolution image will first be processed by the Spatial Transformer Network (STN) [9] to align the spatially deformed texts. The convolution layer will extract features from the image processed by STN to obtain shallow image features. Then the shallow image features, text priors, and style information will be sent to several stacked style-content fusion blocks for style and content fusion. Each block contains a Style Enhancement Block, an adaptive instance normalization (AdaIN) [8], and a content embedding block. Finally, PixelShuffle is applied as an upsampling operator to generate $\mathbf{I}_{\text{SR}}$.

3.2. Style Enhancement Block

The style of an image has been traditionally viewed as the correlation between different features, as demonstrated in the advancements of style transfer [6, 8, 12]. While previous approaches, such as [12], directly utilized ResNet features to obtain style vectors, our approach introduces a novel enhancement by applying channel-wise attention to these feature maps, thereby better capturing and preserving the stylistic characteristics in scene text images.

Inspired by the Efficient Channel Attention (ECA)

mechanism [29], we designed a Style Enhancement Block (SEB) to improve the model's capacity for capturing the correlation between different channel features. Let the input feature map be denoted as $\mathbf{X}_f \in \mathbb{R}^{C\times H\times W}$. We begin by applying Global Average Pooling (GAP) to the input feature map to obtain channel-wise statistics $\mathbf{m} \in \mathbb{R}^C$. To further improve the model's capacity for capturing local cross-channel interactions at multiple scales, we employ multiple 1D convolution kernels of varying sizes, in contrast to the fixed-size kernels used in [29]. Specifically, let n denote the number of different kernel sizes, each represented by k_l, where $l \in 1, 2, \ldots, n$. The convolutional operation at each scale can be expressed as:

$$\mathbf{m'}_c^{(l)} = \sum_{j=-\frac{k_l}{2}}^{\frac{k_l}{2}} b_j^l \mathbf{m}_{c+j}, \tag{1}$$

where $\mathbf{b}^l \in \mathbb{R}^{k_l}$ represents the weight vector for the convolution kernel of size k_l. This multi-scale approach allows the model to capture a more nuanced set of cross-channel dependencies, enhancing the representation of style information. The outputs from the multiple convolutional scales are aggregated by computing the final channel attention weight $\mathbf{m'}_c$ as a weighted sum:

$$\mathbf{m'}_c = \sum_{l=1}^{n} \alpha_l \mathbf{m'}_c^{(l)}, \tag{2}$$

where α_l denotes the weighting coefficient for the output of the l-th convolutional scale. These weights are normalized using a Sigmoid function:

$$\mathbf{X'}_f^{c,i,j} = \sigma(\mathbf{m'}_c) \cdot \mathbf{X}_f^{c,i,j}, \tag{3}$$

where $\sigma(\cdot)$ represents the Sigmoid activation function, and $\mathbf{X'}$ is the final output feature map with the applied channel-wise attention.

The enhanced feature map is passed through several SEBs, which refines the correlation between different feature channels, ensuring that the model pays more attention to the features of important channels. This process effectively enhances the model's capability to capture and preserve the intricate stylistic details in scene text images. Subsequently, global average pooling (GAP) is applied to the final enhanced feature map to obtain a latent vector $\mathbf{z}$ in the latent space $\mathcal{Z}$. The latent vector is then processed by a non-linear style mapping network $f : \mathcal{Z} \to \mathcal{W}$, which produces a vector $\mathbf{w} \in \mathcal{W}$ that decouples the latent space. The style mapping network consists of an 8-layer MLP with LeakyReLU activation functions. Finally, learned affine transformations are applied to transform $\mathbf{w}$ into style parameters $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$.

3.3. Style Content Fuse Block

In the field of scene text image super-resolution, existing approaches primarily focus on embedding content information, including deep character category information derived from text probability priors. While this approach effectively guides the super-resolution network in restoring character structure, it often results in the loss of crucial stylistic features. To address this limitation, we propose a novel Style Content Fuse Block (SCFB) that allows the super-resolution network to restore character structures while preserving the stylistic consistency of the characters.

Our innovation lies in embedding style information directly into the super-resolution network alongside content information. This dual embedding approach ensures that both character structure and style are preserved during the super-resolution process. Specifically, for the shallow image features $\mathbf{X}_s \in \mathbb{R}^{C'\times H'\times W'}$ obtained through a Spatial Transformer Network (STN) and a convolutional layer, we use Adaptive Instance Normalization (AdaIN) [8] to modulate their style. AdaIN has been shown to be effective in transferring style information across different domains. Given a feature map $\mathbf{X}_s$ and a style vector $\mathbf{y}$, the AdaIN operation can be defined as follows:

$$\text{AdaIN}(\mathbf{X}_s, \mathbf{y}) = \mathbf{y}_s \left(\frac{\mathbf{X}_s - \mu(\mathbf{X}_s)}{\sigma(\mathbf{X}_s)} \right) + \mathbf{y}_b, \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and variance of $\mathbf{X}_s$. Each AdaIN layer within the SCFB normalizes feature maps independently, with the dimensionality of $\mathbf{y}$ being twice the number of feature map channels C' in that layer. This embedding ensures that style information is seamlessly integrated into the spatial features of the image, thus preserving the character's stylistic attributes.

In addition to embedding style information, content information, such as text probability distribution priors, should also be embedded into the super-resolution network to guide the learning of character structures. To bridge the domain gap between text priors and image features, we adopt the TP Interpreter architecture from [17] to extract the content embedding features. These interpreted features are then concatenated with the style-embedded features and subsequently passed through a Sequential Residual Block [30] for further refinement.

3.4. Text Style Loss

Although the proposed StyleSRN network effectively embeds the text's style features, these features are derived from the low-resolution image itself. When the style information in the low-resolution image is degraded, even if the network extracts these style features, it cannot fully reconstruct the style of the corresponding high-resolution image. Therefore, the reconstructed text image requires supervision at

the style level to effectively learn the degradation of style information in the low-resolution image.

To supervise the style in the reconstructed image, we introduce a Text Style Loss based on the Gram matrix. Unlike traditional methods [6] that use VGG networks to capture style information, our approach leverages a text recognizer to extract feature maps for Gram matrix computation. This enables better capture of text-specific style features, which are crucial for scene text images.

The Gram matrix $\mathbf{G}_l$ for a given feature map $\Phi_l(\mathbf{I})$ at the l-th layer is computed as follows. First, we reshape the feature map $\Phi_l(\mathbf{I})$ of dimensions $C_l \times H_l \times W_l$ into a matrix $\mathbf{F}_l$ of size $C_l \times (H_l \times W_l)$, where each row corresponds to a channel and each column corresponds to a pixel. The Gram matrix $\mathbf{G}_l$ is the inner product between the different channels in layer l:

$$\mathbf{G}^l_{ij} = \sum_{k=1}^{H_l \times W_l} \mathbf{F}^l_{ik} \mathbf{F}^l_{jk} \tag{5}$$

The Gram matrix captures the correlations between different channels in the feature map, thereby encoding the style information of the image.

The Text Style Loss $\mathcal{L}_{\text{style}}$ is defined as the Mean Squared Error (MSE) between the Gram matrix of the high-resolution (HR) ground truth and the super-resolved (SR) output:

$$\mathcal{L}_{style} = \sum_{l=1}^{L} \frac{1}{4C_l^2 H_l^2 W_l^2} \left\| \mathbf{G}(\Phi_l(\mathbf{I}_{\text{HR}})) - \mathbf{G}(\Phi_l(\mathbf{I}_{\text{SR}})) \right\|_2^2 \tag{6}$$

where $\mathbf{G}(\cdot)$ denotes the Gram matrix, $\Phi_l(\cdot)$ represents the feature map obtained from the l-th layer of the text recognizer, and C_l, H_l, W_l are the dimensions of the feature map at layer l.

3.5. Overall Loss Function

In training, the overall loss function includes a Mean Square Error (MSE) loss $\mathcal{L}_2$, a Text-Focus Loss $\mathcal{L}_{TP}$ [30], and the proposed Text Style Loss $\mathcal{L}_{style}$. The $\mathcal{L}_2$ measures the difference between the super-resolution (SR) output and the ground-truth high-resolution (HR) image. The overall loss function is described as follows:

$$\mathcal{L} = \mathcal{L}_2 + \alpha \mathcal{L}_{TP} + \beta \mathcal{L}_{style} \tag{7}$$

where the α and β are the balancing parameters.

4. Experiments

4.1. Datasets

TextZoom TextZoom [30] is a scene text super-resolution dataset designed for real-world scenarios, containing 21,740 pairs of low-resolution and high-resolution images captured by cameras with various focal lengths in natural environments, with 17,367 samples used for training. The test set is divided into easy, medium, and hard subsets, comprising 1,619, 1,411, and 1,343 LR-HR pairs, respectively, based on the camera focal lengths.

Scene Text Recognition Datasets To assess the robustness and generalization capability of StyleSRN, we conducted comprehensive evaluations on five benchmark scene text recognition datasets: IC13 [10], IC15 [11], CUTE80 [24], IIIT5K [19], and SVTP [22]. These datasets were manually degraded to create low-resolution counterparts. Detailed descriptions of the datasets and degradation procedures are provided in the supplementary materials.

4.2. Metric

Following previous works, we adopt two types of metrics: super-resolution image text recognition accuracy and image quality. Specifically, to evaluate the generalization capability of StyleSRN, we employed three scene text recognizers: CRNN [25], MORAN [16], and ASTER [26]. For the image quality metric, we adopt the widely used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM)[31].

4.3. Implementation Details

We implement our model using PyTorch. All experiments are conducted on a single RTX 3090 GPU. We train our model with a batch size of 64 for 500 epochs using AdamW [15] for optimization. The learning rate is set to 1e-3 and decayed by a factor of 0.5 after 400 epochs. The number of SEBs and SCFBs are set to 3 and 5, respectively. In the SEBs, the sizes of the convolution kernels are 3×3, 5×5, and 7×7, respectively. We use CRNN to generate the features for computing the text style loss and empirically set the number of feature maps L used to calculate the style loss to 5. We set the parameters α and β to 0.5 and 1.0, respectively.

4.4. Ablation Study

In this section, we conduct ablation studies to investigate the effectiveness of SEBs and SCFBs, the effectiveness of training with Text Style Loss, the impact of numbers of SEBs and SCFBs. For STISR methods that do not use text priors, such as TSRN [30] and TBSRN [2], using SCFB means that image features are not used for text prior fusion, but are directly concatenated with style-embedded features.

Effectiveness of SEBs and SCFBs. We compared the performance of different STISR methods using three strategies: No Style Embedding (NSE), only SCFB , and SCFB

Method	NSE	SCFB	SEB	Image Quality		Accuracy(%)
				PSNR	SSIM	
TSRN [30]	✔	-	-	21.42	0.7690	41.4
	-	✔	-	21.54	**0.7699**	43.5
	-	✔	✔	**21.56**	0.7695	**44.4**
TBSRN [2]	✔	-	-	20.91	0.7603	48.1
	-	✔	-	21.04	**0.7623**	50.3
	-	✔	✔	**21.22**	0.7614	**51.1**
TPGSR [18]	✔	-	-	20.97	0.7719	51.8
	-	✔	-	21.04	**0.7723**	52.4
	-	✔	✔	**21.17**	**0.7723**	**52.7**
TATT [17]	✔	-	-	21.52	**0.7930**	52.6
	-	✔	-	21.66	0.7790	54.1
	-	✔	✔	**21.87**	0.7785	**54.9**
LEMMA [7]	✔	-	-	20.88	0.7760	56.3
	-	✔	-	20.94	**0.7771**	56.7
	-	✔	✔	**21.05**	0.7765	**57.0**

Table 1. Effectiveness of the Style Content Fuse Block (SCFB) and the Style Enhancement Block (SEB). NSE refer to No Style Embedding. Accuracy represents the average accuracy with CRNN [25].

Method	$\mathcal{L}_{\text{style}}$	PSNR	SSIM	Accuracy(%)
TSRN [30]	-	21.42	0.7690	41.4
	✔	**21.56**	**0.7711**	**43.7**
TPGSR [18]	-	20.97	0.7719	51.8
	✔	**21.12**	**0.7724**	**52.4**
TATT [17]	-	21.52	0.7930	52.6
	✔	**21.81**	**0.7938**	**53.8**
StyleSRN	-	21.54	0.7758	54.9
	✔	**21.82**	**0.7778**	**57.4**

Table 2. Ablation of the Text Style Loss. Accuracy represents the average accuracy with CRNN [25].

combined with the SEB in Table 2. NSE means removing both SEB and SCFB. SCFB can be used independently of SEB, but when using SEB, SCFB must be used to embed style information. Removing SEB means that the features extracted by Resnet18 are directly fed into the style mapping network through global pooling. Removing SCFB means that the features directly passed through the text interpreter do not need to be cascaded with the style features. As presented in Table 1, using only SCFB generally improves image quality and recognition accuracy over NSE. In addition, SCFB has a greater improvement over methods without text priors such as TSRN [30] and TBSRN [2]. For instance, with TSRN, the introduction of SCFB increases accuracy by 2.1%, along with notable gains in PSNR and SSIM. We speculate that this may be because these methods do not have any prior guidance, while SCFB introduces the guidance of style information. Using SCFB and SEB at the same time can effectively improve PSNR and text recognition accuracy, but slightly reduce SSIM. This shows that SEB can effectively enhance style information, but slightly reduce structural consistency.

Discriminator	PSNR	SSIM	Accuracy (%)
VGG	21.31	0.7761	56.3
CRNN	**21.82**	**0.7778**	**57.4**

Table 3. Ablation of the discriminator used in text style loss.

N	Quality Metric		Accuracy(%)		
	PSNR	SSIM	CRNN [25]	MORAN [16]	ASTER [26]
0	20.76	0.7618	54.6	61.8	65.0
1	21.82	0.7778	56.3	63.3	66.2
2	21.77	0.7764	56.9	64.0	66.7
4	21.77	0.7770	57.1	64.0	67.0
5	21.73	0.7768	55.9	63.2	66.7
3	**21.82**	**0.7778**	**57.4**	**64.1**	**67.3**

Table 4. Ablation of the number of Style Enhancement Blocks.

N	Quality Metric		Accuracy(%)		
	PSNR	SSIM	CRNN [25]	MORAN [16]	ASTER [26]
0	20.67	0.7431	53.2	61.0	64.7
3	21.45	0.7722	55.3	62.3	66.2
4	21.43	0.7751	56.7	63.9	66.9
6	21.74	0.7769	56.9	64.0	67.0
7	21.55	0.7756	56.0	62.8	66.2
5	**21.82**	**0.7778**	**57.4**	**64.1**	**67.3**

Table 5. Ablation of the number of Style Content Fuse Blocks.

Effectiveness of training with Text Style Loss. We show the effectiveness of the Text Style Loss in Table 2. Incorporating the Text Style Loss results in improvements across all metrics. For instance, the PSNR of TSRN increases from 21.42 to 21.56, and its SSIM improves from 0.7690 to 0.7711. In terms of text recognition accuracy, models incorporating the Text Style Loss demonstrate enhanced performance. The accuracy of TSRN improves from 41.4% to 43.7%. These results highlight that the Text Style Loss effectively enhances both image quality and text recognition accuracy. Besides, we also compared the impact of using different discriminators in Table 3. It can be seen that using CRNN performs better than VGG.

Numbers of SEBs. We show the effect of the number of SEBs in Table 4. As the number of SEBs increases, there is a noticeable improvement in both the PSNR and SSIM metrics. Specifically, using 3 SEBs achieves the highest PSNR of 21.82 and an SSIM of 0.7778, indicating the optimal enhancement of image quality. In terms of text recognition accuracy, models incorporating 3 SEBs also demonstrate superior performance across all recognition models. These findings suggest that SEBs significantly enhance both image quality and text recognition performance, with 3 SEBs being the optimal configuration.

Method	Accuracy of CRNN [25] (%)				Accuracy of MORAN [16] (%)				Accuracy of ASTER [26] (%)			
	Easy	Medium	Hard	**Avg**	Easy	Medium	Hard	**Avg**	Easy	Medium	hard	**Avg**
BICUBIC	36.4	21.1	21.1	26.8	60.6	37.9	30.8	44.1	67.4	42.4	31.2	48.2
HR	76.4	75.1	64.6	72.4	91.2	85.3	74.2	84.1	94.2	87.7	76.2	86.6
TSRN [30]	52.5	38.2	31.4	41.4	70.1	55.3	37.9	55.4	75.1	56.3	40.1	58.3
TBSRN [2]	59.6	47.1	35.3	48.1	74.1	57.0	40.8	58.4	75.7	59.9	41.6	60.1
TG [3]	61.2	47.6	35.5	48.9	75.8	57.8	41.4	59.4	77.9	60.2	42.4	61.3
TPGSR [18]	63.1	52.0	38.6	51.8	74.9	60.5	44.1	60.5	78.9	62.7	44.5	62.8
TATT [17]	62.6	53.4	39.8	52.6	72.6	60.2	43.1	59.5	78.9	63.4	45.4	63.6
C3-STISR [34]	65.2	53.6	39.8	53.7	74.2	61.0	43.2	59.5	79.1	63.3	46.8	64.1
TSAN [37]	64.6	53.3	38.8	53.0	78.4	61.3	45.1	62.7	79.6	64.1	45.3	64.1
LEMMA [7]	67.1	58.8	40.6	56.3	77.7	64.4	44.6	63.2	81.1	66.3	47.4	66.0
DPMN [36]	64.4	54.2	39.2	53.4	73.3	61.5	43.9	60.4	79.3	64.1	45.2	63.9
TCDM [20]	67.3	57.3	**42.7**	55.7	77.6	62.9	45.9	62.2	81.3	65.1	**50.1**	65.5
RTSRN [33]	67.0	59.2	42.6	57.0	77.1	63.3	**46.5**	63.2	80.4	66.1	49.1	66.2
StyleSRN	**68.1**	**59.4**	42.4	**57.4**	**78.6**	**65.1**	45.7	**64.1**	**82.7**	**67.4**	48.7	**67.3**

Table 6. Comparison with the existing methods in terms of the recognition accuracy on TextZoom[30]. BICUBIC means LR images are directly upsampled by the bicubic interpolation, and the same definition applies to BICUBIC in subsequent tables.

Method	Loss	PSNR				SSIM			
		Easy	Medium	Hard	**Avg**	Easy	Medium	Hard	**Avg**
BICUBIC	-	22.35	18.98	19.39	20.35	0.7884	0.6254	0.6592	0.6961
TSRN [30]	$\mathcal{L}_2 + \mathcal{L}_{GP}$	**25.07**	18.86	19.71	21.42	0.8897	0.6676	0.7302	0.7690
TBSRN [2]	$\mathcal{L}_{POS} + \mathcal{L}_{CON}$	23.46	19.17	19.68	20.91	0.8729	0.6455	0.7452	0.7603
TPGSR [18]	$\mathcal{L}_2 + \mathcal{L}_{TP}$	23.73	18.68	20.06	20.97	0.8805	0.6738	0.7440	0.7719
TATT [17]	$\mathcal{L}_2 + \mathcal{L}_{TP} + \mathcal{L}_{TSC}$	24.72	19.02	20.31	21.52	**0.9006**	**0.6911**	**0.7703**	**0.7930**
LEMMA [7]	$\mathcal{L}_2 + \mathcal{L}_{TP} + \mathcal{L}_{finetune}$	23.70	19.37	19.84	20.88	0.8753	0.6905	0.7491	0.7760
StyleSRN	$\mathcal{L}_2 + \mathcal{L}_{TP} + \mathcal{L}_{style}$	24.69	**19.70**	**20.58**	**21.82**	0.8844	0.6791	0.7531	0.7778

Table 7. Comparison of the image quality on TextZoom [30]. Avg represents the average value of three subsets.

Numbers of SCFBs. We further examined the impact of the numbers of SCFBs , as detailed in Table 5. While increasing the number of SCFBs improves recognition accuracy metrics, diminishing returns are observed beyond a certain point. The optimal performance, consistent with the SEB ablation study, is achieved with 5 SCFBs, yielding a PSNR of 21.82 and an SSIM of 0.7778. Moreover, text recognition accuracy peaks with CRNN (57.40%), MORAN (64.14%), and ASTER (67.31%) when 5 SCFBs are utilized. This suggests that while the SCFB is essential for balancing style and content, its optimal configuration involves a moderate number of blocks, beyond which performance gains begin to plateau.

Method	Accuracy(%)				
	IC13	IC15	CT80	SVTP	III5K
BICUBIC	76.16	38.21	55.21	35.50	59.70
TSRN [30]	75.67	41.52	54.86	35.97	62.60
TBSRN [2]	76.95	43.07	54.86	37.36	63.60
TPGSR [18]	77.64	42.85	53.47	38.29	62.07
TATT [17]	76.06	48.48	51.04	42.02	62.67
LEMMA [7]	79.80	49.70	51.04	43.72	65.07
StyleSRN	**80.99**	**50.47**	**55.56**	**46.51**	**65.63**

Table 8. Comparison of the accuracy on manually degrade scene text recognition datasets. Accuracy represents the average accuracy with CRNN [25].

4.5. Comparison with State-of-the-Arts

Results on TextZoom. We evaluate the recognition accuracy of text images generated by various STISR methods using three text recognizers: CRNN [25], MORAN [16], and ASTER [26]. As shown in Table 7, our proposed StyleSRN consistently outperforms state-of-the-art methods across all recognizers, achieving average accuracy rates of 57.4% with CRNN, 64.1% with MORAN, and 67.3% with ASTER. On the hard subset, StyleSRN performs slightly lower than some methods, such as TCDM and DPMN, likely due to challenges in balancing structure and style preservation in extremely blurred or occluded cases. In terms of image quality, as shown in Table 7, our method achieves the highest average PSNR of 21.82, surpassing other methods like TATT (21.52) and LEMMA (20.93). Our method achieves an SSIM of 0.7778, which is comparable to TATT (0.7930). The visual comparison is presented in Figure 3, which demonstrates that our method achieves superior visual quality. Our method not only restores the structure of the characters but also preserves their style information. For instance, for certain stereoscopic scene texts shown in Figure 3, our method more effectively restores their stereoscopic effect and shadows details.

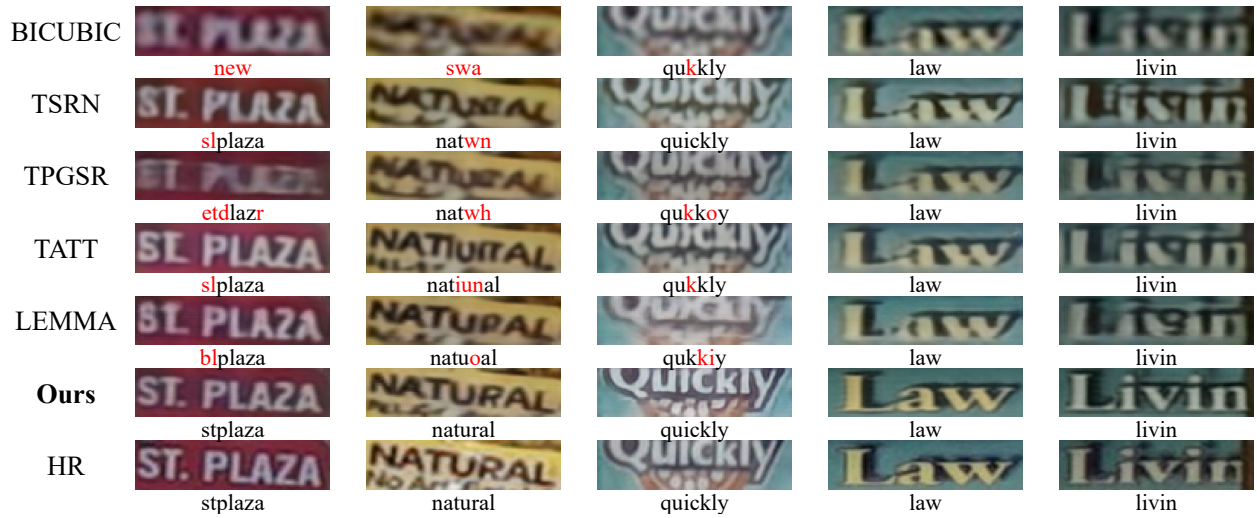

Figure 3. Visualization of SR images and their recognition result. Each text image below uses CRNN [25] for recognition. Characters marked in red indicate incorrect recognition results.

Figure 4. Visualization of the STISR results on scene text with uneven illumination and scene text with severe blur.

Generalization to recognition datasets. To further evaluate the generalization capability and robustness of the proposed StyleSRN, we also conduct experiments on five scene text recognition datasets. These datasets cover a broad spectrum of image styles and label distributions, presenting substantial challenges for the generalization performance of super-resolution networks. Due to the absence of paired low-resolution and high-resolution images in these datasets, we manually degrad the original images before applying StyleSRN for preprocessing. As illustrated in Table 2, our method consistently surpasses previous approaches across these datasets, confirming its strong generalization ability.

5. Discussion

While StyleSRN effectively improves the structural integrity and style consistency of STISR ,as shown in Figure 4, it struggles to recover character structures in images with uneven illumination or severe blur, highlighting the need for further refinement to address degraded visual conditions. Another limitation is the absence of a standardized quantitative metric for evaluating style degradation in STISR methods. While our method restores the style of low-resolution scene text images, no existing metric specifically measures style preservation or degradation. This lack of objective criteria makes it difficult to rigorously assess style restoration performance, as current evaluations are largely subjective and based on visual inspection. Future research should focus on developing standardized, quantitative metrics to objectively evaluate style preservation in STISR, enabling more precise assessments and meaningful comparisons across methods.

6. Conclusions

In this work, we propose StyleSRN to address the limitations of current STISR methods that neglect the stylistic details of scene text images. StyleSRN incorporates Style Enhancement Blocks and Style Content Fuse Blocks to effectively capture and integrate text style with content, ensuring both structural integrity and stylistic consistency. Additionally, the proposed Text Style Loss supervises reconstruction at the style level, which effectively enhances both image quality and text recognition accuracy. Results on TextZoom and five STR benchmarks demonstrate that StyleSRN consistently enhances downstream recognition accuracy and image quality.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.61502164, 62176097), the Natural Science Foundation of Hunan Province (No.2020JJ4057, 2024JJ10034), the Key Research and Development Program of Changsha Science and Technology Bureau (No.kq2004050), and the Scientific Research Foundation of the Education Department of Hunan Province of China (No.21A0052).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1

[2] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2021. 1, 2, 5, 6, 7

[3] Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 285–293, 2022. 2, 7

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 184–199, 2014. 2

[5] 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

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2, 3, 5

[7] Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. Towards robust scene text image super-resolution via explicit location enhancement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 782–790, 2023. 1, 2, 6, 7

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3, 4

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Proceedings of the Advances in Neural Information Processing Systems*, 28, 2015. 3

[10] 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 *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013. 5

[11] 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 *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015. 5

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2, 3

[13] 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. 1

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[16] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 5, 6, 7

[17] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2022. 1, 2, 4, 6, 7

[18] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023. 1, 2, 6, 7

[19] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2012. 5

[20] Chihiro Noguchi, Shun Fukuda, and Masao Yamanaka. Scene text image super-resolution based on text-conditional diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1474–1484, 2024. 2, 7

[21] 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. 2

[22] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013. 5

[23] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2119–2127, 2023. 3

[24] 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. 5

[25] 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. 5, 6, 7, 8

[26] 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. 5, 6, 7

[27] Tonghua Su, Fuxiang Yang, Xiang Zhou, Donglin Di, Zhongjie Wang, and Songze Li. Scene style text editing, 2023. 3

[28] Hanh TM Tran and Tien Ho-Phuoc. Deep laplacian pyramid network for text images super-resolution. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6, 2019. 2

[29] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020. 4

[30] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 650–666, 2020. 1, 2, 4, 5, 6, 7

[31] 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. 5

[32] Chongsheng Zhang, Weiping Ding, Guowen Peng, Feifei Fu, and Wei Wang. Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4727–4743, 2020. 1

[33] Wenyu Zhang, Xin Deng, Baojun Jia, Xingtong Yu, Yifan Chen, Jin Ma, Qing Ding, and Xinming Zhang. Pixel adapter: A graph-based post-processing approach for scene text image super-resolution. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2168–2179, 2023. 7

[34] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: scene text image super-resolution with triple clues. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1707–1713, 2022. 7

[35] Yuxuan Zhou, Liangcai Gao, Zhi Tang, and Baole Wei. Recognition-guided diffusion model for scene text image super-resolution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2940–2944, 2024. 2

[36] Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. Improving scene text image super-resolution via dual prior modulation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3843–3851, 2023. 7

[37] Xiangyuan Zhu, Kehua Guo, Hui Fang, Rui Ding, Zheng Wu, and Gerald Schaefer. Gradient-based graph attention for scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3861–3869, 2023. 7