Received: 20 November 2022 | Revised: 4 March 2023 | Accepted: 10 April 2023

IET Image Processing

DOI: 10.1049/ipr2.12827

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

TLWSR: Weakly supervised real-world scene text image super-resolution using text label

Qin Shi[1] | Yu Zhu[1,3] | Chuantao Fang[1] | Dawei Yang[2,3]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China

[2]Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai, China

[3]Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai, China

Correspondence
Yu Zhu and Dawei Yang, School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.
Email: zhuyu@ecust.edu.cn and yang.dawei@zs-hospital.sh.cn

Funding information
Science and Technology Commission of Shanghai Municipality, Grant/Award Numbers: 20DZ2254400, 20DZ2261200; Fujian Province Department of Science and Technology, Grant/Award Number: 2022D014

Abstract

Scene text image super-resolution (STISR) has recently received considerable attention. Existing STISR methods are applicable to the situation that all the LR-HR pairs are available. However, in real-world scenarios, it is difficult and expensive to collect ground-truth HR labels and align them with LR images, and thus it is essential to find a way to implement weakly supervised learning. We investigate the STISR problem in the situation that only a subset of HR labels is available and design a weak supervision framework using coarse-grained text labels named TLWSR, which combines incomplete supervision and inexact supervision. Specifically, a lightweight text recognition network and connectionist temporal classification loss are used to guide the super-resolution of text images during training. Extensive experiments on the benchmark TextZoom demonstrate that TLWSR generates distinguishable text images and exceeds the fully supervised baseline TSRN in boosting text recognition accuracywith only 50% HR labels available. Meanwhile, TLWSR can be applied to different super-resolution backbones and significantly improves their performance. Furthermore, TLWSR shows good generalization capability to low-quality images on scene text recognition benchmarks, which verifies the effectiveness of this framework. To the authors' knowledge, this is the first work exploring the problem of STISR in weakly supervised scenarios.

1 | INTRODUCTION

Scene text recognition (STR) has been widely used in various domains such as text retrieval [1], license plate recognition [2], signature identification [3], autonomous driving [4]. With the rapid development of deep learning, text recognizers have achieved remarkable advancements in recent years by exploiting visual feature extraction [5–7], language modelling [8–11], loss function [12, 13], etc. However, how to effectively recognize low-resolution (LR) text images under real-world circumstances remains challenging [14].

Single image super-resolution (SISR) is an important technology in low-level computer vision, aiming at estimating a high-resolution (HR) image from its given LR counterpart. Recent studies have witnessed the promising improvement in SISR [15–20].

As a subfield of SISR, the purpose of scene text image super-resolution (STISR) is to improve the visual quality of LR text images by accurately reconstructing the blurry and illegible characters. It would be promising to introduce STISR methods as a preprocessor to enhance text recognition [14, 21]. STISR has been an active research topic recently. Early STISR works [22–24] are based on synthetic data where LR images are usually generated by applying uniform degradation (e.g. bicubic down-sampling) from HR images. Compared to the promising SR results on synthetic text images, their SR performances would drop sharply on real-world text images due to the domain gap between real-world and synthetic data. The authors in [14] constructs the first real-world paired STISR dataset named TextZoom where LR–HR pairs are collected by different cameras with different focal length. Figure 1 shows samples of real-world LR images and synthetic LR images, which are bicubic downsampled from HR images in TextZoom. It is obvious that recovering real-world scene text LR images is more challenging than synthetic LR images. Meanwhile, [14] designs Text Super-Resolution Network (TSRN) for STISR. Inspired by the

FIGURE 1 Comparisons between synthetic LR images, real-world LR images and HR images in TextZoom. HR, high resolution; LR, low resolution.

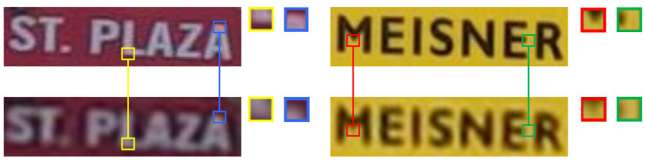

FIGURE 2 Illustration of misalignment between real-world low-resolution–high-resolution (LR-HR) pairs.

success of TSRN, many researchers have started to investigate real-world STISR to improve the quality of LR text images, thus improving recognition accuracy. However, all of the current works concentrate on recovering LR scene text images in a fully supervised manner, that is, with all the LR-HR pairs being used [25–29].

Despite the fact that real-world SR datasets are more practical in reality than synthetic SR datasets, there still exist some limitations. First, as depicted in Figure 2 , there always exists misalignment of pixels between real-world LR-HR pairs due to lens distortion and optical center shift when switching focal lengths, which would introduce blurry artifacts in the reconstructed SR images. Second, in previous non-real-world STISR tasks, HR images are the source used to synthesize LR images, and therefore, the cost of constructing datasets is relatively low. Nonetheless, real-world LR-HR pairs are often captured by adjusting the focal length of different digital single lens reflex (DSLR) cameras [30, 31], which is expensive and time-consuming. Recently, many researchers have proposed some methods to overcome the absence of LR-HR image pairs [32–34]. Nevertheless, these methods aim at natural images, which cannot apply to STISR directly due to the text-specific characteristics in text images. This paper focuses on STISR under the weakly supervised case aiming at reducing the reliance upon HR labels. The author in[35] classifies weak supervision learning into three typical types: incomplete supervision where only a subset of labels are given, inexact supervision where only coarse-grained labels are available and inaccurate supervision where the labels are not always ground-truth. In reality, the three types often occur simultaneously. Our proposed TLWSR is actually a framework combining incomplete and inexact supervision. To our knowledge, this is the first work to perform weakly supervised super-resolution on real-world scene text images.

As depicted in Figure 3a, the fully supervised STISR manner employs all the ground-truth HR labels to calculate L_2 loss between SR images and HR images. Considering that the text in the text images can be regarded as a natural coarse-grained label and is much easier to obtain compared with fine-grained HR labels, it is feasible and worth exploring to use the text label as a weak supervision signal to train the STISR model, as shown in Figure 3b. Specifically, the reconstructed SR image produced by the SR model is directly fed into a text recognizer (CRNN) [12]. Then, the text prediction result and the corresponding text label are supervised by connectionist temporal classification (CTC) loss [12], which can be back-propagated to guide the training of the SR network. Meanwhile, only a subset of HR labels is used to calculate L_2 loss. Our major contributions are listed as follows:

- To reduce dependencies on expensive and rare paired real-world text images, we propose a novel weak supervision framework named TLWSR. This is the first study investigating weak supervision in STISR.
- Given a subset of HR labels in the training set, this paper proposes a weak supervision framework which incorporates a text recognition network into the super-resolution network and utilizes CTC loss to facilitate generating sharp and identifiable text images for text recognition.
- Extensive experiments on TextZoom demonstrate that our weakly supervised framework with 50% HR labels used outperforms the fully supervised baseline TSRN and can be applied to different super-resolution backbones, bringing obvious improvements. In addition, it can be well generalized to images on scene text recognition benchmarks.

The overview of the paper is organized as follows. Section 2 introduces the related work in scene text recognition, scene text image super resolution and weak supervised learning. Then,

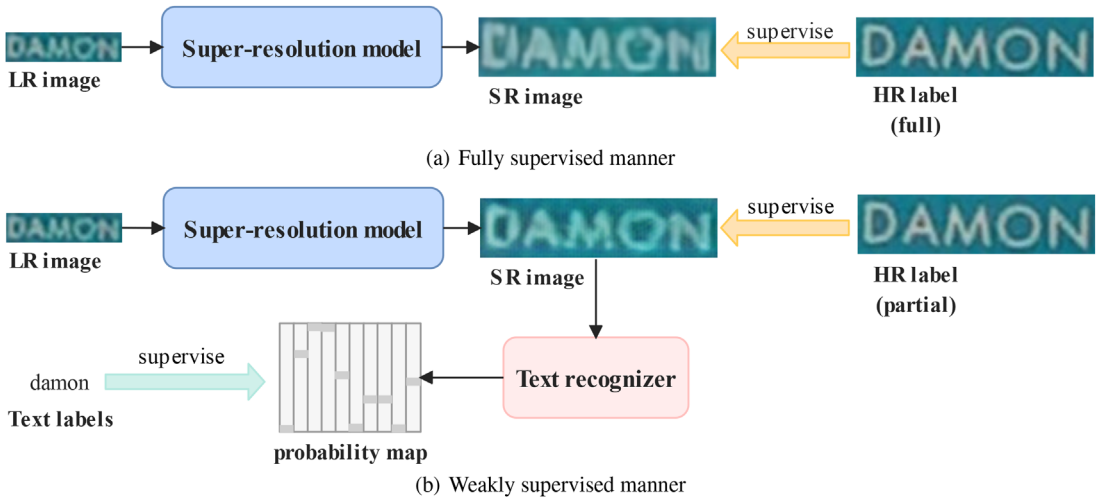

FIGURE 3 Schematic illustration of STISR networks dividing into two manners. HR, high resolution; LR, low resolution; SR, super resolution.

the proposed weak supervision frameworks are described in Section 3. Next, comprehensive experiments are conducted in Section 4. Finally, Section 5 concludes the whole work.

2 | RELATED WORK

2.1 | Scene text recognition (STR)

Scene text recognition (STR) is a widely concerning task in the field optical character recognition (OCR), aiming to identify text in natural images. Traditional text recognition methods [36, 37] mainly adopt a bottom-up strategy, that is, localizing individual characters using sliding window and then grouping them into words or sentences with lexicon search or graph models. These traditional methods show poor performances when confronted with some difficulties, that is, low-quality images, irregular text appearance and complex backgrounds. Benefiting from the development of deep neural networks, STR methods further progress into a top-down manner where text sequences are regarded as a whole and end-to-end predicted. Based on the kind of loss function, existing STR methods can be broadly categorized into CTC-based and attention-based methods [9]. CTC-based methods [12, 38] combine convolutional neural network (CNN) and recurrent neural network (RNN) to extract visual features and model the semantic information, respectively. Then, CTC loss is employed to align the predicted sequence and ground-truth sequence. ASTER [39] and MORAN [40] are representative attention-based methods, which rectify irregular text images and then utilize an attention-based bidirectional decoder to predict the character sequence. This paper employs CRNN and CTC loss to assist the reconstruction of SR text images. Experiments prove that our method can reduce dependencies on HR labels and exhibit superior performance compared with fully supervised methods when 50% HR images are available.

2.2 | Scene text image super resolution (STISR)

The purpose of STISR is to improve the readability of texts on LR images by recovering the blurry text images. TextSR [22] proposes an end-to-end network that utilizes the feedback of the text recognition network to guide the training of the super-resolution network. PlugNet [23] combines a pluggable super-resolution unit to recognize low-quality scene text which improves recognition accuracy significantly. The authors in[24] enhance the original cGAN model by introducing effective channel and spatial attention mechanisms, which enables the SR model to achieve better text image super-resolution results. The above studies focus on synthetic text data, which cannot be generalized to complex real scenarios. The authors in[14] build TextZoom, which is a real-world STISR dataset, and propose a TSRN using sequential residual block to capture sequential information and boundary-aware loss to sharpen the character boundaries. TSRGAN [25] introduces generative adversarial network (GAN) to prevent the SR network from generating over-smoothed results and incorporate triplet attention into the SR module for better representational ability. STT [26] proposes a text-focused super-resolution network to highlight the position and the content of each character. PCAN [27] designs effective attention mechanisms, aiming to learn sequence-dependent features and extract high-frequency information. TPGSR [28] employs a text prior generator to extract categorical probability distribution as guidance for the text image reconstruction process. Text Gestalt [29] pre-trains a text recognizer to highlight the stroke-level details. All of the previous works concentrate on recovering SR text images in a fully supervised manner, that is, with all the LR-HR pairs being used. As we know, this paper is the first attempt to perform weakly supervised super-resolution on real-world text images.

2.3 | Weakly supervised learning

Though current techniques have achieved great success, it is noteworthy that in many tasks it is difficult to get strong supervision information like fully ground-truth labels due to the high cost of the data-labelling process. Thus, it is desirable for machine-learning techniques to work with weak supervision [35]. Recently, many researchers have proposed SR methods to overcome the absence of HR-LR image pairs. The authors in[32] propose a two-stage algorithm which first employs a high-to-low network to learn how to degrade high-resolution images using unpaired HR-LR images and then use the output of this network to train a low-to-high network. The authors in[33] design a cycle-in-cycle network (CinCGAN), which contains two coupled CycleGANs to learn the mapping from degraded LR images to clean LR images and clean LR images to clean HR images, respectively. In order to reduce the dependence of existing STISR methods on expensive and rare paired real-world image super-resolution datasets, this paper concerns the situation where only a subset of HR ground-truths are given, which is more suitable and practical in real-world scenarios because there is often a lack of HR image corresponding to the given LR one.

3 | METHOD

In this section, the detailed introductions of the overall weak supervision framework TLWSR can be found in Section 3.1. Then, Sections 3.2 and 3.3 introduce the architecture of super-resolution network and text recognizer, respectively. Finally, text recognition loss and overall loss function are introduced in Section 3.4.

3.1 | Overall framework

Aiming at the situation where only a subset of HR ground-truths is available, this paper proposes a weak supervision STISR framework using text labels , namely, TLWSR. Figure 4 depicts the overall architecture. Moreover, the SR image generated by the SR model is input into a text recognition network, which predicts a probability map of the character sequence. In this case, low-level supervision is provided by reconstruction loss between a small number of HR labels and SR images. Meanwhile, text recognition loss is computed using all the text labels provided by TextZoom dataset.

3.2 | Super-resolution network

As shown in Figure 4, general super-resolution network based on convolutional neural networks (CNNs) can be divided into three main components, that is, a shallow feature extraction module (head), deep feature extraction module (body) and a HR image reconstruction module (tail) [18]. The input low-resolution image is denoted as $I_{LR} \in \mathbb{R}^{h \times w \times c}$, where h, w and c are the height, width and channel number of the image. A convolutional layer H_{head} is employed to extract shallow features $X_s \in \mathbb{R}^{h \times w \times c}$ from I_{LR} as:

$$X_s = H_{head}(I_{LR})$$

The deep feature extraction module usually consists of a series of stacked SR modules which depend on different SR backbones, such as the residual block in EDSR [41], residual dense block (RDB) in RDN [42], sequential residual block (SRB) in TSRN [14]. Considering our proposed weak supervision framework is a general framework that can be applied to different SR backbones, we do not depict the body in detail in Figure 2. The process is formulated as:

$$X_d = H_{body}(X_s).$$

Then, a global residual path is added to aggregate shallow features X_s and deep features X_d. Finally, the SR image I_{SR} is reconstructed as follows:

$$I_{SR} = H_{tail}(X_s + X_d),$$

where H_{tail} is the reconstruction methods, for example, pixel-shuffle operation followed by a convolutional layer.

Previous methods optimize the parameters of SR model using all of the HR images, bringing about high cost of the data-labelling process. This paper computes loss between a subset of HR-LR pairs, which is more practical in reality. Considering TSRN [14] is the most commonly used text image super-resolution baseline network, we adopt TSRN as the super-resolution network in our proposed framework as default.

3.3 | Text recognition network

This paper employs CRNN [12] as the text recognition network, which is a widely used lightweight network. Before feeding into CRNN [12], I_{SR} is resized to $H \times W$ using bicubic interpolation where H, W is set to 32, 100 following [12]. First, the ConvNet based on VGG architecture [43] extracts visual feature, denoted by $F_v \in \mathbb{R}^{1 \times \frac{W}{4} \times D}$ ($D = 512$). Then, sequential features $F_s \in \mathbb{R}^{\frac{W}{4} \times D}$ are obtained after the map-to-sequence operation on F_v. Thereafter, a two-layer bidirectional LSTM [44] with 512 hidden units conducts on the sequential features to capture contextual information from both directions. Finally, a linear layer and a softmax function are utilized, generating a per-frame probability map $P \in \mathbb{R}^{T \times C}$ where the horizontal axis indicates the sequence in left-to-right order and the vertical axis represents the categories of the alphabet set in the order of "a" to "z", "0" to "9" and blank "-". The length T and category C of alphabet set are set to 25 and 37 as default, respectively. Finally, CTC is adopted for character alignment and generating the final sequence, which will be given details in the following section.

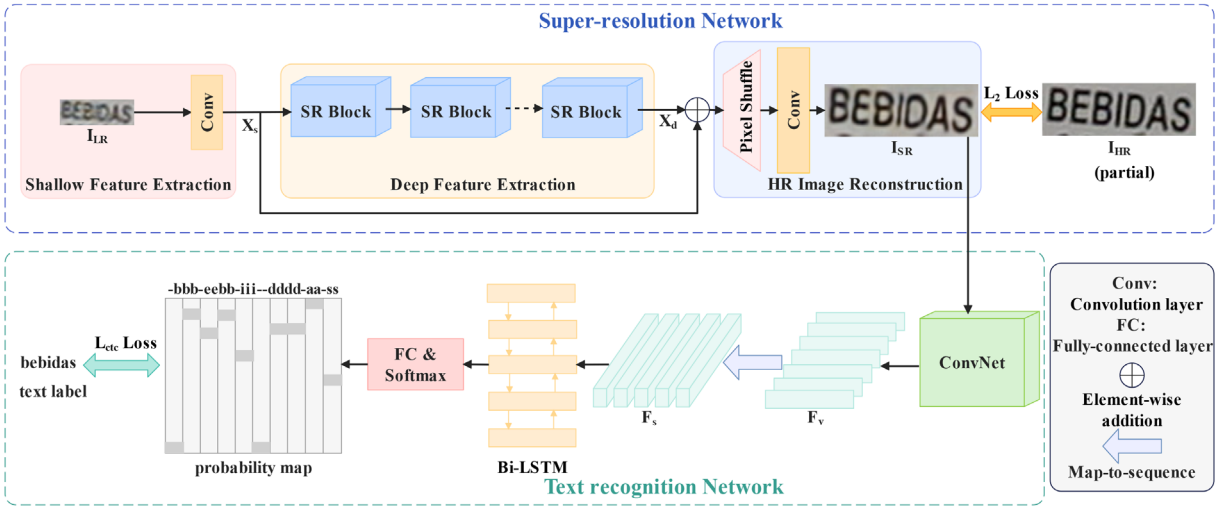

FIGURE 4 Diagram of TLWSR.

3.4 | Loss function

3.4.1 | Text recognition loss

The length of text labels is variable while the output of many text recognition networks is fixed length sequences. These methods usually require transcription layers to convert the fixed length output into a variable length prediction string, and then apply non-aligned classification loss functions. Here, the CTC loss used in CRNN [12] is adopted as the text recognition loss.

Input text images y into CRNN. The output is a sequence of T frames in length, denoted as $z = (z^1, z^2, \ldots, z^T)$, where $z^t = (z^t_1, z^t_2, \ldots, z^t_C)$ and C represents the number of character categories. Z^t_i represents the probability that the t-th frame is predicted to be the character with category index i, and the sum of probabilities of all categories $\sum_{i=1}^{C} Z^t_i = 1$. Generally, $C = 37$, including 26 English letters, 10 Arabic numerals and the empty character "-". Select a character for each frame, and the string formed is called a path π. Assuming that the output at each time is independent, the probability of the path is defined:

$$p(\pi \mid y) = \prod_{t=1}^{T} z^t_{\pi_t}, \tag{1}$$

where π_t represents the category index of the character corresponding to the path π at frame t. Then define B-transform that combines adjacent identical characters and deletes empty characters on a string. Take the following equation as an example:

$$B(-stta - t - e) = state. \tag{2}$$

The last transcription layer of CRNN is realized by B-transform on the path with the highest probability. Obviously, the B-transform is a many-to-one mapping. For a text label, there are multiple corresponding paths. All reachable paths can be listed through a dynamic programming algorithm. Given the input of CRNN y, the probability that the final output prediction is label l equals to the sum of the probabilities of all paths that can be converted into label l by B-transform, as shown in the following equation (3):

$$p(l \mid y) = \sum_{B(\pi)=l} p(\pi \mid y). \tag{3}$$

Take the negative logarithm of the probability and average it on a training batch, that is the text recognition loss, as shown in Equation (4):

$$L_{ctc} = -\frac{1}{N}\sum_{i=1}^{N} \ln\left(p(l_i \mid y_i)\right) = -\frac{1}{N}\sum_{i=1}^{N} \ln\left(p(l_i \mid G(x_i))\right), \tag{4}$$

where N represents the number of images contained in a training batch, and y_i and l_i represent the text image to be recognized and its corresponding text label, respectively. In the proposed weak supervision framework, the input of CRNN y_i is the output of the SR model $G(x_i)$, where x_i represents the LR image to be super-resolved.

3.4.2 | Overall loss

In the proposed weak supervision framework TLWSR, two types of loss functions are used, namely, the reconstruction loss L_{rec} defined by the mean square error and the text recognition loss L_{ctc} introduced in Section 3.4.1. The overall loss is calculated as :

$$\begin{aligned} L &= L_{rec} + \lambda \cdot L_{ctc} \\ &= \sum_{i=1}^{\alpha \cdot N} \left\| \hat{y}_i - G(x_i) \right\|_2 - \lambda \sum_{i=1}^{N} \ln\left(p(l_i \mid G(x_i))\right) \end{aligned} \tag{5}$$

ALGORITHM 1 TLWSR.

Input: LR training images $(x_1, x_2, \ldots x_N)$, partial HR ground-truths $(\hat{y_1}, \hat{y_2}, \ldots \hat{y}_{\alpha N})$, text labels $(l_1, l_2, \ldots l_N)$, batch size N_b, training epochs N_e

Output: θ_G, the parameters of the SR network G

```
for i = 1: N_e do
    for j = 1: N_b do
        Input LR images x to the super-resolution network G
        Input SR images y to the recognition network R
        Compute loss according to Equation (5)
        Update parameters of G and R
    end for
end for
return G
```

where N represents the number of LR images in a training batch and x_i represents the LR image. $\hat{y}_i$ and l_i represent the HR ground-truth and the text label of the LR image, respectively. Partial LR images with HR labels are used in Equation (5), where α represents the ratio of HR labels. λ represents the coefficient of text recognition loss and is used to balance the weight of two loss items. The training process of TLWSR is presented in Algorithm 1.

4 | EXPERIMENTS AND ANALYSIS

4.1 | Datasets

TextZoom The images in TextZoom [14] come from two real-world SISR datasets, namely, RealSR [30] and SRRAW [31]. The training set contains 17,367 LR-HR image pairs and corresponding text labels. The testing set can be divided into three subsets. The easy, medium and hard subset contains 1619, 1411 and 1343 LR-HR image pairs and corresponding text labels, respectively.

Scene text recognition datasets In addition to conducting experiments on TextZoom, we also verify the robustness of TLWSR on several scene text recognition datasets, including ICDAR 2013 (IC13), ICDAR 2015 (IC15), Street View Text Perspective (SVTP) and CUTE80 (CUTE). IC13 contains 1095 testing images. Most of the images are clear and some of them are under uneven illumination. IC15 consists of 1811 images and many images are blurry and rotated, which are challenging for existing recognition methods. SVTP has 649 images for evaluation, most of which are curve-shaped text. Most of samples in CUTE are curved. The test set has 288 images in all.

4.2 | Settings

All the LR images are resized to 16×64 and HR images to 32×128. The learning rate is set as 3×10^{-4} and batch size as 128. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. In the testing phase, we use a scene text recognizer ASTER [39] to evaluate the SR models. All of our models are trained on a single Nvidia RTX Titan GPU for 500 epochs.

TABLE 1 Results of TLWSR on TextZoom.

Type	HR ratio	Loss function	Recognition accuracy (%) Easy	Medium	Hard	Average
BICUBIC	-	-	64.7	42.4	31.2	47.2
Weakly supervised	0(%)	L_{ctc}	23.5	30.7	45.4	32.5
	10(%)	$L_2(10\%) + L_{ctc}$	72.0	54.6	37.0	55.6
	25(%)	$L_2(25\%) + L_{ctc}$	73.3	56.2	38.5	57.1
	50(%)	$L_2(50\%) + L_{ctc}$	**74.7**	**58.7**	**41.4**	**59.3**
Fully supervised	100(%)	L_2	74.8	55.7	39.6	57.8

4.3 | Experiments on TextZoom

4.3.1 | Results on TextZoom

As shown in Table 1, the training process does not converge without HR ground-truths, and the SR results lag far behind bicubic upsampled images. The main reason is that the prediction of the text recognition network only contains high-level semantic information, while the key to a generative task such as super-resolution is to reconstruct low-level pixel information. Therefore, it is not sufficient to only use coarse-grained text labels for supervision. For the proposed weak supervision framework TLWSR, HR labels are partially available and three HR label ratios (10%, 25%, 50%) are selected. The best results are highlighted in **bold**. For fair comparison, TSRN [14] is adopted as the SR model and we reproduce TSRN as the fully supervised baseline. As shown in Table 1, our method achieves comparable performance to the fully supervised baseline with 25% HR labels available on TextZoom. Furthermore, 50% of HR labels lead to a significant gain of 1.5% average recognition accuracy on ASTER [39].

In addition, we retrain different super-resolution backbones under our proposed weak supervision framework with 10%, 25% and 50% of HR labels used and compare with their reported results in [14]. From Table 2, we observe that our framework can achieve reasonable performance when HR labels account for 10% and 25%. Furthermore, our framework with 50% of HR labels can effectively enhance the performance of each backbone. For example, EDSR [41] under the proposed weak supervision framework boosts average accuracy by 5.9% on CRNN [12].

4.3.2 | Loss balance

We explore the choice of λ from $\{0, 0.01, 0.1, 1\}$ with 50% HR labels available and Table 3 shows the results. The recognition

TABLE 2 Comparisons of different backbone networks retrained on the proposed weak supervision framework.

			ASTER[39](%)				MORAN[40](%)				CRNN[12](%)			
Backbone	**HR ratio**	**Loss fuction**	**Easy**	**Medium**	**Hard**	**Average**	**Easy**	**Medium**	**Hard**	**Average**	**Easy**	**Medium**	**Hard**	**Average**
VDSR[45]	100%	L_2	71.7	43.5	34.0	51.0	62.3	42.5	30.5	46.1	41.2	25.6	23.3	30.7
	10%	$L_2(10\%) + L_{ctc}$	69.4	47.3	34.6	51.6	61.2	41.8	30.9	45.6	41.0	25.2	23.0	30.4
	25%	$L_2(25\%) + L_{ctc}$	70.1	48.4	34.7	52.2	62.1	42.2	31.2	46.2	41.5	26.2	23.5	31.0
	50%	$L_2(50\%) + L_{ctc}$	69.8	48.6	35.0	52.3	63.1	42.5	31.4	46.7	43.0	26.4	23.8	31.7
SRResNet[46]	100%	L_2	69.6	47.6	34.3	51.3	60.7	42.9	32.6	46.3	39.7	27.6	22.7	30.6
	10%	$L_2(10\%) + L_{ctc}$	69.7	49.6	34.3	52.3	62.3	43.4	30.7	46.5	45.3	27.7	22.6	32.6
	25%	$L_2(25\%) + L_{ctc}$	69.9	50.0	35.2	52.8	63.9	43.5	31.6	47.4	46.2	31.2	25.5	35.0
	50%	$L_2(50\%) + L_{ctc}$	70.1	51.0	35.5	53.3	65.4	43.9	32.8	48.5	46.5	31.6	26.2	35.5
EDSR[41]	100%	L_1	70.4	49.1	34.2	52.4	64.2	44.9	31.3	47.9	42.5	29.1	23.2	32.2
	10%	$L_2(10\%) + L_{ctc}$	70.8	50.6	35.4	53.4	67.5	45.1	33.2	49.7	48.7	31.8	25.6	36.2
	25%	$L_2(25\%) + L_{ctc}$	69.8	48.6	35.0	52.3	63.1	42.5	31.4	46.7	43.0	26.4	23.8	31.7
	50%	$L_1(50\%) + L_{ctc}$	72.1	49.9	36.0	53.9	68.5	45.2	33.4	50.2	52.3	33.3	27.5	38.6
RDN[42]	100%	L_1	70.0	47.0	34.0	51.5	61.7	42.0	31.6	46.1	41.6	24.4	23.5	30.5
	10%	$L_2(10\%) + L_{ctc}$	68.8	47.0	32.4	50.6	60.9	41.9	30.7	45.5	40.2	24.2	23.2	29.8
	25%	$L_2(25\%) + L_{ctc}$	69.3	47.5	34.3	51.6	63.2	41.8	31.2	46.5	42.2	25.0	24.1	31.1
	50%	$L_1(50\%) + L_{ctc}$	70.4	49.3	35.0	52.7	64.7	42.2	31.3	47.2	45.9	29.1	24.6	33.9
TSRN[14]	100%	L_2	**75.1**	56.3	40.1	58.3	70.1	**55.3**	37.9	55.4	52.5	38.2	31.4	41.4
	10%	$L_2(10\%) + L_{ctc}$	72.0	54.6	37.0	55.6	68.1	49.8	36.7	52.3	49.9	38.6	30.7	40.4
	25%	$L_2(25\%) + L_{ctc}$	73.3	56.2	38.5	57.1	69.2	51.9	37.9	54.0	51.7	39.8	32.8	42.1
	50%	$L_2(50\%) + L_{ctc}$	74.7	**58.7**	**41.4**	**59.3**	**71.5**	53.9	**39.4**	**56.0**	**52.8**	**40.4**	**33.2**	**42.8**

TABLE 3 Ablation study on the choices of λ.

	Accuracy (%) of ASTER[39]			
λ	**Easy**	**Medium**	**Hard**	**Average**
0	69.2	52.0	36.5	53.6
0.01	72.0	52.9	37.5	55.2
0.1	**74.7**	**58.7**	**41.4**	**59.3**
1	68.3	50.1	35.7	52.4

accuracy reaches best when λ is set to 0.1. Therefore, λ is set to 0.1 in 4.3.

4.3.3 | Visualization on TextZoom

Figure 5 compares the visual super-resolution results of TLWSR (50% HR labels available) with the fully supervised baseline TSRN on the test dataset. It shows that our method can effectively improve the visual quality of SR images with more readable characters. The above part of Figure 5 indicates that more details in text region are recovered by TLWSR (e.g. **e**, **b**). More visual comparisons are shown in the below part of Figure 5, which can verify the ability of TLWSR. (e.g. **a**, **n** in the first column, second column of the below part are better recovered using our method).

4.3.4 | Generalization to scene text recognition datasets

We evaluate the robustness of TLWSR using 50% HR labels as a preprocessor on low visual quality scene text images, which are chosen from IC13, IC15,SVTP and CUTE. We only pick 546 low-resolution images with resolution lower than 16×64 from these datasets. Furthermore, some manual degradation is added to these picked images to increase the difficulty of super-resolution. First, we blur the original images. Specifically, a 5×5 sized Gaussian kernel is used to convolve these images with $\sigma = 1$. Then, Gaussian noise with $\sigma = 50$ is added to the blurred images. Examples of the original images and corresponding degraded images are shown in Figure 6. Table 4 shows the experimental results. Compared with preprocessing the original images using fully supervised baseline TSRN [14], our weakly supervised framework boosts the recognition accuracy of ASTER [39] by 3.3%, MORAN [40] by 2.1%, CRNN [12] by 1.7%. After blurring the images and adding Gaussian noise, our method improves the performances of all three recognizers as well. It demonstrates that the proposed method can be well generalized to recover low-quality text images in other datasets.

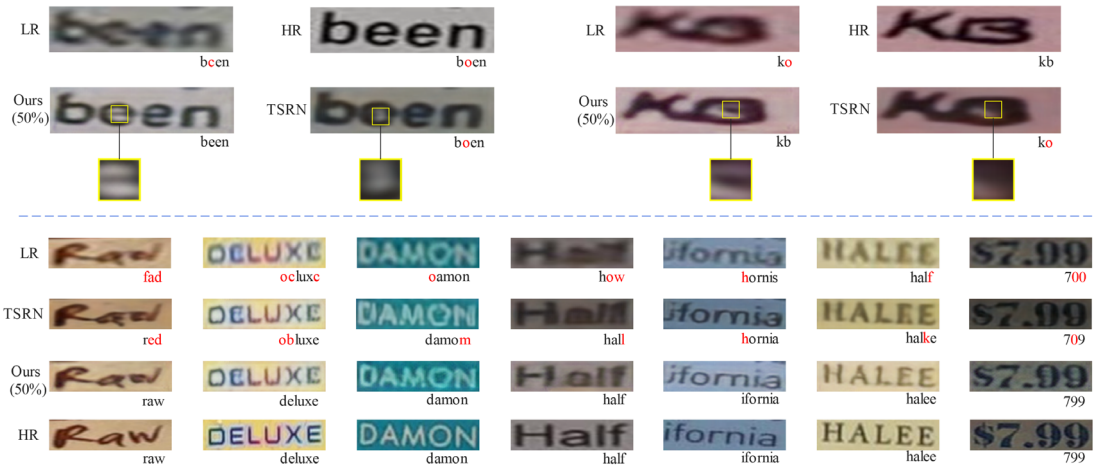

FIGURE 5 Comparisons of super-resolution visualization. Characters in red are missing or wrong. HR, high resolution; LR, low resolution; TSRN, text super-resolution network.

FIGURE 6 Examples of images after degradation on scene text recognition datasets.

TABLE 4 Results on scene text recognition benchmarks.

	Method	HR ratio	Loss function	ASTER [39] (%)	MORAN [40] (%)	CRNN [12] (%)
Original	Bicubic	-	-	69.4	65.4	53.3
	TSRN	100%	L_2	70.7	69.6	57.8
	TSRN	50%	$L_2(50\%) + L_{ctc}$	**74.0**	**71.7**	**59.5**
Blur	Bicubic	-	-	43.4	43.5	31.5
	TSRN	100%	L_2	51.3	47.9	37.9
	TSRN	50%	$L_2(50\%) + L_{ctc}$	**55.0**	**49.6**	**41.1**
Blur+Noise	Bicubic	-	-	37.9	35.2	24.2
	TSRN	100%	L_2	48.7	42.1	32.7
	TSRN	50%	$L_2(50\%) + L_{ctc}$	**50.9**	**44.3**	**34.1**

4.3.5 | Visualization on scene text recognition benchmarks

The super-resolution results after blurring and adding Gaussian noise are shown in Figures 7 and 8, respectively. After blurring the images, some characters are mixed together (e.g. **i** and **l** in the second sample "textiles"). Noise further degrades the visual quality of blurry images, making recognition more challenging. After processing by our proposed weak supervision framework, these blurry and noisy images can achieve better visual quality compared with the fully supervised baseline (e.g. **f** and **i** in the first sample "office").

4.3.6 | Failure Cases

Several failure cases are visualized in Figure 9. Long and oblique texts bring difficulties to our method. We can observe that

FIGURE 7 Super-resolution results of blurred images on scene text recognition datasets. Characters in red are missing or wrong.TSRN, text super-resolution network.

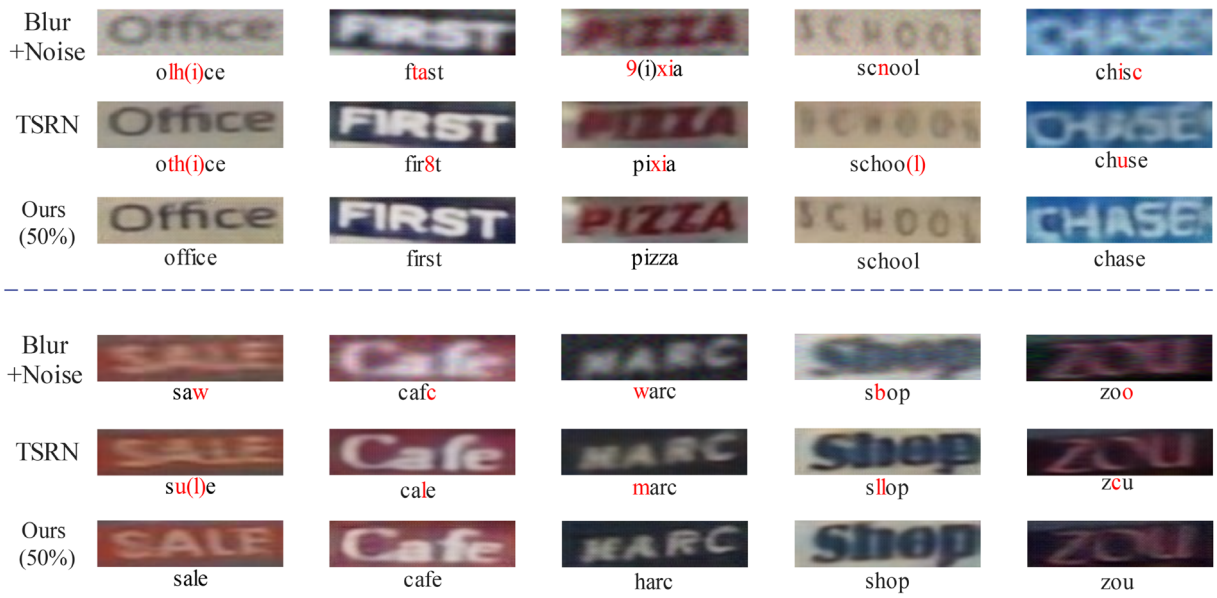

FIGURE 8 Super-resolution results of blurred and noisy images on scene text recognition datasets. Characters in red are missing or wrong. TSRN, text super-resolution network.

the super-resolved images suffer from blurry characters leading to wrong text recognition. We look forward to addressing the problem in the future work.

5 | CONCLUSION

This paper explores weakly supervised super-resolution of real-world text images and designs a novel framework called TLWSR, aiming to reduce the dependency on HR labels which are hard and costly to collect. Particularly, the super-resolution network is combined with a text recognition network. CTC loss is utilized to facilitate the process of super-resolution, and thus concentrate more on the text regions. Extensive experiments show that TLWSR reconstructs the blurry pixels better and outperforms fully supervised baseline method in boosting recognition performance of LR images. Moreover, TLWSR can be well generalized to low-quality images in multiple public text recognition datasets, which further verifies the effectiveness and generalization of the proposed frame-

FIGURE 9 Visualization of several failure cases. HR, high resolution; LR, low resolution.

work. Compared with previous STISR methods, TLWSR is more practical and potential in realistic scenarios. This paper conducts initial research on weak supervision framework in STISR. As a new problem to be explored, weakly supervised scene text image super-resolution deserves more attention in the future.

AUTHOR CONTRIBUTIONS

Qin Shi: Conceptualization, formal analysis, visualization, writing - original draft, writing - review and editing. Yu Zhu: Conceptualization, funding acquisition, project administration, supervision, writing - review and editing. Chuantao Fang: Conceptualization, methodology, writing - original draft. Dawei Yang: Funding acquisition, resources, supervision.

ACKNOWLEDGEMENTS

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grants 20DZ22544000, and 20DZ2261200; and Fujian Province Department of Science and Technology (2022D014).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The training data presented in the study are openly available at: https://github.com/JasonBoy1/TextZoom

ORCID

Yu Zhu https://orcid.org/0000-0003-1535-6520

REFERENCES

1. Karaoglu, S., Tao, R., Gevers, T., Smeulders, A.W.: Words matter: scene text for image classification and retrieval. IEEE transactions on multimedia 19(5), 1063–1076 (2016)
2. Silva, S.M., Jung, C.R.: License plate detection and recognition in unconstrained scenarios. In: In Proceedings of the European Conference on Computer Vision (ECCV), pp. 580–596. Springer, Cham (2018)
3. Ren, H., Pan, M., Li, Y., Zhou, X., Luo, J.: St-SiameseNet: spatio-temporal siamese networks for human mobility signature identification. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1306–1315. ACM, New York (2020)
4. Hou, J.-B., Zhu, X., Liu, C., Yang, C., Wu, L.-H., Wang, H., Yin, X.-C.: Detecting text in scene and traffic guide panels with attention anchor mechanism. IEEE Trans. Intell. Transp. Syst. 22(11), 6890–6899 (2020)
5. Qiao, Z., Zhou, Y., Wei, J., Wang, W., Zhang, Y., Jiang, N., Wang, H., Wang, W.: PIMNet: a parallel, iterative and mimicking network for scene text recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2046–2055. ACM, New York (2021)
6. Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: International Conference on Document Analysis and Recognition, pp. 319–334. Springer, Cham (2021)
7. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.-G.: SVTR: Scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159 (2022)
8. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Deep structured output learning for unconstrained text recognition. arXiv preprint arXiv:1412.5903 (2014)
9. Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: Seed: semantics enhanced encoder-decoder framework for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13528–13537. IEEE, Piscataway, NJ (2020)
10. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12113–12122. IEEE, Piscataway, NJ (2020)
11. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: 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, pp. 7098–7107. IEEE, Piscataway, NJ (2021)
12. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. 39(11), 2298–2304 (2016)
13. Nguyen, N., Nguyen, T., Tran, V., Tran, M.-T., Ngo, T.D., Nguyen, T.H., Hoai, M.: Dictionary-guided scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7383–7392. IEEE, Piscataway, NJ (2021)
14. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. In: European Conference on Computer Vision, pp. 650–666. Springer, Cham (2020)
15. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision, pp. 184–199. Springer, Cham (2014)
16. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In:

Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301. Springer, Cham (2018)
17. Ji, J., Zhong, B., Ma, K.-K.: Single image super-resolution using asynchronous multi-scale network. IEEE Signal Process Lett. 28, 1823–1827 (2021)
18. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844. IEEE, Piscataway, NJ (2021)
19. Zou, W., Ye, T., Zheng, W., Zhang, Y., Chen, L., Wu, Y.: Self-calibrated efficient transformer for lightweightsuper-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition, pp. 930–939. IEEE, Piscataway, NJ (2022)
20. Zhu, H., Tang, H., Hu, Y., Tao, H., Xie, C.: Lightweight single image super-resolution with selective channel processing network. Sensors 22(15), 5586 (2022)
21. Blanco-Medina, P., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R., Jáñez-Martino, F., Bonnici, A.: Rectification and super-resolution enhancements for forensic textrecognition. Sensors 20(20), 5850 (2020)
22. Wang, W., Xie, E., Sun, P., Wang, W., Tian, L., Shen, C., Luo, P.: Textsr: Content-aware text super-resolution guided by recognition. arXiv preprint arXiv:1909.07113 (2019)
23. Mou, Y., Tan, L., Yang, H., Chen, J., Liu, L., Yan, R., Huang, Y.: Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In: European Conference on Computer Vision, pp. 158–174. Springer, Cham (2020)
24. Wang, Y., Su, F., Qian, Y.: Text-attentional conditional generative adversarial network for super-resolution of text images. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1024–1029. IEEE, Piscataway, NJ (2019)
25. Fang, C., Zhu, Y., Liao, L., Ling, X.: TSRGAN: real-world text image super-resolution based on adversarial learning and triplet attention. Neurocomputing 455, 88–96 (2021)
26. Chen, J., Li, B., Xue, X.: Scene text telescope: text-focused scene image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition, pp. 12026–12035. IEEE, Piscataway, NJ (2021)
27. Zhao, C., Feng, S., Zhao, B.N., Ding, Z., Wu, J., Shen, F., Shen, H.T.: Scene text image super-resolution via parallelly contextual attention network. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2908–2917. ACM, New York (2021)
28. Ma, J., Guo, S., Zhang, L.: Text prior guided scene text image super-resolution. arXiv preprint arXiv:2106.15368 (2021)
29. Chen, J., Yu, H., Ma, J., Li, B., Xue, X.: Text gestalt: stroke-aware scene text image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 285–293. AAAI, Washington, DC (2022)
30. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: a new benchmark and a new model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3086–3095. IEEE, Piscataway, NJ (2019)
31. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3762–3770. IEEE, Piscataway, NJ (2019)
32. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 185–200. Springer, New York (2018)
33. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710. IEEE, Piscataway, NJ (2018)
34. Chen, H., Dong, L., Yang, H., He, X., Zhu, C.: Unsupervised real-world image super-resolution via dual synthetic-to-realistic and realistic-to-synthetic translations. IEEE Signal Process Lett. (2022)
35. Zhou, Z.-H.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5(1), 44–53 (2018)
36. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision, pp. 1457–1464. IEEE, Piscataway, NJ (2011)
37. Wang, K., Belongie, S.: Word spotting in the wild. In: European Conference on Computer Vision, pp. 591–604. Springer, New York (2010)
38. Su, B., Lu, S.: Accurate recognition of words in scenes without character segmentation using recurrent neural network. Pattern Recognit. 63, 397–405 (2017)
39. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. 41(9), 2035–2048 (2018)
40. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognit. 90, 109–118 (2019)
41. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144. IEEE, Piscataway, NJ (2017)
42. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481. IEEE, Piscataway, NJ (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 855–868 (2008)
45. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654. IEEE, Piscataway, NJ (2016)
46. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690. IEEE, Piscataway, NJ (2017)

How to cite this article: Shi, Q., Zhu, Y., Fang, C., Yang, D.: TLWSR: Weakly supervised real-world scene text image super-resolution using text label. IET Image Process. 17, 2780–2790 (2023).
https://doi.org/10.1049/ipr2.12827