Multimodal Semantic Interaction for Text Image Super-Resolution

Xin Jiang[1]

North China Electric Power University, School of Control and Computer Engineering, Beijing, 102206, China

Abstract: ***To address the issues in existing methods where text image feature representation lacks scale adaptability and suffers from insufficient resolution, which makes it difficult for the recognizer to extract the correct textual information for guiding the reconstruction network, we propose a multimodal semantic interaction-based text image super-resolution reconstruction method. By using the attention mask in the semantic reasoning module, we correct the textual content information, obtain semantic prior knowledge, and constrain and guide the network to reconstruct semantically accurate text super-resolution images. To enhance the network's representation capability and adapt to text images of different shapes and lengths, we design a multimodal semantic interaction block. Its basic components include a visual dual-stream integration block, a cross-modal adaptive fusion block, and an orthogonal bidirectional gated recurrent unit. Experimental results show that, on the Textzoom test set, our proposed method outperforms other mainstream methods in terms of PSNR and SSIM quantitative metrics, with average recognition accuracy improvements of 2.9%, 3.6%, and 3.7% on three recognizers (ASTER, MORAN, and CRNN) compared to the TPGSR model. These results demonstrate that the text image super-resolution reconstruction method based on multimodal semantic interaction can effectively improve text recognition accuracy.***

Keywords: super-resolution reconstruction; text image; feature semantic prior; multi-modal

1. Introduction

Scene Text Recognition (STR) has wide applications in fields such as autonomous driving, mobile payments, education, and services for visually impaired individuals. With the development of deep learning, significant progress has been made in STR research. Currently, most STR algorithms0[2] rely on high-resolution text images with clear character shapes. However, due to factors such as lighting conditions, zooming, long-distance transmission, and capturing devices, the real-world scene images collected are often low-resolution (Low Resolution, LR) images with blurry characters and missing details, which severely affect the text recognition performance. Super Resolution (SR) reconstruction techniques can effectively address these issues.

With the development of convolutional neural networks and attention mechanisms, image super-resolution reconstruction techniques have achieved remarkable progress, further driving the application of super-resolution reconstruction. SRCNN[3] apply convolutional neural networks to super-resolution reconstruction. HAN[4] enhances reconstruction performance by incorporating attention modules at various layers. Compared to traditional methods, deep learning-based approaches have significantly improved reconstruction results. However, these reconstruction methods are often general models for natural scenes, lack the capability to handle specific scene text content. To address this issue, Scene Text Image Super Resolution (STSR) has emerged, aiming to enhance the resolution of low-resolution (LR) text images, improve their visual effects, and reconstruct semantically accurate text structure and shapes, thereby improving downstream scene text recognition accuracy [5].

TextSR[6] utilizes text-aware loss to guide network training, enabling the model to focus on the text information in the image. CGAN[7] combines dense residual connections with a channel attention mechanism in a generative adversarial network to learn more effective text feature representations. PlugNet[8] introduces a lightweight, pluggable super-resolution unit to handle blurred scene text images, reducing the network model's complexity and the number of parameters. Although these methods achieve excellent performance, most are trained using high-low resolution image pairs generated by bicubic downsampling. The domain gap between artificially blurred text images and real low-resolution (LR) text images makes it difficult to generalize to complex real-world scenarios.

To address this issue, TSRN[9] constructed the first real-world high-low resolution text image pair dataset, named TextZoom. In recent years, research has shown that utilizing prior information helps restore object shapes and textures. Consequently, an increasing number of studies have incorporated various text attributes as priors to guide text image reconstruction networks. TPGSR[10] uses text category information as a prior and embeds it into the reconstruction network to guide model training. TATT[11] uses Transformer to align deformed text images with text priors, further improving model performance. DPGSR[12] utilizes a designed degradation prior extractor to capture text prior information from LR images, guiding the SR module to generate recognizable SR images. TextSRNet[13] applies the Otsu method for thresholding text images, extracting text image contour prior information through a convolutional network to capture fine character details.

Text prior information has further improved the super-resolution reconstruction of text images. However, most models have not fully considered the semantic information brought by the text content in the image. Instead, they often rely on simple linear operations such as element-wise addition or concatenation to fuse with visual features, lacking an adaptive alignment mechanism between modalities, which limits the guiding role of text prior information. Additionally, many studies focus too much on text prior information and neglect the extraction of text visual features, with most using Sequential Residual Blocks (SRB)[9] as feature extraction modules. These modules typically only utilize two simple CNN layers for feature extraction. Due to the inherent limitations of convolutional computations, they struggle to capture long-range dependencies and subtle spatial variations across different granularities of text images, which especially hampers text feature representation, as they lack multi-granularity visual feature representations.

To address these issues, we propose a method for text image super-resolution reconstruction based on multimodal semantic interaction (MSISR), which jointly incorporates text semantic and visual high-level semantic information. The method corrects the text content information using the Semantic Reasoning Module (SRM) to obtain semantically accurate prior information, which then guides the reconstruction network. We introduce the Visual Dual-Flow Integration Module (VDFI), which learns the dependencies at different distances by focusing on feature maps at different layers, capturing multi-granularity high-level visual semantic features at both inter-character and intra-character granularities. Furthermore, we propose the Cross-modal Adaptive Fusion Module (CAFM), it can be seen that deeply mines the correlation between visual features and semantic priors, narrowing the feature gap between modalities. Purpose: This study aims to design and evaluate a multimodal semantic interaction network that simultaneously enhances text-image resolution and recognition accuracy. The work addresses a persistent gap between generic SR models and OCR-specific needs, thereby offering practitioners a deployable pathway for real-time, language-aware enhancement.

2. Methods

2.1 Overall Architecture

Inspired by TPGSR, we propose MSISR, whose overall framework is shown in Figure 1. It mainly consists of four parts: semantic prior generation, shallow feature extraction, deep feature extraction, and image reconstruction. Compared to ordinary natural images, text images contain important information brought by the text content. To address this characteristic of scene text images, the MSISR network utilizes a text recognizer and semantic reasoning module to extract semantic information from the text, which is then used as a prior to guide the reconstruction network in building deep features. This approach not only improves the visual quality of the reconstructed images but also further enhances the accuracy of subsequent text image recognition.

2.2 Semantic Prior Generation

The semantic prior generation mainly consists of two parts: the text recognizer and the semantic reasoning module.

2.2.1 Text Recognizer

The text recognizer utilizes a pre-trained Convolutional Recurrent Neural Network (CRNN)[14]. Compared to attention-based text recognizers, the CRNN model is simpler and considers background regions when predicting characters, which helps the model understand the boundaries between characters. The specific structure of the CRNN is shown in Figure 1, where a CNN convolutional structure is used to extract features from the input image, resulting in a feature map. Then, a bidirectional RNN is employed to predict the feature sequence and output the predicted labels. The network uses a Connectionist Temporal Classification (CTC) loss function in the transcription layer, converting the obtained label distribution into the final label sequence. For the input low-resolution text image, the text recognition probability sequence is obtained using the CRNN:

$$I_{\mathrm{RK}} = \mathrm{CRNN}(I_{\mathrm{LR}}), \tag{1}$$

where CRNN is the text recognizer CRNN operator.

2.2.2 Semantic Reasoning Module

We use a pre-trained Bidirectional Cloze Network (BCN)[15] as the semantic reasoning module to model the character correlations in the text sequence, predict contextual information, and then correct the text recognition probability sequence I_{RK}. The specific structure is shown in Figure 1. BCN consists of a series of multi-head attention layers and feedforward networks, and in the multi-head attention, it incorporates an attention mask to prevent excessive attention to the current character. For the text recognition probability sequence I_{RK}, the semantic information obtained after semantic reasoning through the SRM is:

$$I_{\mathrm{SP}} = SRM(I_{\mathrm{RK}}), \tag{2}$$

Where SRM is semantic reasoning module.

2.3 Shallow Feature Extraction

As shown in Figure 1, the shallow feature extraction module of MSISR consists of an alignment module and a 9x9 convolutional layer. The dataset we use consists of real-world LR-HR image pairs obtained by changing the focal length of a camera. There inevitably exists a misalignment between the pixels of the LR and HR images. Therefore, we use Thin Plate Spline (TPS) transformation based on a Spatial Transformer Network (STN) as the alignment module. The TPS transformation achieves non-rigid deformation by solving a thin plate spline interpolation function, converting the corresponding character regions in the LR and HR images into uniform size and shape areas. This prevents the network from learning incorrect correspondence information and alleviates pixel misalignment issues, such as horizontal, vertical, and skewed misalignments, between the LR and HR images:

$$F_0 = STN(I_{\mathrm{LR}}) \bullet W_s^{9\times 9} + b_s, \tag{3}$$

where F_0 represent the superficial features of images, I_{LR} represent LR image.

2.4 Shallow Feature Extraction

The deep feature extraction module is a residual group constructed by several Multimodal Semantic Interaction Blocks (MSIB), which facilitates effective feature transmission while preventing instability during network training. As an important component of feature extraction, MSIB mainly consists of the Visual Dual Flow Integration Block, the Cross-modal Adaptive Fusion Block, and the orthogonal Bidirectional Gated Recurrent Unit (BiGRU).

2.4.1 Visual Dual Flow Integration Block

Convolutional Neural Networks (CNNs) have local connectivity and translation invariance, which allow them to effectively capture local correlations in input images. However, they lack long-range dependencies. While global attention mechanisms excel at capturing global features, they often come with high computational costs and greater resource consumption. The window-based self-attention and shifting window mechanisms in the Swin Transformer[15] can effectively capture pixel correlations within a window and enable information exchange across windows, thereby enhancing the network's ability to model global relationships. To address this, we propose an efficient Visual Dual-Flow Integration Block (VDFI) to focus on different levels of feature map information, modeling both local pixel correlations and global semantic dependencies. This combines local and global information, including multi-grained features such as the overall layout and local structural details of text characters, to provide rich visual information for image reconstruction, which is helpful for processing deformed and distorted text images.

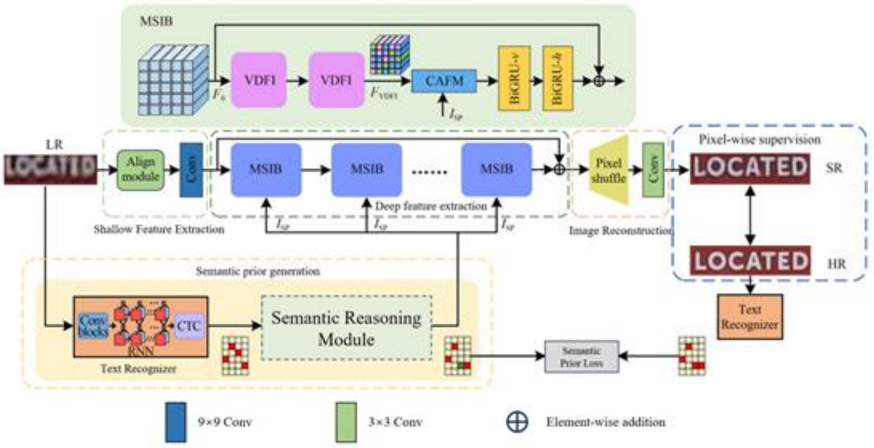

Figure 1: Overall architecture of MSISR

In text, coarse-grained refers to inter-character information, which includes spatial deformation of text lines. We model this using the Transformer self-attention mechanism and the shifting window mechanism to learn long-range dependencies between characters. Fine-grained refers to intra-character information, which uses convolutional networks and the self-attention of the Transformer’ s local window to collaboratively learn short-range dependencies between characters.

The specific structure of the VDFI module is shown in Figure 2. In the Swin Transformer layer, we have added a convolution block (CB), composed of convolution, batch normalization (BN), activation function (GELU), and Efficient Channel Attention (ECA), to enhance the network's representation capability. The CB module uses two 3x3 convolutional layers for local feature extraction. To reduce computational cost, the number of channels is compressed through the first convolution layer and then restored after the second convolution layer. Finally, the ECA adaptively adjusts channel features to refine them.

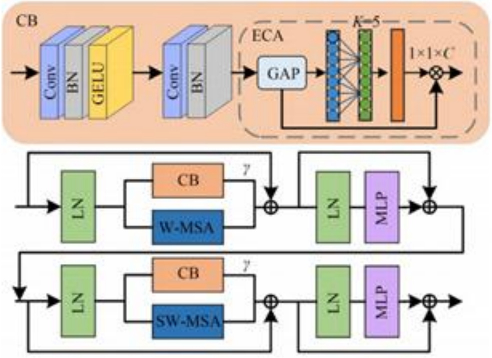

Figure 2: Visual Dual Flow Integration Module

In the VDFI module, after the first Layer Normalization (LN) layer, the CB block and Multi-Head Self-Attention (MSA) module are parallel to each other, leveraging the complementary advantages of global statistical features and strong local fitting capability. To ensure coordination and stability between the CB and MSA blocks during optimization, the output of the CB is regulated using a balancing parameter yyy. Then, the LN layer is followed by a Multilayer Perceptron (MLP) and residual connections on the outer layers.

$$\begin{gathered} F_1 = MSA\left(LN\left(F_0\right)\right) + \gamma CB\left(LN\left(F_0\right)\right) + F_0, \\ F_{VDFI} = MLP\left(LN\left(F_1\right)\right) + F_1 \end{gathered} \tag{4}$$

2.4.2 Cross-modal Adaptive Fusion Block

Text images contain important content information, which plays a crucial role in guiding image reconstruction tasks. However, most STSR methods tend to overlook this information. To address this issue, we use a CRNN and semantic reasoning predictions to obtain rich content information in the form of semantic features I_{SP}. These semantic features provide a high-level understanding of the text image content and are of a different modality compared to multi-grained visual features F_{VDFI} .To overcome the underlying feature gap between these different modalities and to adaptively learn the information correlation between visual and semantic features, we introduce the CAFM to integrate semantic information into the deep feature construction. The specific structure of the CAFM module is shown in Figure 3. The CAFM module consists of four key components: feature conversion, refinement, aggregation, and dual-scale channel attention. To match the dimensions of the visual features F_{VDFI}, the semantic features I_{SP} undergo feature conversion through three deconvolution layers with strides of (2, 2) and one deconvolution layer with a stride of (2, 1), resulting in a semantic feature map F_{SP}. The visual feature map F_{VDFI} and the semantic feature map F_{SP} are then concatenated along the channel dimension and refined through serial Spatial-Channel Reconstruction Convolution (SCConv)[16] blocks.

The SCConv blocks consist of Spatial Reconstruction Units (SRU) and Channel Reconstruction Units (CRU) that reduce redundant information along the spatial and channel dimensions, respectively, minimizing the interference from background and incorrect semantic information. Next, three parallel 1x1 convolutions are applied to compress the channels and perform feature aggregation. The aggregated features are projected onto three feature spaces. Global Average Pooling (GAP) is used to adjust and perform parallel local and global attention weight calculations. These attention weights are multiplied with the features to adaptively select the most relevant features and assign different weights based on the information differences. Finally, the enhanced features are obtained through a residual connection, allowing for the fusion of both semantic and visual features for improved text image reconstruction.

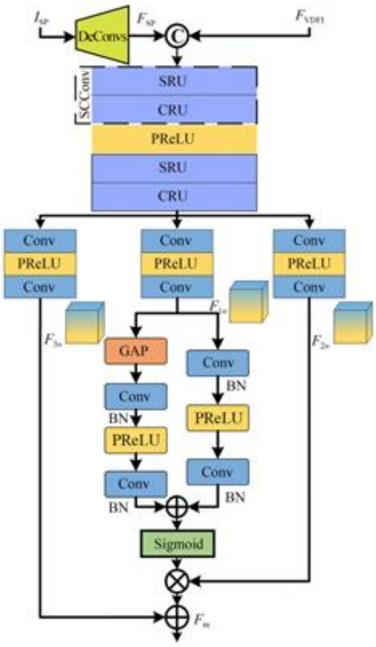

Figure 3: Cross-modal Adaptive Fusion Module

2.4.3 Bidirectional Gated Recurrent Unit

In scene text images, the textual information is mainly concentrated in two directions: horizontal and vertical. The horizontal context provides semantic relationships between characters, while the vertical context offers internal features of characters, such as strokes. As shown in Figure 1, based on the sequential nature of text data, we use Bidirectional Gated Recurrent Units (BiGRU) in both vertical and horizontal directions to capture the multimodal features in these two directions. Specifically, we employ BiGRU-v to capture the vertical direction information and BiGRU-h to capture the horizontal direction information, thereby establishing the text dependencies in both directions.

2.5 Loss Function

MSISR is trained using multiple loss functions, mainly including pixel loss, edge-aware loss, and semantic prior loss. The pixel loss primarily uses the L_2 loss between the corresponding pixels of the SR and HR images, as follows:

$$L_2 = \left\| I_{SR} - I_{HR} \right\|^2 \tag{5}$$

The edge information of text contains key features such as text shape, contour, and structure, which are crucial for understanding and processing the text content. To avoid excessive smoothing of text character edges in the SR image, we propose an edge-aware loss L_{EP}, as follows:

$$L_{EP} = \left\| f\left(I_{HR}\right) - f\left(I_{SR}\right) \right\|_1 \tag{6}$$

where $f\left(\cdot\right)$ is the edge extraction operator.

The semantic prior loss L_{TP} is used to enhance the guiding role of text semantic information, further improving the image reconstruction effect, as follows:

$$L_{TP} = \lambda_1 \left\| L_{SP} - H_{HP} \right\| + \lambda_2 D_{KL}\left(L_{SP} \,||\, H_{HP}\right) \tag{7}$$

where λ and β are balancing parameters, $D_{KL}\left(L_{SP} \,||\, H_{HP}\right)$ is the KL divergence between I_{SP} and I_{HP}.

Based on the above, the total loss of the MSISR network can be expressed as:

$$L = \alpha L_2 + \beta L_{EP} + \lambda L_{SP} \tag{8}$$

where α , β , and λ are balancing parameters, and we set them as 1, 1×10^{-4}, and 1, respectively.

3. Experiments

3.1 Datasets and Evaluation indicators

We use the TextZoom dataset, which is specifically designed for scene text image super-resolution reconstruction. This real-world scene text image dataset contains 21,740 high and low-resolution image pairs taken by a digital camera. Among these, 17,367 pairs are used for training, and the rest are used for testing. The test images are typically divided into three subsets based on the focal length of the camera: easy (1,619 pairs), medium (1,411 pairs), and hard (1,343 pairs). For the fixed input size of 32x128 for the text recognizer, a 2x super-resolution reconstruction is performed, with the LR size being 16x64 and the HR size being 32x128.

The core goal of scene text image super-resolution reconstruction is to improve the text recognition model's accuracy on LR text images. Therefore, we use three mainstream text recognition networks. ASTER, CRNN, and MORAN to recognize the reconstructed text images, with the recognition accuracy serving as the main evaluation metric for the reconstruction network.

3.2 Environment and Parameter Settings

All experiments were implemented on a single NVIDIA GTX 3090 GPU using PyTorch 1.9 and Python 3.9. The Adam optimizer was used for parameter optimization, with a momentum of 0.9 and a batch size of 48. The learning rate was set to 10^{-3}, and the training was conducted for 500 epochs.

3.3 Compare Experiments and Results

3.3.1 Objective Index Analysis

To validate the effectiveness of our proposed method, we conducted 2x super-resolution reconstruction experiments on the publicly available TextZoom dataset and compared it with 11 state-of-the-art super-resolution methods, including Bicubic, SRCNN[3], HAN[4], TSRN[9], PCAN**Error! Reference source not found.**, TBSRN[20], TG[21], MTSR**Error! Reference source not found.**, TATT[12], DPGSR[13], and TPGSR[10]. For TPGSR, we compared the TPGSR and TPGSR-3. Table 1 lists the average text recognition accuracy of the reconstructed images by different methods, with the highest accuracy in each group indicated in bold and the second-best method underlined. "Avg" represents the weighted average calculated based on the number of samples in each subset. The results show that the traditional Bicubic method has the lowest recognition rate. The recognition accuracy of images reconstructed by SRCNN and HAN methods is higher than that of Bicubic, but as general models for image super-resolution reconstruction, their results are not optimal due to their lack of ability to handle specific scene text images. Compared with SRCNN and HAN, TSRN and PCAN methods, which use LSTM to capture text context information, show a significant improvement in text recognition accuracy. TBSRN and MTSR methods, based on self-attention mechanisms to capture long-range dependencies in text images, achieve relatively good results but did not see significant improvements in recognition accuracy due to the lack of local detail information. Benefiting from the incorporation of text prior information, methods like TG, TATT, DPGSR, and TPGSR, which apply various text attributes to the SR network, achieve relatively better recognition accuracy. Our method performs the best, with the weighted average recognition accuracy improving by 16.9%, 17.6%, and 26.8% for ASTER, MORAN, and CRNN, respectively, compared to Bicubic upsampling. Compared to the mainstream TPGSR, our model increases the average recognition accuracy by 2.9%, 3.6%, and 3.7%.

Table 1: Recognition accuracy of different methods on TextZoom(%)

Method	ASTER				MORAN				CRNN			
	easy	medium	hard	avg	easy	medium	hard	avg	easy	medium	hard	avg
Bicubic	64.7	42.4	31.2	47.2	60.6	37.9	30.8	44.1	36.4	21.1	21.1	26.8
SRCNN[3]	69.4	43.4	32.2	49.5	63.2	39.0	30.2	45.3	38.7	21.6	20.9	27.7
HAN[4]	71.1	52.8	39.0	55.3	67.4	48.5	35.4	51.5	51.6	35.8	29.0	39.6
TSRN[9]	75.1	56.3	40.1	58.3	70.1	53.3	37.9	54.8	52.5	38.2	31.4	41.4
PCAN**Error! Reference source not found.**	77.5	60.7	43.1	61.5	73.7	57.6	41.0	58.5	59.6	45.4	34.8	47.4
TBSRN[20]	75.7	59.9	41.6	60.0	74.1	57.0	40.8	58.4	59.6	47.1	35.3	48.1
TG[21]	77.9	60.2	42.4	61.3	75.8	57.8	41.4	59.4	61.2	47.6	35.5	48.9
MTSR**Error! Reference source not found.**	75.6	59.8	43.4	58.9	73.9	57.2	41.8	56.0	56.2	47.0	35.3	45.4
TATT[12]	78.9	63.4	45.4	63.6	72.5	60.2	43.1	59.5	62.6	53.4	39.8	52.6
DPGSR[13]	75.5	57.8	41.9	59.4	69.7	53.4	39.7	55.2	57.6	43.0	33.4	45.5
TPGSR[10]	77.0	60.9	42.4	61.2	72.2	57.8	41.3	58.1	61.0	49.9	36.7	49.9
TPGSR-3[10]	78.9	62.7	44.5	62.8	74.9	60.5	44.1	60.5	63.1	52.0	38.6	51.8
Ours	**80.0**	**63.6**	**45.6**	**64.1**	**76.5**	**60.9**	**44.8**	**61.7**	**64.8**	**54.0**	**39.8**	**53.6**

3.3.2 Comparison of Different Methods

To visually demonstrate the advantages of our method, we performed a visualization comparison. Note that the TBSRN and MTSR methods did not provide related resources, so they were not included in the visual comparison. We selected two images from each of the three subdatasets of the TextZoom test set for visual effect comparison, as shown in Figure 4. In the figure, red characters represent incorrectly recognized characters (for color images, please refer to the electronic version of the journal).The images reconstructed by the Bicubic method suffer from excessive smoothing, leading to an overall blurry visual effect, and they fail to capture sharp character edges. The visual effect of the SCRNN and HAN methods does not show significant improvements, with poor edge integrity. Although the TSRN, PCAN, TBSRN, and TG methods can achieve relatively good text image reconstruction results, they still face issues with detail handling, such as fuzzy boundaries between characters and adhesion between adjacent characters. Compared to the earlier methods, while TPGSR and TATT methods can generate clearer text images, they still reconstruct incorrect character information, lack detail reconstruction in text areas, and exhibit artifacts in the text edges. Our proposed method is able to better reconstruct semantically accurate text images, restore sharp character edges, improve the visual quality of scene text images, and is more suitable for the STISR (Scene Text Image Super-Resolution) reconstruction task.

3.4 Ablation Study

We employ VDFI to extract multi-granularity features from text images, uses CAFM to learn the information correlation between different modalities, and combines edge loss L_{EA} to supervise the reconstruction of text edges. To investigate the impact of different modules on the final reconstruction results, this section analyzes the effectiveness of VDFI, CAFM, L_{EA}, and semantic prior information on three test subsets: easy, medium, and hard. The experimental results are presented in Table 2. Here, avg represents the weighted average based on the sample sizes of each subset, the best recognition rate is highlighted in bold, × indicates that the corresponding module was not used, and √ indicates that the corresponding module was used. Swin refers to the use of Swin Transformer to extract deep features from images. The experimental results show that the model achieves its best performance when all modules are included. Compared to Swin Transformer, the VDFI module model can improve recognition accuracy by 0.7%

Table 2: Recognition accuracy of different modules

Swin	Semantic prior	VDFI	CAFM	L_{EA}	avg/%
×	×	√	×	×	44.2
×	√	√	×	×	52.3
×	√	×	√	×	52.0
×	√	×	×	√	51.4
×	√	√	√	×	53.2
√	√	×	√	√	52.9
×	√	√	√	√	53.6

4. Conclusion

We propose a multimodal semantic interaction-based scene text image super-resolution network for reconstructing semantically accurate text images. By leveraging a semantic reasoning module and a multimodal feature extraction backbone, our approach effectively integrates local and global information, learns semantic interactions between modalities, and captures horizontal and vertical textual dependencies. Experiments on the TextZoom dataset show that our method outperforms state-of-the-art models, with improvements in PSNR, SSIM, and recognition accuracy. Our approach enhances text readability and sets the foundation for future work on non-Latin scripts super-resolution.

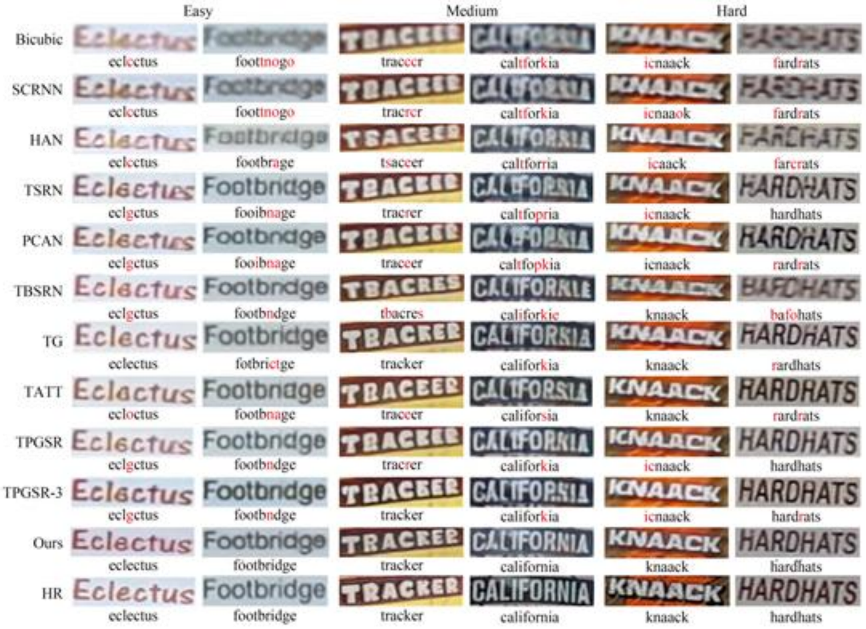

Figure 4: Visualization Results of Different Methods on TextZoom

References

[1] Guan T, Shen W, Yang X, et al. Self-supervised character-to-character distillation for text recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19473-19484.

[2] Li M, Lv T, Chen J, et al. Trocr: Transformer-based optical character recognition with pre-trained models[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(11): 13094-13102.

[3] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. Springer International Publishing, 2014: 184-199.

[4] Niu B, Wen W, Ren W, et al. Single image super-resolution via a holistic attention network[C]//Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020: 191-207.

[5] Zhu S, Zhao Z, Fang P, et al. Improving scene text image super-resolution via dual prior modulation network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(3): 3843-3851.

[6] Wang W, Xie E, Sun P, et al. Textsr: Content-aware text super-resolution guided by recognition[J]. arXiv preprint arXiv:1909.07113, 2019.

[7] Wang Y, Su F, Qian Y. Text-attentional conditional generative adversarial network for super-resolution of text images[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 1024-1029.

[8] Mou Y, Tan L, Yang H, et al. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit[C]//Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020: 158-174.

[9] Wang W, Xie E, Liu X, et al. Scene text image super-resolution in the wild[C]//Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part X 16. Springer International Publishing, 2020: 650-666.

[10] Ma J, Guo S, Zhang L. Text prior guided scene text image super-resolution[J]. IEEE Transactions on Image Processing, 2023, 32: 1341-1353.

[11] Ma J, Liang Z, Zhang L. A text attention network for spatial deformation robust scene text image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5911-5920.

[12] Yang H, Zhou H. Degradation Prior Guided Scene Text Image Super-Resolution[C]//2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC). IEEE, 2022: 170-175.

[13] Ma J, Jin L, Zhang J, et al. Textsrnet: scene text super-resolution based on contour prior and atrous convolution[C]//2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022: 3252-3258.

[14] Fu X, Ch'ng E, Aickelin U, et al. CRNN: a joint neural network for redundancy detection[C]//2017 IEEE international conference on smart computing (SMARTCOMP). IEEE, 2017: 1-8.

[15] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[16] Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6153-6162.

[17] Shi B, Yang M, Wang X, et al. Aster: An attentional scene text recognizer with flexible rectification[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(9): 2035-2048.

[18] Luo C, Jin L, Sun Z. Moran: A multi-object rectified attention network for scene text recognition[J]. Pattern Recognition, 2019, 90: 109-118.

[19] Zhao C, Feng S, Zhao B N, et al. Scene text image super-resolution via parallelly contextual attention network[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 2908-2917.

[20] Chen J, Li B, Xue X. Scene text telescope: Text-focused scene image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12026-12035.

[21] Chen J, Yu H, Ma J, et al. Text gestalt: Stroke-aware scene text image super-resolution[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 285-293.

[22] Honda K, Kurematsu M, Fujita H, et al. Multi-task learning for scene text image super-resolution with multiple transformers[J]. Electronics, 2022, 11(22): 3813.