PEAN: A Diffusion-Based Prior-Enhanced Attention Network for Scene Text Image Super-Resolution

Zuoyan Zhao
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
zuoyanzhao@seu.edu.cn

Hui Xue*
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
hxue@seu.edu.cn

Pengfei Fang
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
fangpengfei@seu.edu.cn

Shipeng Zhu
School of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
Nanjing, China
shipengzhu@seu.edu.cn

Abstract

Scene text image super-resolution (STISR) aims at simultaneously increasing the resolution and readability of low-resolution scene text images, thus boosting the performance of the downstream recognition task. Two factors in scene text images, visual structure and semantic information, affect the recognition performance significantly. To mitigate the effects from these factors, this paper proposes a **P**rior-**E**nhanced **A**ttention **N**etwork (**PEAN**). Specifically, an attention-based modulation module is leveraged to understand scene text images by neatly perceiving the local and global dependence of images, despite the shape of the text. Meanwhile, a diffusion-based module is developed to enhance the text prior, hence offering better guidance for the SR network to generate SR images with higher semantic accuracy. Additionally, a multi-task learning paradigm is employed to optimize the network, enabling the model to generate legible SR images. As a result, PEAN establishes new SOTA results on the TextZoom benchmark. Experiments are also conducted to analyze the importance of the enhanced text prior as a means of improving the performance of the SR network. Code is available at https://github.com/jdfxzzy/PEAN.

CCS Concepts

• **Computing methodologies → Reconstruction**.

* Corresponding author: Hui Xue.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680974

Keywords

Scene Text Image, Super-Resolution, Vision Backbone, Diffusion Models

ACM Reference Format:
Zuoyan Zhao, Hui Xue, Pengfei Fang, and Shipeng Zhu. 2024. PEAN: A Diffusion-Based Prior-Enhanced Attention Network for Scene Text Image Super-Resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680974

1 Introduction

Scene text recognition (STR) focuses on extracting text from images, which has been widely applied in automatic driving [39], intelligent transportation [1], *etc.* However, in real-world applications, a variety of reasons result in captured images being low-resolution (LR), such as the quality of the lens, motion blur, and shaking when capturing photos, leading to blurred text in images. To better read text from such images, researchers formulate the STISR task to reconstruct missing text details in LR images, as a pre-processing step for STR.

For scene text images, two crucial factors determine whether they could be correctly recognized, *i.e.*, visual structure and semantic information [12, 59]. Early attempts at STISR concentrate on adequately recovering the visual structure of LR scene text images [11, 53, 62]. Composed of several CNN-BiLSTM layers, these methods can learn from paired LR-HR images to improve the resolution and readability of scene text images simultaneously. However, the performance is limited due to the fact that they ignore the semantic information of scene text images. This factor has been utilized in recent advancements. These works observe that semantic information plays an important role in guiding the restoration of correct visual structure, and propose numerous text

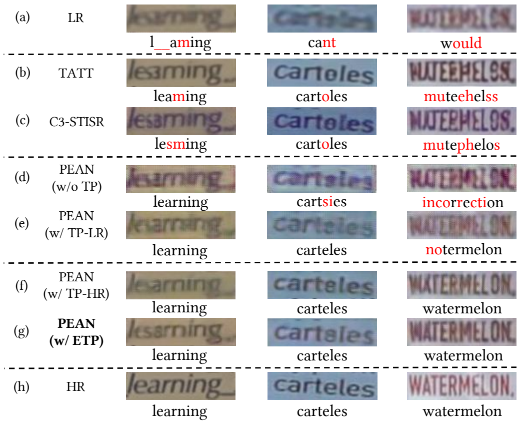


Figure 1: Comparison between previous text prior-based STISR methods (row (b, c)) and PEAN. The incorporation of AMM enables PEAN to restore the visual structure of lengthy text in images. However, its performance is limited by the absence of semantic information (row (d)). The introduction of TP-LR partially addresses this limitation, yet its efficacy remains inadequate, leading to several failure cases (row (e)). Considering that TP-HR is a robust alternative, we conduct an exploratory experiment by substituting TP-HR with TP-LR, resulting in superior performance (row (f)). This inspires us to design a module for enhancing the TP-LR so as to obtain the ETP, which demonstrates comparable effectiveness to TP-HR in guiding the SR process (row (g)).

prior-based methods [16, 29, 30, 63]. That is, the text prior, generated by pre-trained STR models, is leveraged to facilitate the SR process [16, 29, 30, 63], thereby generating correct characters of text in SR images.

Despite improved performance achieved by these approaches, the dominance of visual structure and semantic information persists, as two critical issues in previous studies remain unresolved. Firstly, previous STISR methods [3, 4, 30, 53, 63] rely on Sequential Residual Blocks (SRB) to extract visual features. This module, containing several CNN-BiLSTM layers, *has difficulty in restoring the complete visual structure of images containing long or deformed text string due to its inherent demerits*, *i.e.*, the performance bottleneck of capturing long-range dependence [10, 38]. Secondly, *the introduction of the primary text prior, originating from the interference of low-quality images on recognizers, prevents the SR network from generating images that contain correct semantic information.* Recently, C3-STISR [63] has employed a language model [12] into STISR, utilizing its learned linguistic knowledge to rectify the text prior. Although the rectified prior demonstrates some effectiveness, it lacks sufficient strength in guiding the SR network to produce images with high semantic accuracy.

We propose a **P**rior-**E**nhanced **A**ttention **N**etwork (**PEAN**) to tackle issues caused by the two factors. To begin with, an Attention-based Modulation Module (AMM) is proposed to substitute the SRB, endowing the network with a larger receptive field to images, thereby restoring the visual structure of images with text in various shapes and lengths (shown in Figure 1(d)). Horizontal and vertical strip-wise attention mechanisms [9, 18, 49] are employed in AMM. Among them, the horizontal attention mechanism can capture the dependence between characters while the vertical attention mechanism can capture the structural information within a character [62]. However, the lack of semantic information limits the capability of such model. As demonstrated by previous works [6, 8], leveraging strong prior information to restrict the solution space plays a vital role in SR problems. Notably, the text prior derived from high-resolution (HR) images is a robust choice for STISR, in view of the high recognition accuracy of HR images. Consequently, we conduct an exploratory experiment wherein we substitute the text prior from LR images (TP-LR) with the text prior from HR images (TP-HR) within such model, yielding superior outcomes (see Figures 1(e) and (f) for comparison, details can be found in § 4.4.1). This inspires the design of a module for enhancing the primary text prior, resulting in the creation of the Enhanced Text Prior (ETP), which is comparable in effectiveness to TP-HR (shown in Figures 1(f) and (g)). The ETP provides valuable guidance to the SR network, promoting the generation of SR images with high semantic accuracy. Given the remarkable performance of diffusion models [17, 47], we propose a diffusion-based Text Prior Enhancement Module (TPEM) to obtain the ETP owing to their ability to map complex distributions [57]. In addition, considering that the goal of STISR is to increase the resolution and readability of LR scene text images, we adopt the Multi-Task Learning (MTL) paradigm in the training phase, where the image restoration task aims at generating high-quality SR images, and the text recognition task stimulates the model to generate more readable SR results. In a nutshell, main contributions of our work are three-fold:

- We devise an AMM containing horizontal and vertical attention mechanisms to model the long-range dependence in scene text images, thereby recovering the visual structure of images with long or deformed text.
- A diffusion-based TPEM is further proposed to enhance the primary text prior. The resulting ETP guides the SR network to generate SR images with improved semantic accuracy.
- Empirical studies show that PEAN attains the SOTA performance on the TextZoom [53] benchmark. We also conduct experiments to explore the reasons behind the performance improvement of the SR network.

2 Related Work

2.1 Scene Text Image Super-Resolution

Scene text image super-resolution (STISR) has received surging attention in the computer vision community. Different from the classic single image super-resolution (SISR) task, STISR aims at increasing the resolution and legibility of scene text images simultaneously [69], serving as a pre-processing method for the downstream recognition task.

The milestone works in STISR are the TextZoom benchmark and the TSRN model [53], which promote the development of follow-up approaches. We roughly classify them into two categories. One category of methods focuses on recovering the visual structure of LR scene text images. Among them, TSRN [53] and PCAN [62] use several CNN-BiLSTM blocks to complete the SR process. TSAN [70] adopts a gradient-based graph attention method to extract more

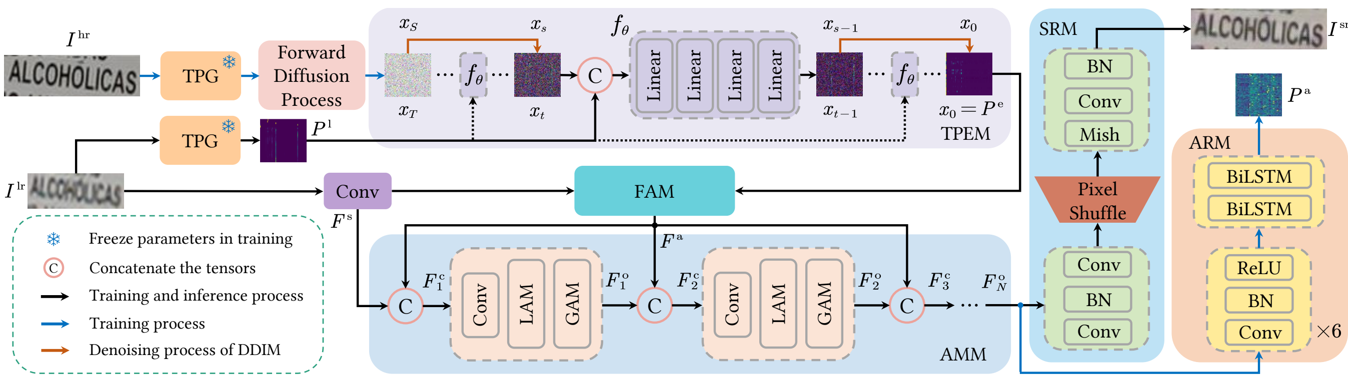


Figure 2: Overview of the architecture of our proposed Prior-Enhanced Attention Network (PEAN).

effective representations for STISR. Another category considers the semantic information as the text prior to guide the SR process. In this category, TPGSR [29] and TATT [30] utilize the pre-trained recognizer to generate the text prior from LR images, boosting the model to generate SR images with correct text. C3-STISR [63] employs a language model [12] to rectify the text prior and uses triple clues to realize STISR.

2.2 Scene Text Recognition

Scene text recognition (STR) aims at reading text contents from natural images. The pioneering work CRNN [43] uses the CNN-BiLSTM framework and the CTC [15] loss to perform STR for the first time. ASTER [44] further exploits the Thin-Plate Spline (TPS) transformation [19] to understand scene text images with deformed or irregular layouts. The concurrent work MORAN [28] also handles these cases via a multi-object rectification network.

Recently, language models have been integrated into STR models and the fusion of vision and language features shows a great potential to improve the scene text understanding. SRN [59] employs an autoregressive language model to rectify the recognition results generated by visual features. Additionally, ABINet [12] shows that masked language models [7], capable of providing bidirectional representations, constitute another effective option for rectification. Furthermore, PARSeq [2] creatively adopts an internal language model as a spell-checker, eliminating the need for the pre-training process in ABINet [12].

2.3 Diffusion Models

In computer vision, diffusion models [17] emerge as robust probabilistic generative models, facilitating tasks like image synthesis [14, 40], text-to-image synthesis [33, 41], image restoration [13, 57] and image inpainting [68] through the iterative diffusion of information among pixels. Recently, diffusion models have also been employed in the super-resolution task. SR3 [42] is the pioneering work that applies diffusion models to SISR. TextDiff [25] represents the initial attempt at a diffusion-based model designed specifically for STISR, focusing on enhancing the visual structure of text within images by refining their contours for a more natural appearance.

In contrast to existing methods, the proposed PEAN uses the diffusion-based TPEM to provide the SR network with enhanced semantic guidance, further resulting in SR images with heightened semantic accuracy.

3 Methodology

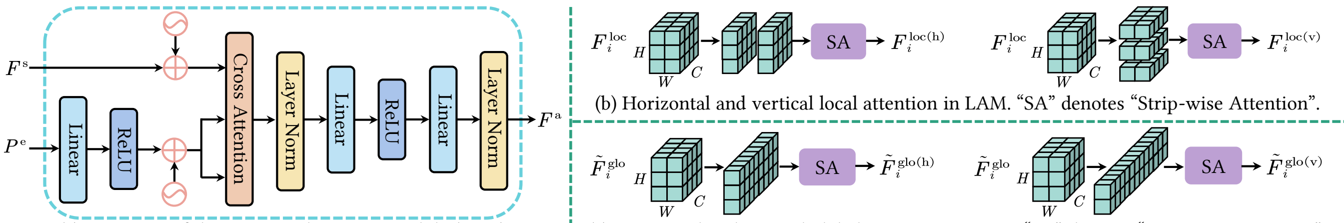
This section first gives an overview of the Prior-Enhanced Attention Network (PEAN). Then we present the proposed Text Prior Enhancement Module (TPEM), Attention-based Modulation Module (AMM) and the Multi-Task Learning (MTL) paradigm.

3.1 Overall Architecture

The pipeline of our proposed PEAN is shown in Figure 2. Given one LR image $I^{\mathrm{lr}} \in \mathbb{R}^{H\times W\times C}$, the Text Prior Generator (TPG) outputs the recognition probability sequence as the primary text prior P^{l}. Then, the diffusion-based TPEM refines it to obtain the ETP denoted as P^{e}, which can assist the SR network to generate SR images with improved semantic accuracy. Concurrently, a convolutional layer is adopted to extract the shallow visual feature F^{s} from I^{lr}, which is then aligned with the refined text prior by a Feature Alignment Module (FAM). Then an AMM with N blocks is introduced to mine the internal dependence between characters in the image, thereby facilitating the SR process. For the i^{th} block of AMM (*i.e.*, B_i), its output F_i^{o} is firstly concatenated with the aligned feature (*i.e.*, F^{a}) in the channel dimension to get F_{i+1}^{c}. The fusion feature is then sent into B_{i+1} for further processing. Finally, a Super-Resolution Module (SRM) containing several convolutional and batch normalization layers, receives F_N^{o} as input and utilizes a PixelShuffle [45] operation to generate the SR image $I^{\mathrm{sr}} \in \mathbb{R}^{2H\times 2W\times C}$. Notably, in the training phase, F_N^{o} is also sent into an Auxiliary Recognition Module (ARM), which outputs the recognition probability sequence of the SR image. The outputs of TPEM, SRM and ARM enable the optimization of the model in an Multi-Task Learning (MTL) paradigm, steering the model to generate plausible and readable SR images.

3.2 Text Prior Enhancement Module

As demonstrated by previous works [6, 8], strong prior information plays a pivotal role in solving SR problems, while the primary text prior applied in previous works is not powerful enough because it originates from LR scene text images. Our exploratory experiments, detailed in § 4.4.1, also underline the influential role of TP-HR in



(a) Overview of the Feature Alignment Module (FAM).

(b) Horizontal and vertical local attention in LAM. "SA" denotes "Strip-wise Attention".

(c) Horizontal and vertical global attention in GAM. "SA" denotes "Strip-wise Attention".

Figure 3: Overview of the architecture of the FAM and the strip-wise attention mechanism inside LAM and GAM.

guiding the SR network to generate images with improved semantic accuracy for the model. Therefore, we introduce the TPEM to obtain the ETP, which can effectively guide the SR network, similarly to the efficacy of TP-HR. The core component of TPEM is the denoiser, denoted as f_θ, which leverages the reverse diffusion process [17] to estimate the enhanced prior, providing substantial semantic guidance to the SR network.

3.2.1 ***Forward Diffusion Process***. In the training phase, with a given HR image, denoted as $I^{\mathrm{hr}} \in \mathbb{R}^{2H\times 2W\times C}$, the TPG generates a sequence of recognition probabilities, referred to as $P^{\mathrm{h}} \in \mathbb{R}^{L\times|\mathcal{A}|}$, serving as our ground truth. L is the length of the sequence and $|\mathcal{A}|$ is the cardinality of the recognizable letter set. Consequently, in line with Ho *et al.* [17], we incrementally introduce Gaussian noise denoted as ϵ to the initial variable $x_0 = P^{\mathrm{h}}$ based on the timestamp, as follows:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, \left(1-\alpha_t\right)\boldsymbol{I}\right), \tag{1}$$

Here, α_t is a hyperparameter that controls the variance of the added Gaussian noise at each time step. Leveraging the reparameterization trick [21], we can express x_t as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{2}$$

where $\alpha_i \in [0,1]$, $\bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i$, $t = 1, 2, \cdots, T$. As $T \to \infty$, x_T converges to an isotropic Gaussian distribution. Consequently, during the inference phase, the forward diffusion process simplifies to initializing $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

3.2.2 ***Reverse Diffusion Process***. In the reverse diffusion process, Gaussian noise gradually transforms into the ETP, denoted as $P^{\mathrm{e}} \in \mathbb{R}^{L\times|\mathcal{A}|}$, conditioned on the primary text prior $P^{\mathrm{l}} \in \mathbb{R}^{L\times|\mathcal{A}|}$. The latter is the output recognition probability sequence of the TPG, with I^{lr} as input. This process can be formulated as follows:

$$p_\theta\left(x_{t-1} \mid x_t, P^{\mathrm{l}}\right) = q\left(x_{t-1} \mid x_t, f_\theta\left(x_t, P^{\mathrm{l}}, t\right)\right), \tag{3}$$

where f_θ is an MLP-based denoising network, the architecture of which can be found in the Supplementary Material. Similar to previous works [25, 57, 58, 64], we opt to directly estimate P^{e} instead of ϵ for performance improvement. Experiments in § 4.4.1 verifies the effectiveness of this design. The whole process is supervised by the MAE and CTC loss [15], given by:

$$\mathcal{L}_{\mathrm{diff}} = \lambda_1\underbrace{\left\|P^{\mathrm{h}} - P^{\mathrm{e}}\right\|_1}_{\mathcal{L}_{\mathrm{mae}}} + \lambda_2\mathcal{L}_{\mathrm{ctc}}^{\mathrm{t}}, \tag{4}$$

where λ_1 and λ_2 are the weight of the two losses. During the training phase, this reverse sampling process is a Markov process containing T steps, which is computationally intensive. Therefore, in the inference phase, we adopt the sampling strategy of DDIM [47] with S steps ($S \ll T$) following previous works [68]. Experiments in the Supplementary Material validate the effectiveness and efficiency of this design.

Besides, a feature alignment process between the shallow visual feature and the ETP is required to facilitate the SR process. We design a Feature Alignment Module (FAM) to obtain the aligned feature $F^{\mathrm{a}} \in \mathbb{R}^{H\times W\times C_1}$, where C_1 is the dimension of the shallow visual feature. The architecture of the FAM is shown in Figure 3(a).

3.3 Attention-Based Modulation Module

We design an Attention-based Modulation Module (AMM) to capture long-range dependence, thereby effectively restoring the visual structure of images with long or deformed text. As can be seen in Figure 2, AMM contains N blocks, each block (*i.e.*, B_i) including a simple convolutional layer, a Local Attention Module (LAM) and a Global Attention Module (GAM). For B_i, it receives the output of B_{i-1}, which is then concatenated with F^{a} channel-wisely to get F_i^{c}, serving as the input of this block. Since the dimension of F_i^{c} is located in $\mathbb{R}^{H\times W\times 2C_1}$, a convolutional layer is adopted to project it to the same space as F_{i-1}^{o}, as:

$$F_i^{\mathrm{loc}} = \mathrm{Conv}\left(\mathrm{Concat}\left(F_{i-1}^{\mathrm{o}}, F^{\mathrm{a}}\right)\right) \in \mathbb{R}^{H\times W\times C_1}. \tag{5}$$

Firstly, a LAM is adopted to model the local similarity between intra- and inter-character features. Considering that for a scene text image, the horizontal contexts contain correlation information between characters while the vertical contexts contain internal features inside a character, such as the stroke information [53, 62], we propose to perform the strip-wise attention mechanism [9, 18, 49] on the fusion feature to capture long-range dependence. Taking the horizontal attention as an example. As shown in Figure 3(b), F_i^{loc} is split into W strips in the width dimension and the attention mechanism is applied to each of the W strips. The resulting feature is then concatenated to form the horizontal feature $F_i^{\mathrm{loc(h)}} \in \mathbb{R}^{H\times W\times C_1}$. Likewise, the vertical attention mechanism can also result in the vertical feature $F_i^{\mathrm{loc(v)}} \in \mathbb{R}^{H\times W\times C_1}$. $F_i^{\mathrm{loc(h)}}$ and $F_i^{\mathrm{loc(v)}}$ are concatenated in the channel dimension and a convolutional layer is used to fuse them. Then the FFN with residual connection is introduced to perform non-linear transformation, given by:

$$F_i^{\mathrm{glo}} = F_i^{\mathrm{loc}} + \mathrm{Conv}\left(\mathrm{Concat}\left(F_i^{\mathrm{loc(h)}}, F_i^{\mathrm{loc(v)}}\right)\right), \tag{6}$$

$$\tilde{F}_i^{\text{glo}} = F_i^{\text{glo}} + \text{FFN}\left(F_i^{\text{glo}}\right) \in \mathbb{R}^{H\times W\times C_1}. \quad (7)$$

However, the interaction between character-level features is limited in a local manner [67]. To encourage global interaction, we employ the strip-wise GAM as the complementation of the LAM, thus further modeling the global similarity between intra- and inter-character features. Specifically, the width dimension of $\tilde{F}_i^{\text{glo}}$ is mer-ged with the channel dimension, leading to a feature map located in $\mathbb{R}^{H\times C_1 W}$, as can be seen in Figure 3(c). The strip-wise attention mechanism is imposed on it to capture the global inter-character dependence, resulting in the feature $\tilde{F}_i^{\text{glo(h)}} \in \mathbb{R}^{H\times C_1 W}$. Likewise, the height dimension of $\tilde{F}_i^{\text{glo}}$ is merged with the channel dimension, and the attention mechanism is applied to this feature map, leading to the feature $\tilde{F}_i^{\text{glo(v)}} \in \mathbb{R}^{W\times C_1 H}$. With the aid of them, the horizontal and vertical attention mechanisms can work globally to promote global-level interaction of character features. Similar workflow as Eq. (6) and (7) is adopted to get the final feature of B_i, namely $F_i^{\text{o}} \in \mathbb{R}^{H\times W\times C_1}$.

3.4 Multi-Task Learning

While previous researches [16, 29, 30, 60, 63] adopt the MTL paradigm by incorporating an additional text prior loss [30] to fine-tune the TPG for better text priors, our approach differs in the utilization of the MTL paradigm. We employ this paradigm not only to ensure image quality but also to facilitate text recognition. This distinction arises from the fact that STISR aims to simultaneously enhance the resolution and legibility of scene text images [69].

3.4.1 ***Image Restoration Task.*** To begin with, we employ an image restoration task to generate high-quality SR images. As can be seen in Figure 3, the output of the last block of AMM, *i.e.*, F_N^{o}, is sent into a Super-Resolution Module (SRM) to get the SR image $I^{\text{sr}} \in \mathbb{R}^{2H\times 2W\times C}$. It uses several convolutional layers with batch normalization and Mish [32] activation function for feature refinement and adopts a PixelShuffle [45] operation to increase the resolution. To ensure the pixel-level and structure-level coherence between I^{sr} and I^{hr}, the Mean-Square-Error (MSE) loss and the Stroke-Focused Module (SFM) loss [4] are applied between the two images, formulated as:

$$\mathcal{L}_{\text{img}} = \underbrace{\lambda_3 \left\| I^{\text{hr}} - I^{\text{sr}} \right\|_2^2}_{\mathcal{L}_{\text{mse}}} + \underbrace{\lambda_4 \left\| A^{\text{hr}} - A^{\text{sr}} \right\|_1}_{\mathcal{L}_{\text{sfm}}}, \quad (8)$$

where A is the attention map fetched from a Transformer-based recognizer [4]. λ_3 and λ_4 are the weights of the MSE and SFM loss respectively.

3.4.2 ***Text Recognition Task.*** Considering that high image quality may not be equal to superior recognition results [16, 25], we introduce an Auxiliary Recognition Module (ARM), steering the model to generate SR images with promising legibility. In the inference phase, this module is abandoned, causing no extra computation complexity.

We exploit the classic and effective CNN-BiLSTM architecture applied in the STR task to design the ARM for simplicity [43]. Its output is a recognition probability sequence on the basis of F_N^{o} and we denote it as $P^{\text{a}} \in \mathbb{R}^{L\times|\mathcal{A}|}$, where L is the length of the sequence and $|\mathcal{A}|$ is the cardinality of the recognizable letter set. The CTC [15] loss denoted as $\mathcal{L}_{\text{ctc}}^{\text{a}}$ is imposed between P^{a} and the ground truth for better optimization. In a word, the total loss of the text recognition task is:

$$\mathcal{L}_{\text{txt}} = \lambda_5 \mathcal{L}_{\text{ctc}}^{\text{a}}, \quad (9)$$

where λ_5 is the weight of the CTC [15] loss for the text recognition task. In the training phase, we optimize the parameters of PEAN. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}. \quad (10)$$

4 Experiments

In this section, we first introduce the datasets and evaluation metrics. Then the implementation details are thoroughly described. Subsequently, comprehensive experiments are conducted to demonstrate that our proposed PEAN is an effective alternative for STISR.

4.1 Datasets and Evaluation Metrics

We conduct experiments on the TextZoom [53] benchmark. However, due to some inherent drawbacks of it, these experiments are not enough to reflect that PEAN surely has the ability to restore the complete visual structure. Therefore, we select 651 images with the resolution no greater than 16×64 as LR images from the IIIT5K [31], SVTP [37] and IC15 [20] datasets for evaluation. Details of the datasets and drawbacks of TextZoom can be found in the Supplementary Material. For TextZoom, following previous works [3, 4, 16, 29, 30, 53, 60, 63], we adopt ASTER [44], MORAN [28] and CRNN [43] for evaluation. In the Supplementary Material, we also adopt three recent Transformer-based recognizers, namely, MGP-STR [52], ABINet [12] and VisionLAN [55] for evaluation. For the constructed dataset, PSNR and SSIM [56] are selected as metrics to evaluate the image quality.

4.2 Implementation Details

Our model is implemented with the Pytorch 1.10 deep learning library [36]. All of the experiments are conducted on 1 NVIDIA TITAN RTX GPU. In the training phase, the model is trained for 200 epochs and optimized by the AdamW [27] optimizer. The learning rate and the size of the mini-batch are set as 0.001 and 32 respectively. The weight of each loss is set as $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.8$, $\lambda_4 = 75$, $\lambda_5 = 1$. In terms of the network architecture, the number of blocks in AMM is 6. The sampling timestep in the TPEM is set as 1. Following CRNN [43], the number of convolutional layers applied in the ARM is 6. We first drop the TPEM and pre-train the model with TP-HR, then the TPEM is introduced and weights of parameters obtained by pre-training are initialized to continue the fine-tuning process of the model. Of note, if this setting is abandoned, PEAN can still achieve SOTA performance on the TextZoom [53] dataset. Ablation studies about the weight of loss functions and the pre-training and fine-tuning setting can be found in the Supplementary Material. In the main paper, we study our method on top of PARSeq [2] as the TPG. In the Supplementary Material, we also study the cases when CRNN [43] and ABINet [12] are adopted as the TPG.

Table 1: The recognition accuracy of some mainstream STISR methods on the three subsets of TextZoom. Best scores are bold.

Methods	Accuracy of ASTER [44] (%)				Accuracy of MORAN [28] (%)				Accuracy of CRNN [43] (%)			
	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
LR	62.4	42.7	31.6	46.6	59.4	36.0	28.2	42.3	37.5	21.2	21.4	27.3
SRCNN [8]	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 [24]	69.6	47.6	34.3	51.3	60.7	42.9	32.6	46.3	39.7	27.6	22.7	30.6
RDN [61]	70.0	47.0	34.0	51.5	61.7	42.0	31.6	46.1	41.6	24.4	23.5	30.5
RRDB [54]	70.9	44.4	32.5	50.6	63.9	41.0	30.8	46.3	40.6	22.1	21.9	28.9
LapSRN [23]	71.5	48.6	35.2	53.0	64.6	44.9	32.2	48.3	46.1	27.9	23.6	33.3
ESRT [8]	69.8	49.1	35.2	52.5	61.9	41.7	32.2	46.3	48.2	27.9	25.8	34.8
Omni-SR [51]	71.2	52.3	38.1	54.9	66.7	47.9	36.5	51.4	54.8	37.4	29.4	41.4
SRFormer [66]	69.0	45.1	32.8	50.2	61.3	39.6	29.9	44.7	41.0	22.8	22.9	29.6
TSRN [53]	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 [3]	75.7	59.9	41.6	60.1	74.1	57.0	40.8	58.4	59.6	47.1	35.3	48.1
PCAN [62]	77.5	60.7	43.1	61.5	73.7	57.6	41.0	58.5	59.6	45.4	34.8	47.4
TG [4]	77.9	60.2	42.4	61.3	75.8	57.8	41.4	59.4	61.2	47.6	35.5	48.9
SGENet [48]	75.8	60.7	45.0	61.4	71.5	56.2	41.4	57.3	59.4	47.9	37.7	49.0
TPGSR [29]	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 [30]	78.9	63.4	45.4	63.6	72.5	60.2	43.1	59.5	62.6	53.4	39.8	52.6
C3-STISR [63]	79.1	63.3	46.8	64.1	74.2	61.0	43.2	60.5	65.2	53.6	39.8	53.7
TATT + DPMN [69]	79.3	64.1	45.2	63.9	73.3	61.5	43.9	60.4	64.4	54.2	39.2	53.4
TSAN [70]	79.6	64.1	45.3	64.1	78.4	61.3	45.1	62.7	64.6	53.3	38.8	53.0
TEAN [46]	80.4	64.5	45.6	64.6	76.8	60.8	43.4	61.4	63.7	52.5	38.1	52.2
MSPIE [71]	80.4	63.4	46.3	64.4	74.0	61.4	44.4	60.8	64.5	54.2	39.6	53.5
TCDM [34]	81.3	65.1	50.1	65.5	77.6	62.9	45.9	62.2	67.3	57.3	42.7	55.7
LEMMA [16]	81.1	66.3	47.4	66.0	77.7	64.4	44.6	63.2	67.1	58.8	40.6	56.3
RTSRN [60]	80.4	66.1	49.1	66.2	77.1	63.3	46.5	63.2	67.0	59.2	42.6	57.0
RGDiffSR [65]	81.1	65.4	49.1	66.2	78.6	62.1	45.4	63.1	67.6	56.5	42.7	56.4
TextDiff [25]	80.8	66.5	48.7	66.4	77.7	62.5	44.6	62.7	64.8	55.4	39.9	54.2
PEAN	**84.5**	**71.4**	**52.9**	**70.6**	**79.4**	**67.0**	**49.1**	**66.1**	**68.9**	**60.2**	**45.9**	**59.0**
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

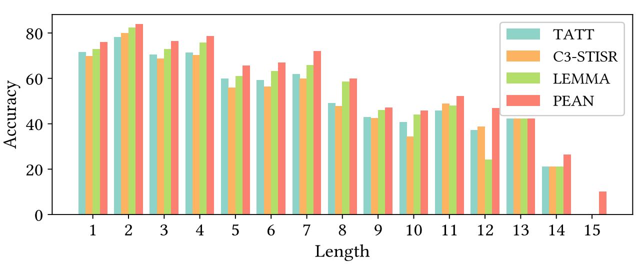


Figure 4: Statistics on the performance of different text prior-based models with publicly available weights on images containing text of different lengths.

4.3 Comparing with State-of-the-Art Methods

We evaluate our proposed PEAN on the TextZoom benchmark [53] and compare its performance with several typical STISR methods. The results are presented in Table 1. It is evident that our proposed PEAN achieves new SOTA performance with significant improvements. For instance, when considering the recognition accuracy of ASTER [44], our model shows an average improvement of **+4.2** compared with the existing SOTA model, namely, TextDiff [25]. Furthermore, we provide statistics on the performance of different text prior-based models [16, 30, 63] on images containing text of different lengths. For a fair comparison, all selected models use publicly available weights. As illustrated in Figure 4, PEAN outperforms other models across nearly every text length. It is particularly superior to previous works with images containing lengthy text, establishing itself as the first model capable of processing images containing text with up to 15 letters.

In addition, we provide some visualizations of samples from TextZoom recovered by several representative STISR models, as shown in Figure 5. With the assistance of the ETP, PEAN is able to generate SR images with improved semantic accuracy compared with previous text prior-based methods [16, 30, 63]. Although early methods such as TSRN [53] can generate SR images with correct semantic information for words like "cooking", it is obvious that the visual structure of the text in images is disastrous.

Additionally, we perform experiments on the dataset we built without retraining the models. Improved performance on PSNR and SSIM in Table 2 shows that compared with previous representative STISR methods, PEAN is better at recovering the visual structure of the text in images.

Figure 5: Visualization of SR images and their recognition results by ASTER. Red characters indicate wrong recognition results.

Table 2: The performance of representative STISR models on the dataset we built.

Methods	PSNR	SSIM
TSRN [53]	21.48	0.7743
TBSRN [3]	23.01	0.7967
TG [4]	21.84	0.7116
TATT [30]	23.31	0.8012
C3-STISR [63]	21.03	0.7602
LEMMA [16]	22.09	0.7549
PEAN	**24.24**	**0.8021**

4.4 Ablation Study

Here we conduct ablation studies to demonstrate the effectiveness of components in our model. All the experiments are conducted on TextZoom and we report the recognition accuracy of ASTER [44].

4.4.1 ***Text Prior and the Enhancement Module.*** We conduct experiments to verify the importance of incorporating the text prior and the enhancement module into our model. The results shown in Table 3 justify that the introduction of the text prior can improve the performance of the SR process. However, this primary text prior from LR images is not robust enough to guide the SR network to generate SR images with high semantic accuracy. Considering that TP-HR can serve as a powerful alternative for guidance, we conduct an exploratory experiment wherein we substitute TP-LR with TP-HR. Results shown in the last row of the table demonstrate that leveraging the powerful prior information from HR images can significantly boost the performance of text prior-based STISR methods. This observation motivates us to design a module aimed at enhancing the primary text prior from LR images, thereby providing further guidance to the SR network for generating images with high semantic accuracy. Results displayed in the table indicate that our proposed TPEM is able to produce the ETP, which significantly improves the performance over the model with TP-LR among all subsets by **+6.8** on average.

Table 3: Analysis of the impact of the text prior and the TPEM. The last row shows the results of the exploratory experiment as illustrated in Figure 1(f).

TP-LR	TPEM	TP-HR	Easy	Medium	Hard	Average
			75.7	60.2	42.1	60.4
✓			79.7	62.3	46.1	63.8
✓	✓		**84.5**	**71.4**	**52.9**	**70.6**
		✓	88.4	75.5	61.3	75.9

4.4.2 ***Impact of the AMM.*** Here we perform experiments to justify the superiority of the AMM adopted in our model. As presented in Table 4, the comparison reveals the following points: (1) Our proposed AMM handles the STISR task in a more effective way than the CNN-BiLSTM-based SRB [53]. The local inductive bias of the CNN [10, 38] and the rigid nature of BiLSTM [50, 69] limit its performance in recovering the visual structure of images with long or deformed text [3, 30], while the AMM can solve this problem. (2) The typical ViT-based architectures [9, 10, 26] show trivial performance, as a result of the tendentious design for high-level vision tasks, which neglects the inherent characteristic of STISR. (3) Although employing the strip-wise attention mechanism, Stripformer [49], which also leverages the conditional positional encodings [5] and several residual blocks for the image deblurring task, fails in the STISR task. On average, PEAN obtains an improvement of the performance of **+5.9** from the AMM. Further experiments in the Supplementary Material show that even without the MTL paradigm, the AMM still outperforms the SRB [53].

Table 4: Ablation study on the impact of the AMM.

Methods	Easy	Medium	Hard	Average
SRB [53]	80.1	64.4	46.4	64.7
ViT [10]	81.8	65.7	49.5	66.7
Swin [26]	73.8	55.1	39.0	57.1
CSWin [9]	70.2	52.9	37.2	54.5
Stripformer [49]	72.9	53.6	37.3	55.7
AMM	**84.5**	**71.4**	**52.9**	**70.6**

4.4.3 ***Effect of the MTL Paradigm***. In this part, we conduct experiments to show the effect of the MTL paradigm introduced in our work. The results shown in Table 5 indicate that: (1) The stroke-based SFM loss [4] provides an average contribution of **+4.6** for the improved performance, which convinces us that the preservation of visual structure plays a vital role in STISR. (2) The text recognition task employed in our work is also indispensable. The CTC [15] loss applied on the output of ARM ($\mathcal{L}^{a}_{ctc}$) brings correct semantic constraint for the AMM. Besides, the CTC loss imposed on the ETP ($\mathcal{L}^{t}_{ctc}$) ensures the coherence between the text prior and the ground truth text label, guiding the training of the enhancement module in a precise way. We provide more detailed experiments in the Supplementary Material.

Table 5: Ablation study on the effect of MTL.

Loss Functions	Easy	Medium	Hard	Average
$\mathcal{L}_{mse}$	76.2	58.8	41.5	59.9
+ $\mathcal{L}_{sfm}$	79.2	64.3	47.0	64.5
+ $\mathcal{L}_{mae}$	79.6	65.1	47.1	64.9
+ $\mathcal{L}^{t}_{ctc}$	81.4	68.8	50.7	67.9
+ $\mathcal{L}^{a}_{ctc}$	**84.5**	**71.4**	**52.9**	**70.6**

4.5 Representation Analysis

In this section, we delve into the reasons about the ability of the ETP that guides the SR network in generating images with improved semantic accuracy. In Table 3, we observe superior performance for PEAN with TP-HR, while PEAN with TP-LR exhibits poorer results compared to PEAN with ETP, as demonstrated in Table 3. To further investigate these observations, following the common analysis setting on Vision Transformers [35, 38], we employ the linear centered kernel alignment (CKA) [22] to assess the similarity of representations among these three models.

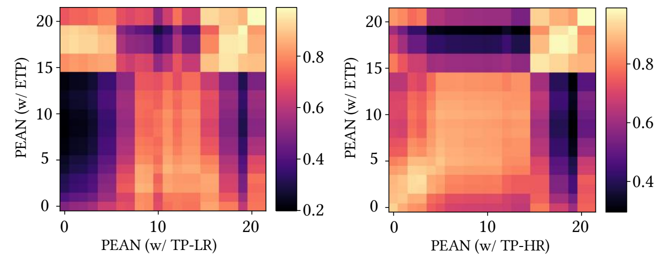


Figure 6: Results of representation comparison by CKA [22].

As depicted in Figure 2, the introduction of the text prior influences the representations of the AMM and SRM during the inference phase. Therefore, we focus on comparing the representational similarity of the layers in these two modules. We categorize the 22 layers into two groups for analysis: layers of AMM (0^{th}~11^{th} layer) and layers of SRM (12^{th}~21^{st} layer). As shown in the diagonal section of Figure 6, distinct differences in representations are evident across all layers of AMM when comparing PEAN (w/ ETP) and PEAN (w/ TP-LR). Conversely, a consistently high similarity is observed between PEAN (w/ ETP) and PEAN (w/ TP-HR). Regarding the layers of SRM, variations in representational similarity mainly exists in the shallow layers (12^{th}~14^{th} layer).

Consequently, we can conclude that the power of the ETP lies in its ability to *make the representations learned by AMM and SRM more similar to those learned by the corresponding modules in PEAN (w/ TP-HR)*, which is known for its superior performance. This study further justifies that the combination of our proposed TPEM and AMM brings out powerful capabilities.

5 Conclusion

In this paper, we propose a Prior-Enhanced Attention Network (PEAN) for scene text image super-resolution (STISR). Specifically, we design a Text Prior Enhancement Module (TPEM) to provide the ETP for the subsequent SR process, enabling SR images to contain accurate semantic information. Moreover, an Attention-based Modulation Module (AMM) is devised to obtain local and global coherence in scene text images, which can recover the visual structure of images with text in various sizes and deformations. Additionally, we introduce the Multi-Task Learning (MTL) paradigm to improve the legibility of LR images. Experiments demonstrate that our proposed PEAN achieves SOTA performance through the interaction of these designs. We believe our work will serve as a strong baseline for future works, and will push forward the research of STISR as well as other sub-fields of scene text images.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62076062 and 62306070) and the Social Development Science and Technology Project of Jiangsu Province (No. BE2022811). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

[1] Meeras Salman Al-Shemarry, Yan Li, and Shahab Abdulla. 2023. Identifying License Plates in Distorted Vehicle Images: Detecting Distorted Vehicle Licence Plates Using a Novel Preprocessing Methods With Hybrid Feature Descriptors. *IEEE Intelligent Transportation Systems Magazine* 15, 2 (2023), 6–25.

[2] Darwin Bautista and Rowel Atienza. 2022. Scene Text Recognition with Permuted Autoregressive Sequence Models. In *Proceedings of the European Conference on Computer Vision*. 178–196.

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

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

[5] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. 2023. Conditional Positional Encodings for Vision Transformers. In *Proceedings of the International Conference on Learning Representations*.

[6] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. 2017. Pixel Recursive Super Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5449–5458.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2016. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2016), 295–307.

[9] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. 2022. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12114–12124.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

[11] Chuantao Fang, Yu Zhu, Lei Liao, and Xiaofeng Ling. 2021. TSRGAN: Real-world text image super-resolution based on adversarial learning and triplet attention. *Neurocomputing* 455 (2021), 88–96.

[12] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7098–7107.

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9935–9946.

[14] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked Diffusion Transformer is a Strong Image Synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23164–23173.

[15] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*. 369–376.

[16] Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. 2023. Towards Robust Scene Text Image Super-resolution via Explicit Location Enhancement. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 782–790.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Proceedings of the Advances in Neural Information Processing Systems*. 6840–6851.

[18] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. 2023. CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 6896–6908.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 2017–2025.

[20] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on Robust Reading. In *Proceedings of IEEE International Conference on Document Analysis and Recognition*. 1156–1160.

[21] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations*.

[22] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the International Conference on Machine Learning*. 3519–3529.

[23] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5835–5843.

[24] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 105–114.

[25] Baolin Liu, Zongyuan Yang, Pengfei Wang, Junjie Zhou, Ziqi Liu, Ziyi Song, Yan Liu, and Yongping Xiong. 2023. TextDiff: Mask-Guided Residual Diffusion Models for Scene Text Image Super-Resolution. *arXiv preprint arXiv:2308.06743* (2023).

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.

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

[29] Jianqi Ma, Shi Guo, and Lei Zhang. 2023. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing* 32 (2023), 1341–1353.

[30] Jianqi Ma, Zhetong Liang, and Lei Zhang. 2022. 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*. 5911–5920.

[31] Anand Mishra, Karteek Alahari, and C. V. Jawahar. 2012. Scene Text Recognition using Higher Order Language Priors. In *Proceedings of the British Machine Vision Conference*. 1–11.

[32] Diganta Misra. 2020. Mish: A Self Regularized Non-Monotonic Activation Function. In *Proceedings of the British Machine Vision Conference*.

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning*. 16784–16804.

[34] Chihiro Noguchi, Shun Fukuda, and Masao Yamanaka. 2024. Scene Text Image Super-resolution based on Text-conditional Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1485–1495.

[35] Namuk Park and Songkuk Kim. 2022. How Do Vision Transformers Work?. In *Proceedings of the International Conference on Learning Representations*.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*. 8024–8035.

[37] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing Text with Perspective Distortion in Natural Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 569–576.

[38] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks?. In *Proceedings of the Advances in Neural Information Processing Systems*. 12116–12128.

[39] Sangeeth Reddy, Minesh Mathew, Lluís Gómez, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. 2020. RoadText-1K: Text Detection & Recognition Dataset for Driving Videos. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 11074–11080.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10674–10685.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems*. 36479–36494.

[42] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2023. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2023), 4713–4726.

[43] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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 (2016), 2298–2304.

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

[45] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1874–1883.

[46] Rui Shu, Cairong Zhao, Shuyang Feng, Liang Zhu, and Duoqian Miao. 2023. Text-Enhanced Scene Image Super-Resolution via Stroke Mask and Orthogonal Attention. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 11 (2023), 6317–6330.

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations*.

[48] LeoWu TomyEnrique, Xiangcheng Du, Kangliang Liu, Han Yuan, Zhao Zhou, and Cheng Jin. 2024. Efficient Scene Text Image Super-Resolution with Semantic Guidance. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 3160–3164.

[49] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. 2022. Stripformer: Strip Transformer for Fast Image Deblurring. In *Proceedings of the European Conference on Computer Vision*. 146–162.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*. 5998–6008.

[51] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. 2023. Omni Aggregation Networks for Lightweight Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22378–22387.

[52] Peng Wang, Cheng Da, and Cong Yao. 2022. Multi-granularity Prediction for Scene Text Recognition. In *Proceedings of the European Conference on Computer Vision*. 339–355.

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

[54] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision Workshops*. 63–79.

[55] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14174–14183.

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

[57] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. DiffIR: Efficient Diffusion Model for Image Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13095–13105.

[58] Zongyuan Yang, Baolin Liu, Yongping Xiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. 2023. DocDiff: Document Enhancement via Residual Diffusion Models. In *Proceedings of the ACM International Conference on Multimedia*. 2795–2806.

[59] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12110–12119.

[60] Wenyu Zhang, Xin Deng, Baojun Jia, Xingtong Yu, Yifan Chen, Jin Ma, Qing Ding, and Xinming Zhang. 2023. Pixel Adapter: A Graph-Based Post-Processing Approach for Scene Text Image Super-Resolution. In *Proceedings of the ACM International Conference on Multimedia*. 2168–2179.

[61] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual Dense Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2472–2481.

[62] Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. 2021. Scene text image super-resolution via parallelly contextual attention network. In *Proceedings of the ACM International Conference on Multimedia*. 2908–2917.

[63] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. 2022. C3-STISR: Scene Text Image Super-resolution with Triple Clues. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1707–1713.

[64] Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, and Matthias Grundmann. 2023. Towards Authentic Face Restoration with Iterative Diffusion Models and Beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7312–7322.

[65] Yuxuan Zhou, Liangcai Gao, Zhi Tang, and Baole Wei. 2024. Recognition-Guided Diffusion Model for Scene Text Image Super-Resolution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 2940–2944.

[66] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, and Qibin Hou Ming-Ming Cheng. 2023. SRFormer: Permuted Self-Attention for Single Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12734–12745.

[67] Yuxuan Zhou, Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Lei Zhang, Margret Keuper, and Xiansheng Hua. 2022. SP-ViT: Learning 2D Spatial Priors for Vision Transformers. In *Proceedings of the British Machine Vision Conference*.

[68] Shipeng Zhu, Pengfei Fang, Chenjie Zhu, Zuoyan Zhao, Qiang Xu, and Hui Xue. 2024. Text Image Inpainting via Global Structure-Guided Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7775–7783.

[69] Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. 2023. Improving Scene Text Image Super-Resolution via Dual Prior Modulation Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3843–3851.

[70] Xiangyuan Zhu, Kehua Guo, Hui Fang, Rui Ding, Zheng Wu, and Gerald Schaefer. 2023. Gradient-Based Graph Attention for Scene Text Image Super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3861–3869.

[71] Zhongjie Zhu, Lei Zhang, Yongqiang Bai, Yuer Wang, and Pei Li. 2024. Scene Text Image Super-Resolution Through Multi-Scale Interaction of Structural and Semantic Priors. *IEEE Transactions on Artificial Intelligence* 1, 1 (2024), 1–11.