Improving Scene Text Image Super-resolution via Dual Prior Modulation Network

Shipeng Zhu[1,2], **Zuoyan Zhao**[1,2], **Pengfei Fang**[1,2], **Hui Xue**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]MOE Key Laboratory of Computer Network and Information Integration (Southeast University), China
{shipengzhu, zuoyanzhao, fangpengfei, hxue}@seu.edu.cn

Abstract

Scene text image super-resolution (STISR) aims to simultaneously increase the resolution and legibility of the text images, and the resulting images will significantly affect the performance of downstream tasks. Although numerous progress has been made, existing approaches raise two crucial issues: (1) They neglect the global structure of the text, which bounds the semantic determinism of the scene text. (2) The priors, e.g., text prior or stroke prior, employed in existing works, are extracted from pre-trained text recognizers. That said, such priors suffer from the domain gap including low resolution and blurriness caused by poor imaging conditions, leading to incorrect guidance. Our work addresses these gaps and proposes a plug-and-play module dubbed Dual Prior Modulation Network (DPMN), which leverages dual image-level priors to bring performance gain over existing approaches. Specifically, two types of prior-guided refinement modules, each using the text mask or graphic recognition result of the low-quality SR image from the preceding layer, are designed to improve the structural clarity and semantic accuracy of the text, respectively. The following attention mechanism hence modulates two quality-enhanced images to attain a superior SR result. Extensive experiments validate that our method improves the image quality and boosts the performance of downstream tasks over five typical approaches on the benchmark. Substantial visualizations and ablation studies demonstrate the advantages of the proposed DPMN. Code is available at: https://github.com/jdfxzzy/DPMN.

Introduction

Scene text images, containing rich linguistic and graphic information, are widely present in our daily life. The understanding of scene text images is an integral part of various high-level applications, like scene text recognition (Fang et al. 2021), scene text retrieval (Wang et al. 2021a), and text-based image captioning (Zhang et al. 2022). However, limited zone in the image and inadequate imaging conditions (Long, He, and Yao 2021) cause an issue of low resolution (LR) for texts. This, in turn, leads to unreliable text understanding in such images, thereby degrading the performance of downstream tasks. In this context, it is necessary to develop a super-resolution (SR) method, to recover LR scene text images to high-quality SR ones.

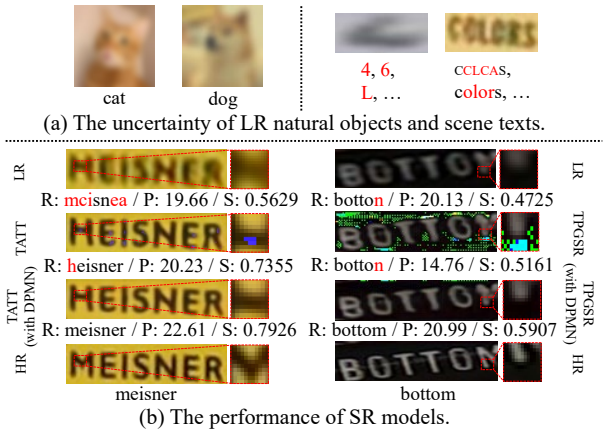

Figure 1: The illustration of effect about global structure and graphic semantic priors on the scene text image. "R", "P" and "S" represent recognition results, PSNR, and SSIM.

To achieve this goal, some early attempts (Xu et al. 2017; Pandey et al. 2018) simply utilize techniques of conventional SR methods, e.g., L_1 loss, to improve the quality of scene text in images. However, they cannot effectively boost the performance of downstream tasks. Subsequently, some methods, tailored for the scene text image super-resolution (STISR) task, benefit from the superficial properties of the scene text. For example, the pioneering work, Text Super-Resolution Network (TSRN), tends to perceive the sequential information of the text via CNN-BiLSTM layers (Wang et al. 2020). The most recent works attempt to leverage various text properties from LR images, as prior, to steer the SR process. Text Gestalt (TG) model (Chen et al. 2022) uses the local stroke structure to capture the stroke-aware prior from a Transformer-based recognizer to prompt the training phase. Text Prior Guided Super-Resolution (TPGSR) model (Ma, Guo, and Zhang 2021) and the following Text ATTention network (TATT) (Ma, Liang, and Zhang 2022) leverage the pre-trained text recognizer to obtain a text prior, i.e., the probability sequence of a scene text image whose length denotes the number of characters learned by the text recognizer. This hence improves the quality of SR images as

*Corresponding author.

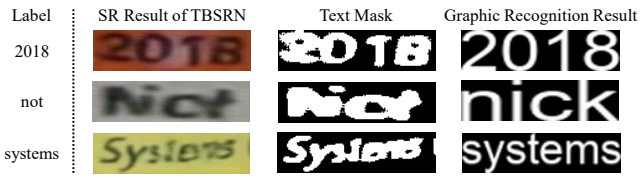

Figure 2: The illustration of global structure and graphic semantic priors in an existing STISR approach.

well as the performance of downstream tasks.

Despite the significant advances that have been made, existing methods ignore two essential facts, limiting their further improvement. **First, the global structure information of scene text plays a vital role in STISR**. In terms of the scene text, global structure contains character strokes and the orientation of the text sequence (see HR images in Figure 1(b)). Unlike natural objects, scene text in the image is the spatial cluster of discrete characters, where each item is a continuous graphic (Ye and Doermann 2014). Moreover, the semantic definiteness of the text is determined by the global structure containing characteristics at multi scales (Yao et al. 2014). That said, natural objects like animals can be recognized by local structure and texture, even omitting most of the global structure information, e.g., shape and profile. However, missing such a global structure brings uncertainty to the scene text (see Figure 1(a)) (Wu et al. 2019). As a direct representation of the global structure, the text mask has been ignored or solely used as the fourth channel of the input for enhancement by existing methods (Wang et al. 2020; Ma, Liang, and Zhang 2022). **Second, the recognizer-based priors have inherent shortcomings**. Concretely, current prior-guided methods merely employ a pre-trained text recognizer to extract the text prior, such that the domain gap of resolution will result in incorrect prior information. Meanwhile, we empirically observe that such methods are prone to generate artifacts. As shown in Figure 1(b), TATT and TPGSR produce amiss images, and characters are falsely predicted by the recognizer, which again shows the necessity of the structure prior to the scene text image.

To address the issues above, we utilize two explicit image-level priors, i.e., text mask and graphical recognition result from low-quality images. These two priors provide complementary information. Specifically, the text mask brings superiority in terms of the global structure, while the graphical recognition result contributes to clear semantic features. That is, the former compensates for the deficiency of character correctness and attributes w.r.t the graphical property, while the latter compensates for the ambiguity of the mask in terms of local essential information. This induction is shown in Figure 2. Having those two priors in mind, we propose Dual Prior Modulation Network (DPMN), a plug-and-play module that enjoys the global structural information and the local semantic information to improve the quality of the SR images produced by existing models. In doing so, two branches of Prior-Guided Refinement Modules (PGRM) are designed to create the text mask prior and the graphical recognition prior, and each processes SR images guided by the global structural information and the local semantic information, respectively. A following Complementation Modulation Module (CMM) is further proposed to modulate and fuse the reconstructed SR images, refined by two PGRM branches. Of note, our work can be understood as a post-processor of STISR networks, such that it can be seamlessly used in existing established works. Our **contributions** are summarized as follows:

- We propose a novel Dual Prior Modulation Network (DPMN), which leverages the text mask and the graphical recognition result as priors, to improve the quality of SR images. By doing so, DPMN benefits from the global structure and semantic information, attaining a superior SR result.
- In DPMN, each PGRM generates a prior and produces the improved SR image via interacting the scene text image and the prior image. Then enhanced SR images are modulated and fused by a CMM.
- Extensive experiments demonstrate that the proposed DPMN can boost image quality and the performance of the text recognition task on the TextZoom benchmark on top of existing methods. Additional analysis exhibits the generalizability of DPMN.

Related Work

Single Image Super-Resolution

Single image super-resolution (SISR) aims to recover HR images from LR ones (Wang, Chen, and Hoi 2020). The pioneering work, called SRCNN (Dong et al. 2015), employs CNNs to regress a complex non-linear mapping to reconstruct the HR images and achieves superior performance. This is the first attempt in the deep learning area. Subsequent approaches improve the quality of SR outputs by developing advanced learning strategies and neural architectures. To enrich the knowledge of the SR model, the content-based approaches develop transfer learning algorithms, delivering knowledge from pre-trained classification networks to SR networks (Ledig et al. 2017; Johnson, Alahi, and Fei-Fei 2016). As another way to learn informative features, a growing number of approaches develop the attention mechanism (Fang et al. 2022) to attend to useful regions on LR images (Zhang et al. 2018; Dai et al. 2019; Mei, Fan, and Zhou 2021). Recent studies also show the Transformer with a self-attention mechanism gains promising performance in SISR (Chen et al. 2021; Liang et al. 2021).

Scene Text Image Super-Resolution

Unlike the SISR task, scene text image super-resolution (STISR) is a more challenging task, which imposes requirements for understanding the text in images. The initial solutions utilize prior statistical knowledge to guide the SR process. In (Capel and Zisserman 2000), the maximum a posterior (MAP) method is adopted to predict new pixels in SR images. The Teager Filter (Mancas-Thillou and Mirmehdi 2005) employs the Taylor decomposition to highlight high frequencies of the text. Recently, many works exploit the properties of the scene text to improve SR networks.

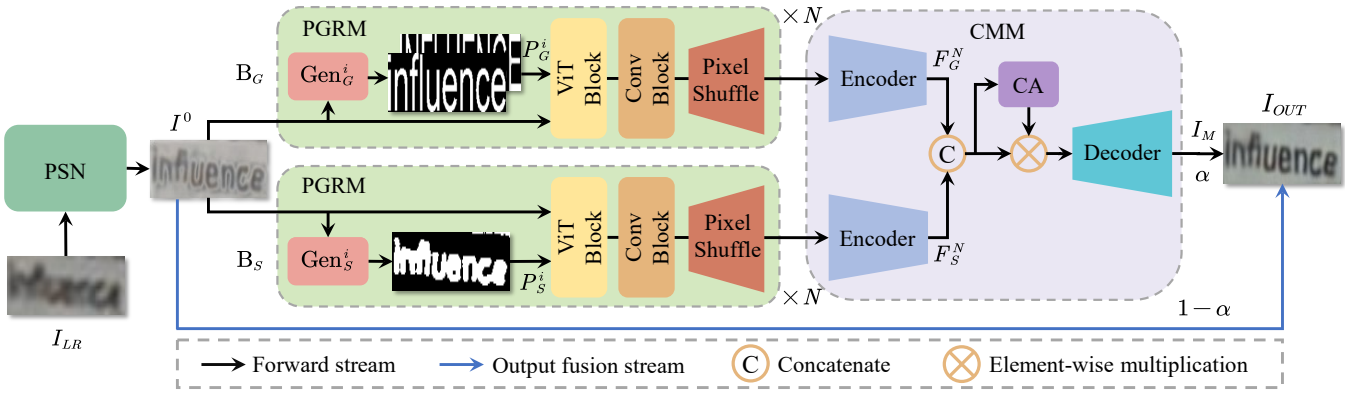

Figure 3: The overall architecture of our proposed Dual Prior Modulation Network (DPMN). It consists of two main modules: Prior-Guided Refinement Modules (PGRMs) and Complementation Modulation Module (CMM). "Gen", "Conv" and "CA" denote Prior Generator, Convolutional Block, and Channel-wise Attention Block, respectively. The input of our DPMN is provided by the Pre-trained STISR Network (PSN) with frozen parameters.

For example, TSRN (Wang et al. 2020) and PCAN (Zhao et al. 2021) apply the CNN-BiLSTM module to perceive sequential features of the scene text. The prior information is also considered as essential auxiliary information for the SR process. Specifically, TBSRN (Chen, Li, and Xue 2021) benefits from the supervision of character-level features, which are developed by a pre-trained text recognizer. The TG model (Chen et al. 2022) uses the local structure prior, i.e., text strokes, to improve the quality of SR images. TPGSR (Ma, Guo, and Zhang 2021) and the following TATT (Ma, Liang, and Zhang 2022) further show that priors with text semantic information are also beneficial for the STISR task.

Although significant progress has been made, existing methods suffer from the lack of constraints on global structure information and imprecise priors from the recognizer, which may limit further performance improvements. Inspired by this, our work proposes a dual-branch network to produce superior SR images, benefiting from complementary priors.

Methodology

This section first provides a sketch of the proposed Dual Prior Modulation Network (DPMN). Then we continue to present a detailed description of two units in DPMN, i.e., the Prior-Guided Refinement Module (PGRM) and the Complementation Modulation Module (CMM). We will also introduce the training objective of the proposed network.

Overall Architecture

The overall architecture of the proposed DPMN is illustrated in Figure 3. Our proposed DPMN is built on top of existing STISR networks. For any pre-trained STISR network, denoted by PSN in Figure 3, it first receives LR images $I_{LR} \in \mathbb{R}^{h\times w\times 3}$ as input, and produces primary SR images $I^0 = \mathrm{PSN}(I_{LR}) \in \mathbb{R}^{2h\times 2w\times 3}$. Then the DPMN further refines I^0 using two branches of networks, with each consisting of N Prior-Guided Refinement Modules (PGRMs). Notably, the two branches, called B_G and B_S, refine I^0 guided by the graphic semantic prior P_G and the global structure prior P_S, respectively. For the i-th PGRM in branch B_G, a generator Gen_G^i first generates a graphic semantic prior P_G^i, given the refined images I_G^{i-1} from the previous PGRM as input. Then both I_G^{i-1} and P_G^i are sent to the following refinement modules to produce a newly refined image I_G^i. In the same vein, the PGRM in branch B_S has a similar workflow using the global structure prior P_S. Having the refined images I_G^N and I_S^N in hand, the Complementation Modulation Module (CMM) aggregates the two images and produces a modulated one I_M. Of note, the parameters of the PSN are frozen while only the DPMN is optimized in the training phase. In the inference phase, we employ a fusion strategy to ensure the robustness of the final output, i.e., $I_{OUT} = \alpha \times I_M + (1-\alpha) \times I^0$ for $0 < \alpha < 1$.

Prior-Guided Refinement Module

Two critical issues arise for the refinement process guided by priors in the STISR task: (1) How to mitigate the influence of imprecise prior information due to defective input (see Figure 2). (2) How to integrate useful prior information, as guidance, into images. Our Prior-Guided Refinement Module (PGRM) is proposed to address those two issues. This is achieved by calibrating the low-quality SR input by information interaction of the input and the prior. Each PGRM (see Figure 3) contains a prior generator, a Vision Transformer (ViT) block, a convolutional block, and a pixel shuffle layer.

As shown in Figure 2, we can empirically find that the global structure of the scene text and graphic semantics can provide complementary guidance for the restoration of the SR image. This motivates us to exploit two representative image-level information as priors, i.e., the text mask and the graphic recognition result.

To utilize these two priors, we propose to extract and use the priors in a parallel manner, such that each branch of PGRM processes a specific prior to the fullest. Different from the feature map or the embedding vector (Ma, Guo, and

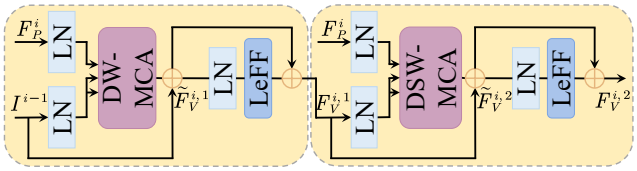

Figure 4: The architecture of the Vision Transformer (ViT) block in i-th PGRM of each branch. "DW-MCA" and "DSW-MCA" are short for Dynamic Window Multi-head Cross Attention and Dynamic Shifted Window Multi-head Cross Attention. "LN" and "LeFF" denote Layer Normalization and Locally-enhanced Feed-Forward Network.

Zhang 2021; Chen, Li, and Xue 2021), the prior in graphic-recognition-guided branch B_G (see Figure 3) can present the semantic information of text images. This can avoid bringing the noise, due to the domain gap, from the ambiguous semantics to the input images. For the i-th PGRM in B_G, a pre-trained recognizer first produces a text result for the input I_G^{i-1}, and the following rendering module (Gupta, Vedaldi, and Zisserman 2016) further transfers the text result to an image-format data containing upper and lower case letters, as the graphic prior, denoted by P_G^i. This processing is formulated as:

$$P_G^i = \mathrm{Ren}(\mathrm{Rec}(I_G^{i-1})) \in \mathbb{R}^{2h\times 2w\times 2}. \quad (1)$$

In the i-th PGRM unit of the global-structure-guided branch B_S, we utilize the text mask generated by the binarization operation of I_S^{i-1} as the structure prior P_S^i. The structure information, e.g., font, size, tendency, et al, can be represented by P_S^i. Its operation is given by:

$$P_S^i = \mathrm{Bin}(I_S^i) \in \mathbb{R}^{2h\times 2w\times 1}. \quad (2)$$

Existing STISR approaches simply stack the Sequential Residual Block (SRB) in the backbone network, to capture and mine the sequence dependence of the text. However, the SRB has difficulty to understand the deformable text due to the rigid nature of the Bi-LSTM. Meanwhile, most prior-based approaches manually fuse the prior information into the input image via simple operations, e.g., addition. In our work, we want the network to learn the fusion protocol adaptively, thereby improving the quality of the SR images.

We address this by interacting the information flow between the prior and the input image via the ViT. Before formulating the information interaction, we first perform the dimensionality matching operation. That is, since the dimensions of the input image are located in the space of $\mathbb{R}^{2h\times 2w\times 3}$, we need to project the two priors (e.g., P_G^i and P_S^i) to the same space. For $P_G^i \in \mathbb{R}^{2h\times 2w\times 2}$, we use a convolutional layer to do the transformation, obtaining input feature $F_{P_G}^i = \mathrm{Conv}(P_G^i) \in \mathbb{R}^{2h\times 2w\times 3}$. For the global structure prior P_S^i, we expand it in the channel wise, as $F_{P_S}^i \in \mathbb{R}^{2h\times 2w\times 3}$.

As shown in Figure 4, the ViT block in the i-th PGRM of arbitrary branch is realized by the cross attention mechanism (Zhou et al. 2020), which takes the prior F_P^i (collective term for $F_{P_G}^i$ and $F_{P_S}^i$) as the query, and inputs low-quality SR image I^{i-1} (collective term for I_G^{i-1} and I_S^{i-1}) as the key and value, to perform the information interaction. Following the recent DW-ViT (Ren et al. 2022), we also process the ViT block in two stages. In the first stage, a multi-head cross attention layer (denoted by DW-MCA) with residual connection receives the F_P^i and I^{i-1} as input, to produce multi-scale correlation by three sizes of windows, given by:

$$\tilde{F}_V^{i,1} = \text{DW-MCA}(\mathrm{LN}(F_P^i), \mathrm{LN}(I^{i-1})) + \mathrm{LN}(I^{i-1}). \quad (3)$$

Along with interacting the prior and the image with dynamic windows mechanism, we also enable the DW-ViT the capacity to capture local contextual information for $F_V^{i,1}$, realized by:

$$F_V^{i,1} = \tilde{F}_V^{i,1} + \mathrm{LeFF}\left(\mathrm{LN}\left(\tilde{F}_V^{i,1}\right)\right), \quad (4)$$

where LeFF is the locally-enhanced feed-forward network with spatial-wise and depth-wise convolutional layer proposed in UFormer (Wang et al. 2022).

In the second stage, we replace the DW-MCA layer with the dynamic shifted window multi-head cross attention (DSW-MCA) layer to build long-range interaction. This stage can be formulated as:

$$\tilde{F}_V^{i,2} = \text{DSW-MCA}(\mathrm{LN}(F_P^i), \mathrm{LN}(F_V^{i,1})) + \mathrm{LN}(F_V^{i,1}), \quad (5)$$

$$F_V^{i,2} = \tilde{F}_V^{i,2} + \mathrm{LeFF}\left(\mathrm{LN}\left(\tilde{F}_V^{i,2}\right)\right). \quad (6)$$

Recent researches have proved that the ViT has its internal drawbacks, e.g., the limitation of encoding the inductive bias, for vision tasks (Li et al. 2022). Thereby, a convolutional block (denoted by Conv), which consists of two convolutional layers, is added after the ViT block. This design can improve the perception ability of locality and spatial invariance (Peng et al. 2021) in PGRM. Subsequently, the widely-used pixel shuffle layer (PS) (Shi et al. 2016) produces the refined image, as:

$$I^i = \mathrm{PS}(\mathrm{Conv}(F_V^{i,2})). \quad (7)$$

To optimize the proposed PGRM block efficiently, we adopt two image-level losses. The pixel loss $\mathcal{L}_{pix}$ constrains the information of image content, and the gradient profile loss $\mathcal{L}_{gp}$ considers to minimize the information loss of edge details (Wang et al. 2020). For the i-th PGRM module, the loss function is given by:

$$\mathcal{L}_{img}^i = \lambda_p \underbrace{||I_{HR} - I^i||_2}_{\mathcal{L}_{pix}} + \lambda_g \underbrace{||\nabla I_{HR} - \nabla I^i||_1}_{\mathcal{L}_{gp}}. \quad (8)$$

Since each branch includes N PGRM blocks, the total loss of each branch is:

$$\mathcal{L}_B = \sum_{i=1}^{N} \mathcal{L}_{img}^i. \quad (9)$$

Complementation Modulation Module

After the enhancement to the quality of SR images in two branches, we have two refined SR images, i.e., I_G^N and I_S^N. A neural module is required to fuse and modulate those

two images to attain a superior SR one. As shown in Figure 3, this is realized by an encoder-decoder architecture. Each encoder is composed of six convolutional layers to extract the key features, as $F_G^N = \text{Encoder}(I_G^N)$ and $F_S^N = \text{Encoder}(I_S^N)$. Then we concatenate two feature maps, as:

$$F_M^N = \text{Concat}(F_G^N, F_S^N). \tag{10}$$

We then employ the channel attention mechanism (Hu, Shen, and Sun 2018) to learn the modulation importance per slice in F_M^N and weight F_M^N in channel-wise. A symmetric decoder is further used to produce the final modulated SR images. This is summarized as:

$$I_M = \text{Decoder}(\text{CA}(F_M^N) \otimes F_M^N + F_M^N). \tag{11}$$

In CMM, we use $\mathcal{L}_{img}$ to restrain I_M by HR images I_{HR}, which can be described as follows:

$$\mathcal{L}_{CMM} = ||I_{HR} - I_M||_2 + ||\nabla I_{HR} - \nabla I_M||_1. \tag{12}$$

Training Objective

In the training phase, we optimize the parameters of the proposed DPMN. The objective loss includes two types of components, i.e., the branch loss $\mathcal{L}_B$ and the CMM loss $\mathcal{L}_{CMM}$. The total loss function is:

$$\mathcal{L}_{Total} = \lambda_C \mathcal{L}_{CMM} + \lambda_G \mathcal{L}_{B_G} + \lambda_S \mathcal{L}_{B_S}. \tag{13}$$

Experiments

In this section, we first introduce the experiment datasets, evaluation metrics, and implementation details. Then we conduct comparison experiments and ablation studies to demonstrate the superiority of our method.

Evaluation Datasets and Metrics

The STISR benchmark **TextZoom** (Wang et al. 2020) is collected in real-world scenarios. It consists of 17,367 LR-HR image pairs for training and 4,373 pairs for testing. Wherein, the test set is divided into three subsets to indicate different levels of blurriness, i.e., easy (1,619 pairs), medium (1,411 pairs) and hard (1,343 pairs). The size of LR images is 16×64, while the size of HR images is 32×128.

We use the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) metrics to evaluate the quality of the SR images. In order to measure the downstream task performance, we calculate the recognition accuracy for the text recognition task. Following the common practice in (Wang et al. 2020), the recognition results are evaluated on ASTER (Shi et al. 2018), CRNN (Shi, Bai, and Yao 2016), MORAN (Luo, Jin, and Sun 2019) models.

Baselines and Implementation Details

We evaluate the proposed DPMN on five STISR models as baselines, including TSRN (Wang et al. 2020), TBSRN (Chen, Li, and Xue 2021), TPGSR (Ma, Guo, and Zhang 2021), TG (Chen et al. 2022), and TATT (Ma, Liang, and Zhang 2022). In our experiments, we directly use models from the official implementation or perform the identical hyper-parameters as reported in the official implementations to train the baseline models. Of note, to understand the net improvement from the proposed DPMN, we respectively select three models with the best performance on ASTER, CRNN, and MORAN, as baselines of each aforementioned STISR method. Then we fix the parameters of the SR models and only train the DPMNs.

We implement our model with PyTorch 1.10 deep learning library (Paszke et al. 2019) and all the experiments are conducted on one RTX 3090 GPU. For each experiment, the DPMN is trained 20 epochs using Adam optimizer (Kingma and Ba 2014). The learning rate is set to 0.001, and the size of the mini-batch is 48. We empirically observe that the loss function is insensitive to the parameter λ, and we set all the λ to 1. In the inference phase, the output fusion ratio α is selected based on the pre-trained baselines. The size of original SR results and modulated images is 32×128. We apply the pre-trained VisionLANs (Wang et al. 2021b) in PGRMs as the text prior generators. In terms of the network architecture, the number of PGRMs in each branch, N, is set to 3. In the ViT block, the window numbers of the DW-MCA and the DSW-MCA are 2, 4, and 8 with patch size 2, while the head number of the MCA is set to 6. Additionally, we exploit the adaptive dense-connection (Xie et al. 2019) and the self-distillation mechanism (Zhang et al. 2019) to ensure the stability of training and speed up the convergence rate.

Experimental Results

We conduct quantitative experiments on the benchmark TextZoom. The comparison results are shown in Table 1 and Table 2. We can clearly observe that our method boosts the image quality and the recognition accuracy of existing baselines even if the data is derived from the best models obtained on different text recognizers. Taking the latest TATT as an example, our method achieves better recognition accuracy and image quality simultaneously (improves average Accuracy/PSNR/SSIM by 0.32/0.28/0.01). Compelled by the observations above, our proposed DPMN has the universal enhancement capability for existing STISR models given the all-around improvement across all metrics.

Meanwhile, we provide some qualitative studies by comparing the visualization of the SR images between the baselines and our work. The results and comparisons shown in Figure 5 reveal that: (1) Existing methods have defects w.r.t. color reproduction and structure retention. In contrast, our method can produce precise details, verifying the potential of the complementary prior information. (2) The baseline models are prone to generate artifacts, which degrade the image quality and the performance of the downstream recognition task, while DPMN can avoid the artifacts due to explicit guidance from the text structure information.

Ablation Study

In this section, we conduct ablation studies to investigate the effectiveness of motivation and model components. All the evaluations are validated on TextZoom. The default baseline is TATT based on ASTER.

Effect of the Two Priors We first perform experiments to demonstrate the necessity of complementary priors in the SR process. This study has two settings, i.e., single-branch

Method	ASTER				CRNN				MORAN			
	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
TSRN	73.32	56.20	39.17	57.31	54.73	41.25	32.24	43.47	67.88	49.96	37.08	52.64
+DPMN	**74.43**	**56.41**	**39.24**	**57.81**	**54.91**	**41.46**	**32.46**	**43.68**	**68.07**	**50.18**	**37.16**	**52.80**
TBSRN	76.71	59.53	43.71	61.03	59.79	45.07	34.18	47.18	70.17	55.63	40.80	56.46
+DPMN	**76.78**	**60.52**	**44.08**	**61.49**	59.79	**45.22**	**34.40**	**47.29**	**70.54**	55.63	**40.95**	**56.64**
TG	77.02	62.58	42.37	61.72	59.42	48.62	34.33	48.23	72.58	57.97	39.76	57.79
+DPMN	**77.39**	**62.72**	**42.96**	**62.08**	**59.48**	**48.76**	**34.40**	**48.32**	**72.88**	**58.19**	**40.21**	**58.11**
TPGSR	78.01	60.67	42.67	61.56	58.62	45.92	33.88	46.92	72.33	55.71	39.91	57.01
+DPMN	**78.13**	60.67	**42.74**	**61.63**	**59.36**	45.92	**34.03**	**47.24**	**72.95**	**56.13**	**40.06**	**57.42**
TATT	78.51	63.29	44.97	63.30	64.30	54.15	39.09	53.28	72.88	61.02	43.78	60.12
+DPMN	**79.25**	**64.07**	**45.20**	**63.89**	**64.36**	54.15	**39.24**	**53.35**	**73.26**	**61.45**	**43.86**	**60.42**
HR	93.39	86.96	75.65	85.87	76.41	75.05	64.56	72.33	89.01	83.13	71.11	81.62

Table 1: The recognition accuracy (%) on TextZoom. The bold numbers denote the better score between the baseline and improved method by DPMN.

Method	PSNR	SSIM	Accuracy
TSRN	20.81	0.7594	51.14
+DPMN	**21.09**	**0.7698**	**51.43**
TBSRN	20.91	0.7625	54.89
+DPMN	**21.11**	**0.7650**	**55.14**
TG	18.80	0.6597	55.91
+DPMN	**20.56**	**0.7472**	**56.17**
TPGSR	21.18	0.7615	55.16
+DPMN	**21.33**	**0.7718**	**55.43**
TATT	21.21	0.7825	58.90
+DPMN	**21.49**	**0.7925**	**59.22**

Table 2: The average image quality scores and average recognition accuracy (%) on TextZoom.

Setting	Prior	PSNR	SSIM	Accuracy
SingleB	Mask	21.12	0.7864	63.50
	GRR	19.59	0.7472	63.62
	Mask+GRR	21.01	0.7859	63.67
DualB	Mask&GRR	**21.49**	**0.7925**	**63.89**
	Mask&GRR*	23.60	0.8875	77.57

Table 3: The performance of different priors. "SingleB" denotes Single-Branch, "DualB" denotes Dual-Branch, "GRR" denotes Graphic Recognition Results, "+" denotes concatenate, and "*" denotes the dual priors are from the HR image.

and dual-branch. The results in Table 3 convince us that: (1) The mask prior performs better in image quality, while the graphic recognition result, including rich semantic features, shows superior performance in the text recognition accuracy. This justifies our motivation for modulating the two complementary priors. (2) The dual-branch setting outperforms the single-branch one, clearly showing our design is reasonable and superior. Specifically, we obtain the two priors from HR images, which are impossible to be acquired in the natural SR process, and evaluate the upper bound of the image quality enhancement. The stunning results again indicate that the two priors play a promising role in modulation.

	PSN	DPMN	Easy	Medium	Hard	Average
(i)	Frozen	None	78.51	63.29	44.97	63.30
(ii)	None	Train	66.09	45.92	32.99	49.42
(iii)	Fine-tune	Train	79.18	63.08	44.38	63.30
(iv)	Frozen	Train	**79.25**	**64.07**	**45.20**	**63.89**

Table 4: The recognition accuracy (%) of different training strategies. "PSN" represents the Pre-trained STSIR Network. "None" denotes not using this module. "Fine-tune" denotes fine-turning based on a pre-trained model. "Frozen" denotes freezing the parameters of the pre-trained model.

N	Easy	Medium	Hard	Average
1	78.81	**64.07**	45.12	63.71
2	78.69	63.78	45.12	63.57
4	78.57	64.00	45.12	63.60
5	78.63	63.86	**45.20**	63.60
3	**79.25**	**64.07**	**45.20**	**63.89**

Table 5: The recognition accuracy (%) of different PGRM numbers in each branch.

Effect of the Training Strategy In this study, we investigate the training strategy of the proposed DPMN. Table 4 shows the results of possible training strategies. We can find that training the DPMN, on top of the pre-trained PSN with fixed parameters, achieves the best performance, showing the flexibility and effectiveness of DPMN. Notably, applying DPMN independently (see (ii)) as an image SR backbone leads to trivial results. One conjecture may lie in that the ambiguous priors from LR images struggle to provide proper guidance for the SR process, thereby leading to problematic initialization and weak complementary information.

Number of PGRMs We implement the proposed DPMN in a symmetric fashion, e.g., each branch containing three PGRMs. In this study, we evaluate the effect of the number of PGRMs. The results are reported in Table 5. We can empirically find that when $N = 3$, the proposed DPMN achieves the overall best performance over three settings, as well as the peak value of the average accuracy. We use $N = 3$ as the default value for the experiments in this paper.

Figure 5: The SR images with recognition results on TextZoom based on ASTER. The red characters mean wrong results.

Variant		Easy	Medium	Hard	Average
FW	2	78.69	63.78	**45.20**	63.59
	4	78.81	63.57	45.12	63.55
	8	79.06	63.64	45.05	63.64
DW (DPMN)		**79.25**	**64.07**	**45.20**	**63.89**

Table 6: The recognition accuracy (%) of different multi-head cross attention layers. "FW" denotes Fixed Window, whose size is set to the subsequent number. "DW" denotes Dynamic Window.

CMM	Easy	Medium	Hard	Average
Encoder-decoder	78.88	63.43	**45.20**	63.55
U-Net	78.94	63.29	45.05	63.48
TSRN	78.75	63.08	45.05	63.34
DPMN	**79.25**	**64.07**	**45.20**	**63.89**

Table 7: The recognition accuracy (%) comparison between different variants of CMM.

Effect of the Dynamic Window Mechanism In this part, we study the dynamic window mechanism in the proposed DPMN. The results in Table 6 show that: (1) The large window helps achieve better performance in the easy subset, while the MCA with a small window performs admirably in the hard subset. (2) Our method exhibits remarkable results in all subsets, which demonstrates the effectiveness of the dynamic window mechanism. These observations verify that the dynamic window can understand the images with different blur levels, thereby enriching its learning capacity.

Design of the CMM We utilize the encoder-decoder architecture with a channel-wise attention (CA) module to fuse the two refined SR images. In this study, we empirically compare it with three methods, i.e., single encoder-decoder without CA mechanism, U-Net (Ronneberger, Fischer, and Brox 2015) with skip connection, and TSRN (Wang et al. 2020) from previous STISR. Table 7 also shows that the architecture of CMM in our work outperforms other vanilla variants, again showing the superiority of our design.

Conclusion

In this paper, we propose a Dual Prior Modulation Network (DPMN) to boost the performance of existing Scene Text Image Super-Resolution (STISR) methods. We leverage the global text structure and graphical semantics as complementary priors to guide the SR image refinement progressively in dual branches. This is realized by the Prior-Guided Refinement Module (PGRM) and the Complementation Modulation Module (CMM). Substantial experiments and ablation studies demonstrate the effectiveness of DPMN, which improves both image quantity and the performance of the downstream recognition task. We believe our work will provide valuable intuition for further improvement of the STISR task. Future work will focus on developing a more efficient and effective backbone network for the STISR task.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62076062) and Collaborative Innovation Center of Wireless Communications Technology.

References

Capel, D.; and Zisserman, A. 2000. Super-resolution enhancement of text image sequences. In *Proceedings of the International Conference on Pattern Recognition*, 600–605.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, J.; Li, B.; and Xue, X. 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.

Chen, J.; Yu, H.; Ma, J.; Li, B.; and Xue, X. 2022. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 285–293.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Fang, P.; Zhou, J.; Roy, S. K.; Ji, P.; Petersson, L.; and Harandi, M. T. 2022. Attention in attention networks for person retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4626–4641.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 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.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2315–2324.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, 694–711. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 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*, 4681–4690.

Li, Y.; Yao, T.; Pan, Y.; and Mei, T. 2022. Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Long, S.; He, X.; and Yao, C. 2021. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1): 161–184.

Luo, C.; Jin, L.; and Sun, Z. 2019. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90: 109–118.

Ma, J.; Guo, S.; and Zhang, L. 2021. Text prior guided scene text image super-resolution. *arXiv preprint arXiv:2106.15368*.

Ma, J.; Liang, Z.; and Zhang, L. 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.

Mancas-Thillou, C.; and Mirmehdi, M. 2005. Super-resolution text using the teager filter. In *Proceedings of the International Workshop on Camera-Based Document Analysis and Recognition*, 10–16. Citeseer.

Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3517–3526.

Pandey, R. K.; Vignesh, K.; Ramakrishnan, A.; et al. 2018. Binary document image super resolution for improved readability and OCR performance. *arXiv preprint arXiv:1812.02475*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; and Ye, Q. 2021. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 367–376.

Ren, P.; Li, C.; Wang, G.; Xiao, Y.; Du, Q.; Liang, X.; and Chang, X. 2022. Beyond Fixation: Dynamic Window Visual Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11987–11997.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Shi, B.; Bai, X.; and Yao, C. 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): 2298–2304.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2035–2048.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 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.

Wang, H.; Bai, X.; Yang, M.; Zhu, S.; Wang, J.; and Liu, W. 2021a. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4558–4567.

Wang, W.; Xie, E.; Liu, X.; Wang, W.; Liang, D.; Shen, C.; and Bai, X. 2020. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision*, 650–666.

Wang, Y.; Xie, H.; Fang, S.; Wang, J.; Zhu, S.; and Zhang, Y. 2021b. 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*, 14194–14203.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3365–3387.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Wu, L.; Zhang, C.; Liu, J.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2019. Editing text in the wild. In *Proceedings of the ACM International Conference on Multimedia*, 1500–1508.

Xie, T.; Yang, X.; Jia, Y.; Zhu, C.; and Xiaochuan, L. 2019. Adaptive densely connected single image super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop*, 3432–3440. IEEE.

Xu, X.; Sun, D.; Pan, J.; Zhang, Y.; Pfister, H.; and Yang, M.-H. 2017. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 251–260.

Yao, C.; Bai, X.; Shi, B.; and Liu, W. 2014. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4042–4049.

Ye, Q.; and Doermann, D. 2014. Text detection and recognition in imagery: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7): 1480–1500.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3713–3722.

Zhang, W.; Shi, H.; Guo, J.; Zhang, S.; Cai, Q.; Li, J.; Luo, S.; and Zhuang, Y. 2022. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3335–3343.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, 286–301.

Zhao, C.; Feng, S.; Zhao, B. N.; Ding, Z.; Wu, J.; Shen, F.; and Shen, H. T. 2021. Scene text image super-resolution via parallelly contextual attention network. In *Proceedings of the ACM International Conference on Multimedia*, 2908–2917.

Zhou, J.; Roy, S. K.; Fang, P.; Harandi, M.; and Petersson, L. 2020. Cross-correlated attention networks for person re-identification. *Image and Vision Computing*, 100: 103931.