Scene Text Image Super-Resolution via Content Perceptual Loss and Criss-Cross Transformer Blocks

Rui Qin  
School of Software  
Tsinghua University  
Beijing, China  
qr20@mails.tsinghua.edu.cn

Bin Wang*  
School of Software  
Tsinghua University  
Beijing, China  
wangbins@tsinghua.edu.cn

Abstract—Text image super-resolution is a unique and vital task aimed at enhancing the readability of text images to humans. It frequently serves as a pre-processing step in scene text recognition. Nevertheless, due to the complex degradation in natural scenes, recovering high-resolution texts from low-resolution inputs is ambiguous and challenging. Predominantly, existing methods employ deep neural networks trained with pixel-wise losses, tailored for natural image reconstruction, yet neglecting the unique characteristics intrinsic to text. While a limited number of studies proposed content-based losses, these primarily concentrate on the accuracy of text recognizers, resulting in reconstructed images that may still be ambiguous to humans. Moreover, these approaches typically exhibit inadequate generalizability when dealing with cross-language cases. To this end, we present TATSR, a Text-Aware Text Super-Resolution framework, which effectively learns the unique text characteristics using Criss-Cross Transformer Blocks (CCTBs) and a novel Content Perceptual (CP) Loss. The CCTB, consisting of two orthogonal transformers, is designed to extract both vertical and horizontal content information from text images. The CP Loss supervises text reconstruction by integrating content semantics through multi-scale text recognition features, thereby embedding content awareness effectively into the framework. Extensive experiments on different language datasets demonstrate that TATSR outperforms state-of-the-art methods in terms of both recognition accuracy and human perception. Codes are released at [https://github.com/Imalne/TATSR.git](https://github.com/Imalne/TATSR.git).

Index Terms—Scene Text Image Super-Resolution, Scene Text Recognition (STR), Text Image, Attention, Convolutional Neural network

I. INTRODUCTION

Scene Text Recognition (STR) aims to extract character sequences from real-world scene text images. This task has gathered significant attention due to its relevance in various scene-based text-related applications such as document retrieval [1] and license plate recognition [2]. However, in real-world scenarios, texts often occupy a small region and can be blurred, resulting in poor STR performance. To tackle these challenges, researchers investigate the Scene Text Image Super-Resolution (STISR) to enhance the quality of low-resolution text images for better recognition performance. Previous works have indicated that the pixel-wise losses commonly employed in natural image super-resolution may not be suitable for STISR as these approaches tend to overlook the significance of text contexts, such as character shape and text sequence information. In an effort to better align with text data, a few recent studies have integrated text-aware characteristics into the supervision process for better restoration and recognition performance. Scene Text Telescope (STT, [3]) and the follow-up works [4], [5] carry out text recognition on the super-resolution results and supervise the outputs and attention maps using the ground truth text labels.

However, we find that 1) the supervision using text recognition accuracy makes the trained network tend to maximize the character probabilities, which is in conflict with the pixel-wise losses that maximize the image similarity with the ground truth text images. 2) The feedback from the recognizer is not consistently reliable. The discrepancy between these two types of losses can destabilize the training process, often culminating in suboptimal outcomes. 3) The losses based on recognition results are significantly constrained by the language type, resulting in their ineffectiveness across different languages. 4) The supervision with only the recognizer’s feedback is insensitive to the local structure of characters (e.g., shapes and strokes), resulting in poor visual quality (e.g., Fig. 1).

To address the aforementioned challenges, we introduce a novel character-shape-sensitive text-aware loss, termed the Content Perceptual (CP) Loss, which aligns well with pixel-wise losses and is capable of cross-language utility. Specifically, rather than relying on the recognizer’s final feedback, we opt for the recognizer’s multi-scale features as the carrier of text information. This selection allows for supervising the
similarity between super-resolution results and high-resolution images within text-oriented feature spaces. The CP Loss can help the network to learn both precise character structure (from the local features) and text-aware content (from the global features) while simultaneously keeping the convergence point consistent with pixel-wise losses. Notably, given the cross-language similarities in parts of the text priors, such as stroke layouts and character shapes, CP Loss can provide effective text-related supervision even in cross-language scenarios.

In addition to text-oriented supervision, recent STISR methods employ specific sequence processing modules for learning text priors, including character shape and text sequence information. This is predominantly achieved through modules based on Recurrent Neural Networks (RNNs). Nevertheless, the restricted memory capacity of RNNs tends to undermine the performance of RNN-based frameworks when dealing with long text samples. To address this problem, we propose a new sequence processing block, the Criss-Cross Transformer Block (CCTB), which is more adaptive to arbitrary-length texts. With the global visibility afforded by multi-head attention, the transformer-based CCTB is capable of modeling content information of arbitrary-length texts and proves especially beneficial in long text cases. Given the typical layout of text characters, we distinguish the learning of character shape and sequence information using a criss-cross sparse strategy, which not only reduces the complexity of learning but also proves more suitable for text content. The contributions of our work can be summarized as follows:

1) We propose a novel character-shape-sensitive, text-aware loss function that aligns with pixel-wise losses and is cross-language utilizable, termed the Content Perceptual (CP) Loss. It calculates the discrepancy between super-resolution results and ground truth within multi-scale Scene Text Recognition (STR) feature spaces, thereby effectively addressing the issues inherent in preceding text-oriented losses.

2) We propose a new sequence processing block called Criss-Cross Transformer Block (CCTB), which utilizes the Transformer to model arbitrary-length text sequence information. In addition, it incorporates a criss-cross sparse design to aptly adapt text data, thereby addressing the limitations of RNN-based blocks in handling lengthy text samples.

3) We construct a new STISR framework combining the CP Loss and CCTB, called Text-Aware Text Super-Resolution (TATSR). Experiments on various language datasets show that our proposed TATSR can effectively restore low-resolution scene text images and has visible improvements in visual perception and text recognition accuracy, achieving new state-of-the-art performance.

II. RELATED WORK

A. Scene Text Recognition

Traditional Scene Text Recognition (STR) methods [6], [7] often adopt a bottom-up strategy by single-character recognition and splicing. In contrast, recent solutions often adopt a top-down strategy that treats the scene text recognition task as an Image-to-Sequence task. CRNN [8] first combined the Convolutional Neural Network (CNN) with Bidirectional Long Short-Term Memory (BLSTM, [9]) for text feature extraction and introduced the Connectionist Temporal Classification (CTC, [10]) Loss for alignment between feature sequences and text labels. Attention-based frameworks [11], [12] have been recently studied because of their robustness to text data with diverse text lengths and shapes.

In practice, CNN features were widely used as the basic feature extractor for both top-down and bottom-up strategies. Thus, it is natural to consider the CNN features extracted from STR models have captured the information closely related to character shape and text sequences. Based on this inspiration, we build our Content Perceptual Loss.

B. Scene Text Image Super Resolution

Unlike the Single Image Super-Resolution (SISR), the Scene Text Image Super-Resolution (STISR) solutions focus more on optimizing human eye perception and recognizability of the text images. Except for the earlier work [13], [14] migrated from SISR, recent STISR works focus on mining the text-specific contextual information such as character shape and text sequence. The main directions include super-resolution models specific to text data and task-specific text-oriented losses.

1) Text-Specific Super-Resolution Model: Like STR, recent STISR models [3]–[5], [15]–[18] are also mainly based on CNN and sequence modeling modules. TSRN [16] and its follow-up works [4], [5], [18] use BLSTM [9] as the main sequence modeling module. However, due to the limitation of the hidden state structure, BLSTM can not handle samples with long texts well. STT [3] introduces the Transformer [19] to replace the BLSTM-based sequence processing blocks. Due to its rough full-image attention calculation, the single dense transformer has to learn all text contexts (the character shape and the sequence information) from a totally-flattened pixel sequence. This increases learning difficulty for the network, causing slight improvement compared to the BLSTM-based modules even with both a larger global receptive field and stronger fitting ability.

2) Text-Oriented Loss: The previous text-oriented losses use the deep feedback of the text recognizer as a provider of text semantics and design metrics based on it. Based on the type of the feedback used, they can be divided into deep-feature-based and recognition-result-based. The deep-feature-based losses use the recognizer’s deep features for loss calculation. For instance, the Position-Aware (PA) Loss in STT [3] calculates the L2 distance of the transformer’s attention map. The recognition-result-based losses [3]–[5], [15], [18] mainly evaluate the difference of distribution between recognition results of super-resolution images and ground truth text labels. As discussed in Sec. I, both types of losses have weak supervision on the fine structure and edges of strokes. Besides, since the recognition results are highly correlated with
language types and recognizer’s accuracy, recognition-result-based losses are naturally unusable across languages and have unavoidable convergence conflicts with pixel-wise losses.

III. METHOD

In this section, we introduce the proposed Text-Aware Text Super-Resolution (TATSR) framework. We first start with a brief overview of our framework. Then, we describe our Content Perceptual (CP) Loss and Criss-Cross Transformer Block (CCTB) in detail, respectively.

A. Overview

Our proposed TATSR framework, depicted in Fig. 2, comprises three key components: the super-resolution network, the pixel-level supervision module, and the text-aware supervision module. In our framework, we initially align the low-resolution (LR) image and its binary mask through the adaptive Thin Plate Spline (TPS, [20]) module. Following alignment, a single convolution layer extracts CNN features, which then traverse a sequence of repeated Criss-Cross Transformer Blocks. Ultimately, the Pixel Shuffle [21] module upsamples the processed feature maps to yield super-resolution results. The training of our network is supervised by the pixel-level supervision module and the text-aware supervision module. The pixel-level supervision module operates in the RGB color space, computing the L2 loss and Gradient Prior Loss [16] directly with the high-resolution image. The text-aware supervision module, functioning in text-oriented feature spaces, calculates the high-level similarity via our Content Perceptual Loss, which integrates text information across multiple scales to address the limitations of previous text-oriented losses.

B. Content Perceptual Loss

In this section, we introduce Content Perceptual (CP) Loss in greater detail. Specifically, we extract the CNN part of the CRNN [8] model as the loss function network. The overall CRNN model has been pre-trained for scene text recognition, and the parameters are fixed during the loss calculation. The CNN part of the CRNN model has five downsampling operations through max-pooling and convolution. We extract the features after these five downsampling layers for loss calculation. Let \( \phi_j(x) \) (\( j = 1, 2, 3, 4, 5 \)) be the activations after the \( j \)th downsampling layer of the loss function network \( \phi \) when processing the input image \( x \in \mathbb{R}^{C_0 \times H_0 \times W_0} \). Assuming that the shape of \( \phi_j(x) \) is \( C_j \times H_j \times W_j \), the single-scale Content Perceptual Loss \( \mathcal{L}_{\text{fus}}^j \) after the \( j \)th downsampling layer between super-resolution (SR) image \( I_S \) and high-resolution (HR) image \( I_H \) can be calculated as follows:

\[
\mathcal{L}_{\text{fus}}^j(\phi, I_S, I_H) = \frac{1}{C_j H_j W_j} || \phi_j(I_S) - \phi_j(I_H) ||_2^2. \tag{1}
\]

The overall CP Loss consists of the weighted sum of each single-scale loss:

\[
\mathcal{L}_{\text{CP}}(\phi, I_S, I_H) = \sum_{j=1}^{5} \lambda_j \cdot \mathcal{L}_{\text{fus}}^j(\phi, I_S, I_H). \tag{2}
\]

As demonstrated in [22], the shallow and deep features from the pre-trained CNN focus on the local structure and global semantics, respectively. Therefore, by simultaneously assessing the similarity between HR images and SR images across multi-scale Scene Text Recognition (STR)-trained features, CP Loss ensures the consistency of high-level text contexts and low-level stroke structures at the same time. Since we use the distance between STR-based features as the similarity measurement instead of the final prediction results, we sidestep the conflict at the convergence point with pixel-wise losses caused by the direct loss calculation with text labels. Furthermore, in contrast to recognition-result-based loss, which heavily relies on text strings, CP Loss can transfer the text priors embedded in the features to languages not present in the training set.
This suggests that even if the recognition model has not been trained in a specific language, it can still be deployed for CP Loss computation.

Similarly, the Perceptual Loss [22] in the Single Image Super-Resolution (SISR) task uses the measurement of image-classification-based features, achieving great performance in SISR. However, it should be noted that we have a clear difference in the scope of the application. Owing to the significant domain gap between natural and text images, the image-classification-trained features focus on the local details of general images, having a limited perception of the text semantics and character shapes. In contrast, the features obtained from pre-trained text recognition models contain more text-oriented information, thereby better gauging the similarity between the foreground characters in the Super-Resolution (SR) and High-Resolution (HR) images. For a more intuitive explanation, we compare the iteration process of Mean Squared Error (MSE) Loss, Perceptual Loss, and CP Loss in Fig. 3. Compared with MSE Loss’s blurry results, Perceptual Loss restores some unrecognizable local details due to its text-unsuitable general priors (e.g., 1.5k iterations in Fig. 3) while CP Loss speeds up the optimization of text-oriented information, resulting in the quickest recovery of the important text foreground.

**Overall Loss Function** The overall loss function is composed of the pixel-wise part $L_{PE}$ and the content-aware part $L_{CA}$. The pixel-wise part is the weighted sum of L2 Loss $L_2$ and the Gradient Prior Loss $L_{GP}$. The Content-aware part is our Content Perceptual (CP) Loss $L_{CP}$. The calculation of the overall loss can be represented as:

$$L_{PE} = \lambda_2 L_2 + \lambda_{GP} L_{GP}, \quad L_{CA} = \lambda_{CP} L_{CP}, \quad L = L_{PE} + L_{CA},$$

where $\lambda_2$, $\lambda_{GP}$, and $\lambda_{CP}$ are hyperparameters to trade off the proportions of three kinds of losses. In experiments, we fix $\lambda_{GP}$ as $10^{-4}$, and further explore the balance of L2 Loss and CP Loss by changing the values of $\lambda_2$ and $\lambda_{CP}$. The experimental results can be found in Sec. IV-C, and we choose the best ($\lambda_2 = 0.1, \lambda_{CP} = 5 \times 10^{-4}$) as the final settings.

**C. Criss-Cross Transformer Block**

In this section, we detail our Criss-Cross Transformer Block (CCTB). The specific structure of CCTB is shown in Fig. 4.

CCTB contains two consecutive transformer encoders, which we call vertical transformer $\phi_v$ and horizontal transformer $\phi_h$, responsible for co-column and co-row learning, respectively. Specifically, we denote $\text{col}^j(X) \in \mathbb{R}^{C \times H \times 1}$, $\text{row}^i(X) \in \mathbb{R}^{C \times 1 \times W}$ as the $j$th column and the $i$th row of the feature map $X \in \mathbb{R}^{C \times H \times W}$. Assuming that the input feature map is $I$, the $j$th column of vertical transformer’s output $O_v$ and the $i$th row of horizontal transformer’s output $O_h$ can be calculated as:

$$\text{col}^j(O_v) = \phi_{col}(\text{col}^j(I)), \quad \text{row}^i(O_h) = \phi_{row}(\text{row}^i(O_v)), \quad (4)$$

The entire output of the vertical (horizontal) transformers is the concatenation of every single column (row) output,

$$O_v = \text{concat}(\text{col}^1(O_v), \text{col}^2(O_v), \text{col}^3(O_v), ..., \text{col}^{W}(O_v)), \quad (5)$$

$$O_h = \text{concat}(\text{row}^1(O_h), \text{row}^2(O_h), \text{row}^3(O_h), ..., \text{row}^{H}(O_h)). \quad (6)$$

This design splits the learning of different contexts into two orthogonal directions separately, which we call criss-cross learning. We opt for the criss-cross strategy because characters in natural scenes are predominantly arranged horizontally. Consequently, this streamlined strategy maximizes the segregation of text sequence and character shape information into horizontal and vertical directions, respectively. Intuitively, the learning of character shapes and text sequence information exerts mutual influence. Better single-character information facilitates the learning of text sequences and vice versa. As such, we elect the alternating layout strategy for the two types of transformers over the parallel one, thereby facilitating a more frequent interchange optimization. Compared to the previous ViT [23]-like all-flattened image-to-sequence strategy in STT [3], this explicit text context separation strategy mitigates the learning difficulty of each text context, effectively boosting the transformer’s performance in fitting different text contexts.

The global visibility of multi-head attention empowers CCTB to capture the complete sequence in the horizontal direction. It is able to memorize entire sequence information equally without forgetting any middle units. Therefore, CCTB can accurately model the sequence contexts of samples with texts of arbitrary length without the problem of RNN’s long-term memory failure, leading to better-predicting character types and fewer restored characters with type errors. This is especially beneficial for processing long text samples, which is more dependent on sequence information.

**IV. EXPERIMENT**

**A. Datasets and Metrics**

1) Datasets: TextZoom [16] dataset is a real-scene text image dataset cropped from RealSR [24] and SRRAW [25]. TextZoom contains 17,367 LR-HR image data pairs for training and three test subsets divided according to the focal lengths when shot, namely easy (1,619 samples), medium (1,411 samples), and hard (1,343 samples).

TextZoom$^1$ To approach the upper-performance limit, we employ a mixed degradation approach combining BSRGAN [26] and Real-ESRGAN [27] to synthesize diverse LR images of the HR samples in TextZoom dataset. For the sake of equitable comparison, this data is only used when explicitly

---

**Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.**
mentioned. By default, in all settings, we rely solely on the basic TextZoom for training purposes.

**ChineseSTR** is a scene text image dataset with Chinese texts as the main content generated by us, consisting of 161,219 train samples and 8,508 test samples. We sample HR images from PaddleOCR Datasets and synthesize the LR version using compound degradations in BSRGAN [26].

2) **Evaluation Metrics:** Since the primary purpose of this task is to improve the text recognition accuracy and human eye perception of text images, we choose the recognition accuracy of the open-source text recognition models and visual perception to measure the model’s performance. More specifically, we use the text recognition results of the pre-trained Aster [11], CRNN [8], and Moran [12] models for evaluation. To be consistent with the previous work, we exclude the influence of punctuation and capitalization when calculating accuracy.

Similar to previous work [3]–[5], [17], we do not use Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) as the main evaluation metrics because they are not suitable for text image evaluation. In text images, the character regions with a small proportion of pixels have a far more significant impact on perception than the large background. However, as mentioned in previous work [3], these metrics are averaged based on all image pixels, causing a much larger divergence than in the general images between the metrics and human perception. More intuitively, a restored character with sharp but misaligned boundaries (e.g., Fig. 5 (c)) tends to have lower PSNR/SSIM than a restored character with a blurry appearance (e.g., Fig. 5 (b)). This problem is more obvious on artificially cropped real scene datasets (e.g., TextZoom). Therefore, visual perception and text recognition accuracy are more accurate metrics for judging the performance of scene text image super-resolution.

3) **Implementation Details:** Our method is implemented using the PyTorch framework. For data preprocessing, following the previous works, we resize the low-resolution images and high-resolution images to $16 \times 64$ and $32 \times 128$. The experiments are trained on 5 NVIDIA RTX 2080 Ti GPUs. Unless otherwise stated, we set the number of CCTBs and feature channels to 4 and 128, respectively. The network is optimized using the Adam optimizer, and the initial learning rate is set to $5 \times 10^{-4}$. Training takes 500 epochs, and the batch size is set to 128. The used STR models are based on the officially released codes and the pre-trained parameters. The codes and weights are available at https://github.com/Imalne/TATSR.git.

**B. Comparison with State-of-the-Art**

In Tab. I and Fig. 6, we list the quantitative and qualitative results of thirteen different competitive methods on the TextZoom [16], including SRCNN [28], SRRResNet [29], RRDB [32], EDSR [30], LapSRN [31] belonging to the Single Image Super-Resolution methods and TSRN [16], STT [3], PCAN [17], TG [3], TATT [4], TPSOSR [33] and TATSR belonging to Scene Text Image Super-Resolution methods. To ensure a fair comparison, we use the basic settings (no TextZoom) in Fig. 6, as well as in the subsequent experiments. The results of the other methods are obtained from their official release codes. We do not provide the visual samples and metrics values on ChineseSTR of TPGSR [33] due to the lack of their official release weight.

First, the quantitative results in Tab. I show that TATSR markedly enhances text recognition accuracy across all text recognition models and test subsets. Compared with bicubic, TATSR achieves 18.5%, 17.3%, and 28.2% improvement under the three text recognition models, respectively. Moreover, it achieves the highest recognition accuracy across all three distinct recognition models, thereby affirming TATSR’s efficacy in augmenting the text recognition quality of low-resolution images. It merits emphasis that, even in the absence of supplementary synthetic training data, our TATSR methodology has already yielded substantial advancements over preceding methods. Furthermore, the incorporation of synthetic data serves to bolster the performance of TATSR, suggesting its potential to optimize generalization when coupled with larger, diversely degraded datasets. Second, as illustrated in Fig. 6, when compared with other methods, TATSR is able to generate more accurate and well-formed strokes and more legible characters, courtesy of our Content Perceptual (CP) Loss. Besides, the Criss-Cross Transformer Blocks (CCTBs) can better handle the text characteristics, especially when dealing with images containing long character sequences (e.g., row 2, 3, 5 and 7 in Fig. 6). In general, TATSR has achieved consistent improvements in both human eye perception and text recognition compared to the previous methods.

**C. Ablation Study**

1) **Effectiveness of Content Perceptual (CP) Loss:** To verify the effectiveness of CP Loss, we incorporated it in the training of five diverse models, and Tab. II presents the results of each model trained with and without CP Loss. Since our CP Loss uses the pre-trained CRNN [8] features, for fairness, we refrain from using the recognition accuracy of CRNN as the evaluation metrics for the comparison of CP Loss and other text-oriented losses. As an alternative, we choose the Aster [11] recognition accuracy as the evaluation metrics in the ablation study of CP Loss. The factor of text-oriented loss will also be blocked in the comparison of CCTB with other blocks (Tab. VII and Fig. 10). Evidently, CP Loss brings obvious improvement to each model, and it is noteworthy that with CP Loss, SRRResNet [29], initially designed for single
image super-resolution, achieves competitive precision with TSRN [16], which was specially designed for scene text image super-resolution, showing CP Loss’s great boost.

We further compare the performance of Perceptual (P) Loss [22], Content/Position-Aware (C/PA) Loss [3], and our CP Loss by training the same models. The quantitative and qualitative results are shown in Tab. III and Fig. 7, which indicate the advantage of our CP Loss over the other text-oriented losses. Compared to text-imperceptible Perceptual Loss, the text-oriented C/PA Loss and CP Loss can provide text-aware supervision, effectively mitigating the misjudgment of character types and improving recognition accuracy. Our CP Loss supervises low-level features while also providing deep text-aware supervision, enabling the recovery of sharper, more legible results.

**TABLE I**

| Method | Loss | easy | medium | hard | all |
|--------|------|------|--------|------|-----|
| BICUBIC | - | 64.7% | 42.4% | 31.2% | 47.2% |
| SRCNN (2016) [28] | $L_1$ | 69.4% | 43.4% | 32.2% | 49.5% |
| SRResNet (2017) [29] | $L_2 + L_{TV}$ | 69.4% | 47.3% | 34.3% | 51.3% |
| EDSR (2017) [30] | $L_1$ | 72.3% | 48.6% | 34.3% | 53.0% |
| LapSRN (2017) [13] | $L_{charbonnier}$ | 71.5% | 48.6% | 35.2% | 53.0% |
| RRDB (2018) [32] | $L_1$ | 70.9% | 44.4% | 32.5% | 50.6% |
| TSRN (2020) [16] | $L_2 + L_{GC}$ | 75.1% | 56.3% | 40.1% | 58.3% |
| STTR (2021) [3] | $L_2 + L_{C/PA}$ | 75.7% | 59.9% | 41.6% | 60.1% |
| PCAN (2021) [7] | $L_2 + L_{GC}$ | 77.5% | 60.7% | 43.1% | 61.5% |
| TG (2022) [5] | $L_2 + L_{C/PA}$ | 77.9% | 60.2% | 42.4% | 61.3% |
| TATT (2022) [4] | $L_2 + L_{TV} + L_{BC}$ | 78.9% | 63.4% | 45.4% | 63.6% |
| TPGRS (2023) [13] | $L_1 + L_{L1} + L_{CP}$ | 78.9% | 62.70% | 44.5% | 62.8% |

**TABLE II**

| Method | $L_{CP}$ | easy | medium | hard | all |
|--------|----------|------|--------|------|-----|
| SRCNN [28] | × | 69.4% | 43.4% | 32.2% | 49.5% |
| SRResNet [29] | ✓ | 70.6% | 45.7% | 33.6% | 51.2% |
| TSRN [16] | × | 75.1% | 56.3% | 40.1% | 58.3% |
| TBSRN [3] | ✓ | 74.9% | 60.6% | 42.1% | 60.5% |
| TATSR | ✓ | 75.6% | 58.9% | 42.7% | 60.1% | 80.4% | 64.6% | 45.9% | 64.7% |

**TABLE III**

| Method | Loss | easy | medium | hard | all |
|--------|------|------|--------|------|-----|
| TSRN [16] | × | 75.4% | 58.3% | 41.7% | 59.5% |
| + $L_{C/PA}$ | ✓ | 73.4% | 59.3% | 39.6% | 58.9% |
| + $L_{CP}$ | ✓ | 74.9% | 60.6% | 42.1% | 60.2% |
| TBSRN [3] | × | 75.2% | 56.7% | 40.2% | 58.5% |
| + $L_{C/PA}$ | ✓ | 73.4% | 59.6% | 41.6% | 58.2% |
| + $L_{CP}$ | ✓ | 77.0% | 59.9% | 41.6% | 60.1% |
| TATSR | × | 75.6% | 58.9% | 42.7% | 60.1% | 80.4% | 64.6% | 45.9% | 64.7% |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
to training failure. Generally, we adopt the settings introduced by all auxiliary losses, are more pronounced when increasing the text relevance of supervision.

2) Interpretability Analysis of Content Perceptual (CP) Loss: For further analysis of CP Loss, we conduct ablative experiments on feature scales and weight distribution. The results are illustrated in Tab. IV and Tab. V. Tab. IV reveals that combining low-level structure with global semantics results in optimal performance, while excessive emphasis on either can lead to the loss of the other type of information. Therefore, we selected the weight proportion that delivers the best results as the final setting. Additionally, we investigate varying proportions between pixel-wise losses and CP Loss. Tab. V shows that increasing the weighting of CP Loss enables the network to focus more on minimizing the distance between HR and SR images in text-aware feature spaces, constructed by multi-scale recognition-based features, and de-emphasizes the color space. This results in improved restoration of text foreground and increased text recognition accuracy. It’s important to note that since CRNN’s input is a grayscale image, relying solely on CP Loss and eliminating pixel-level loss could cause instability in the learning process, potentially leading to training failure. Generally, we adopt the settings that exhibit the highest performance as the final loss settings.

3) Cross-language Generalization of Content Perceptual (CP) Loss: To verify the cross-language generalization performance of CP Loss, we conducted cross-language comparative experiments on the Chinese-main dataset ChineseSTR and the English dataset TextZoom. Since the complete version of CP Loss depends on a pre-trained text recognition model for the corresponding language, we also trained a CRNN [8] model on ChineseSTR for Chinese-based CP Loss calculation.

As shown in Tab. VI, the losses corresponding to their specific languages yield optimal outcomes in both datasets. It is noteworthy that CP Loss achieves the second-best results on cross-lingual datasets, outperforming Perceptual Loss and demonstrating the excellent generalization performance of CP Loss across different languages.

These observations align with the visualizations presented in Fig.8. When contrasted with Perceptual Loss, the stroke shapes of Chinese characters restored by English-based CP Loss are more discernible. However, since English-based CP Loss does not include Chinese character sequence information, it is more prone to misjudging Chinese character types compared to Chinese-based CP Loss (for instance, row 3 in Fig.8). These findings support two conclusions: 1) CP Loss incorporates abundant text priors, including language-specific and universally applicable cross-language priors. The CP loss for a specific language yields optimal performance in the same language. 2) Texts across languages possess stronger domain priors (such as stroke shapes) as compared to generic image priors, and these priors prove to be effective in cross-language scenarios. Our CP Loss effectively transfers these priors to text images across varied languages, thereby securing better outcomes than the generic image prior loss.

4) Effectiveness of CCTB: To verify the effectiveness of our Criss-Cross Transformer Block (CCTB), we choose Sequential Residual Block (SRB, [16]), Transformer-Based Super-Resolution Network (TBSRN, [3]), Text-Prior Guided Blocks (TPGB, [4]), and our CCTB for comparison. For fairness, we compare all sequence modules under the same framework and using the same supervision for training. Specifically, we utilize the vanilla TBSRN [16] framework supervised by L2

| Table V | Comparison of TATSR’s performance with different assignments to $L_{MSE}$, $L_{CP}$, and $L_{CP}$ on TextZoom [16]. |
|---|---|---|---|---|
| $\lambda_{MSE} : \lambda_{CP}$ | easy | medium | hard | all |
| 1.0 : 0.0001 : 0.0001 | 76.6% | 58.9% | 42.7% | 60.1% |
| 1.0 : 0.0001 : 0.0005 | 76.6% | 61.6% | 44.8% | 61.6% |
| 0.1 : 0.0005 : 0.0005 | 80.4% | 64.6% | 45.9% | 64.7% |

| Table VI | Multi-language experiment on TextZoom [16] and ChineseSTR. |
|---|---|---|
| Loss | TextZoom(English) | ChineseSTR(Chinese) |
| $+L_P$ | 61.0% | 60.1% |
| $+L_{CP}^{C/PA}$ | 62.5% | 61.1% |
| $+L_{CP}^{CRNN}$ | 64.7% | 41.9% |

| Table VII | Accuracies of ASTER [11] on super-resolution results generated from different sequence processing blocks on TextZoom [16]. |
|---|---|---|---|
| Sequence Block | easy | medium | hard | all |
| SRB [16] | 75.1% | 56.3% | 40.1% | 58.3% |
| TBSRN [3] | 75.2% | 56.7% | 40.2% | 58.5% |
| TPGB [4] | 74.7% | 57.8% | 41.4% | 58.8% |
| CCTB | 76.8% | 61.0% | 41.3% | 60.8% |

Fig. 7. The super-resolution results of different text-oriented losses and the recognition results of CRNN [9].

TABLE IV
Comparison of the proportions of different single-layer losses $L^j_{feo}$ in $L_{CP}$ on TextZoom [16].

| $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4$ | easy | medium | hard | all |
|---|---|---|---|---|
| 1.6 : 1.6 : 0.0 : 0.0 | 78.0% | 63.9% | 45.6% | 63.5% |
| 1.4 : 1.4 : 0.4 : 0.4 | 80.4% | 64.6% | 45.9% | 64.7% |
| 1.0 : 1.0 : 1.0 : 1.0 | 79.1% | 63.4% | 47.1% | 64.2% |
| 0.5 : 0.9 : 1.2 : 1.2 | 76.2% | 61.9% | 45.7% | 62.2% |
Fig. 8. The examples of super-resolution results on ChineseSTR and their recognition results of Chinese-trained CRNN [8].

Fig. 9. Top: Recognition accuracy of different blocks on samples with varying text lengths on TextZoom [16]; Bottom: Improvement in recognition accuracy of CCTB over other blocks for samples with varying text lengths on TextZoom [16].

Loss and Gradient Prior Loss [16] as the uniform training framework and fix the number of blocks and feature channels to the commonly used 5 and 64, respectively. The quantitative results in Tab. VII validate our CCTB’s general superiority over the competing blocks. For an in-depth comparison, we display the visualization of different blocks’ results on long text samples in Fig. 10. Furthermore, Fig. 9 illustrates the enhancement in recognition accuracy achieved by CCTB over the other blocks for samples with varying text lengths. When dealing with long text images (Fig. 10), RNN-based SRB and TPGB struggle to accurately infer specific character types from the contextual information due to RNN's limited ability to model long-distance relationships. Despite transformer-based TBSRN producing fewer recognition errors, its results’ character shapes still remain coarse. In contrast, our CCTB effectively acquires sequence information and character shape by two text-layout-oriented orthogonal transformers, respectively. Fig. 9 shows the average recognition accuracy against text length; CCTB outperforms the other blocks on almost all samples with different text lengths.

5) Interpretability Analysis of CCTB: To gain a deeper understanding of our CCTB, we replace the two transformers in CCTB with BLSTM and evaluate their recognition accuracy of short ($\text{Len} < 9$) and long text samples ($\text{Len} \geq 9$) in Tab. VIII, respectively. Given the superior long-range modeling capability of the Transformer’s multi-head attention over the hidden state structure of BLSTM, using only the horizontal transformer yields a significant improvement in long text samples, though the enhancement in short texts remains relatively modest. In contrast to the sole use of the horizontal transformer, the vertical transformer in full CCTB enhances the recovery of the character shapes, leading to a performance boost in the overall data. These outcomes demonstrate that the criss-cross strategy effectively decouples the learning of different text priors into distinct directions and that the sparse transformers indeed enhance the learning of each text prior.

We also evaluate the performance of different transformer layout strategies in Tab IX. Due to more frequent iterative feature refinement, both vertical-first and horizontal-first strategies outperform the parallel strategy. As the order of the vertical and horizontal transformers has a minimal impact on the results, we select the configuration that yields the best outcome as the final setting.

D. Generalizability on Scene Text Recognition (STR) Benchmarks

We conduct experiments on six STR benchmark datasets, including IC03 [34], IC13 [35], SVT [37], IIIK50 [36], CUTE80 [38] and COCO [39], among which CUTE80 [38] and COCO [39] contain many samples with complex degradations, challenging for scene text image super-resolution.

1) Datasets: IIIK50 [36] consists of 3000 test instances, taken from street scenes and originally-digital images. SVT [37] SVT consists of 647 test image instances. Some images are severely degraded by noise, blur, and low resolution. IC03 [34], IC13 [35] The datasets used in the Incidental Scene Text Competitions include 867 and 1015 test images.
In this work, we propose a new scene text image super-resolution (STISR) framework named TATSR. By calculating the similarity between recognition-based multi-scale features of high-resolution and super-resolution images, CP Loss overcomes the problems of previous text-oriented losses, which conflict with pixel-wise losses, and cross-language unavailability. Meanwhile, a new sequence processing block called Criss-Cross Transformer Block (CCTB) is introduced to address the problems of previous text-oriented losses, which conflict with pixel-wise losses, and cross-language unavailability. Comprehensive experiments and ablation studies show that our TATSR framework can effectively improve the quality of low-resolution scene text images in terms of both text recognition accuracy and human perception, achieving a new state-of-the-art performance in STISR.

V. CONCLUSION

In this work, we propose a new scene text image super-resolution (STISR) framework named TATSR. By calculating the similarity between recognition-based multi-scale features of high-resolution and super-resolution images, CP Loss overcomes the problems of previous text-oriented losses, which conflict with pixel-wise losses, and cross-language unavailability. Meanwhile, a new sequence processing block called Criss-Cross Transformer Block (CCTB) is introduced to address the weaknesses of previous models on long text samples. Comprehensive experiments and ablation studies show that our TATSR framework can effectively improve the quality of low-resolution scene text images in terms of both text recognition accuracy and human perception, achieving a new state-of-the-art performance in STISR.

**TABLE X**

Text recognition accuracy of CRNN [8] model on six benchmark datasets. "Radius" denotes the radius of Gaussian kernels. "Preprocess" denotes the super-resolution method used before recognition ("-" means bicubic interpolation).

| Radius | Preprocess | IC03 [30] | ICD13 [31] | IIHIK50 [16] | SVT [37] | CUTE80 [38] | COCO [39] |
|--------|------------|------------|------------|--------------|-----------|--------------|------------|
| 0.5*   | +TSRN [16] | 78.1%      | 77.1%      | 70.0%        | 52.6%     | 46.3%        | 55.0%      |
|        | +STT [3]   | 79.8%      | 78.6%      | 71.5%        | 53.9%     | 47.2%        | 54.6%      |
|        | +TATT [4]  | 80.2%      | 79.3%      | 71.6%        | 54.7%     | 48.0%        | 55.2%      |
|        | +TATSR     | 81.6%      | 80.5%      | 72.6%        | 55.3%     | 48.5%        | 55.8%      |
| 1      | +TSRN [16] | 89.0%      | 87.6%      | 80.0%        | 58.4%     | 62.1%        | 65.1%      |
|        | +STT [3]   | 90.5%      | 89.2%      | 82.2%        | 59.3%     | 62.6%        | 65.6%      |
|        | +TATT [4]  | 91.2%      | 90.0%      | 82.6%        | 60.2%     | 63.0%        | 66.0%      |
|        | +TATSR     | 92.8%      | 91.5%      | 83.0%        | 60.9%     | 63.5%        | 66.5%      |
| 1.5*   | +TSRN [16] | 91.8%      | 90.6%      | 83.2%        | 61.2%     | 64.1%        | 67.1%      |
|        | +STT [3]   | 93.0%      | 91.8%      | 84.2%        | 62.1%     | 64.7%        | 67.7%      |
|        | +TATT [4]  | 93.6%      | 92.4%      | 84.5%        | 62.6%     | 65.1%        | 68.1%      |
|        | +TATSR     | 95.2%      | 94.0%      | 85.1%        | 63.3%     | 65.7%        | 68.7%      |

**TABLE XI**

Recognition accuracy on real low-resolution images. "Preprocess" denotes the super-resolution method used before recognition ("-" means bicubic interpolation).

| Preprocess | Aster [11] | Moran [12] | CRNN [8] |
|------------|------------|------------|----------|
| +TSRN [16] | 58.3%      | 51.4%      | 40.0%    |
| +STT [3]   | 61.8%      | 54.2%      | 42.1%    |
| +TATT [4]  | 62.1%      | 55.4%      | 43.5%    |
| +TATSR     | 62.8%      | 55.8%      | 44.8%    |

CUTE80 [38] consists of 288 test images. Since it focuses on curved text recognition, most samples have complex backgrounds, perspective distortion, and poor resolution.

COCO [39] consists of 39K image instances (9835 for evaluation) cropped from the MS COCO [40] dataset. Since MS COCO [40] is not designed for text capturing, many samples contain occluded or low-resolution texts.

2) Experimental Conditions: For fairness, we use all methods trained on TextZoom [16] dataset for comparison. Since STR datasets contain many high-resolution images in addition to low-resolution images, we conduct experiments under two different conditions: 1) All images undergo synthetic downsampling and degradation. 2) Real low-resolution images only.

In detail, for the first condition, we use all the images in six datasets. We use bicubic interpolation to resize images to 16 × 64. Then, we apply Gaussian blur kernels with various radii to simulate degradation found in real-world scenes. For the second condition, we select images smaller than 16 × 64 from all datasets and resize them to 16 × 64, totaling 3404 samples.

3) Results: Under these two experimental conditions, we perform STR on the super-resolution images restored by different preprocessing methods, including bicubic interpolation, TSRN [16], TBSRN [3], TATT [4], and our proposed TATSR.

As shown in Tab. X, all of the tested scene text image super-resolution methods visibly improve text recognition accuracy, especially on the heavily degraded images generated by large Gaussian kernels (e.g., R=1.5* or 2). This proves the effectiveness of scene text image super-resolution as a preprocessing operation of STR. Among all of them, TATSR brings the biggest performance boost on all six benchmarks, showing the superiority of our method. In addition, it is worth noting that under the condition of small Gaussian kernels (e.g., R=0.5* or 1), the text recognition accuracy of TSRN [16] and TBSRN [3] is even slightly inferior to the results of bicubic interpolation. We think this is caused by the domain gap between the TextZoom [16] dataset and the STR datasets. In contrast, TATSR successfully enhances the images under all experimental conditions, showing better generalization. Fig. 11 and Tab. XI show the performance of different methods on real low-resolution scene text images. After the pre-processing of TATSR, the accuracy increases by 4.53% for Aster [11], 7.23% for Moran [12], and 3.67% for CRNN [8].

Fig. 11. Examples of super-resolution results on real low-resolution images (no high-resolution ground truth). The first column shows the low-resolution images, and the four columns on the right show the super-resolution (SR) results. We present the SR images with their text recognition results at the top. The last column shows the ground-truth texts. The text recognition results are predicted by CRNN [8].
REFERENCES

[1] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, “Words matter: Scene-text for image classification,” pp. 12021–12030, 2021.
[2] J. Chen, B. Li, and X. Xue, “Scene text telescope: Text-focused scene image super-resolution,” Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 73–75, 2015.
[3] J. Chen, B. Li, and X. Xue, “Scene text telescope: Text-focused scene image super-resolution,” Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 73–75, 2015.
[4] J. Chen, H. Yu, J. Ma, B. Li, and X. Xue, “Text gestalt: Stroke-aware scene text image super-resolution,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 285–293, 2022.
[5] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Deep features for text spotting,” Proceedings of the European Conference on Computer Vision, pp. 512–528, 2014.
[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading text in the wild with convolutional neural networks,” International Journal of Computer Vision, vol. 116, no. 1, pp. 1–20, 2016.
[7] B. Shi, X. Dai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298–2304, 2016.
[8] R. A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Systems with Applications, vol. 41, no. 18, pp. 8027–8048, 2014.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
[10] M. Jaderberg, K. Simonyan, A. Zisserman et al., “Spatial transformer networks,” Advances in neural information processing systems, vol. 28, 2015.
[11] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1874–1883, 2016.