DCDM: Diffusion-Conditioned-Diffusion Model for Scene Text Image Super-Resolution

Shrey Singh[1], Prateek Keserwani[1], Masakazu Iwamura[2], and Partha Pratim Roy[1]

[1] Indian Institute of Technology, Roorkee, Uttarakhand-247667, India
{ssingh19, pkeserwani, partha}@cs.iitr.ac.in
[2] Osaka Metropolitan University, Sakai, Osaka 599-8531, Japan
masa.i@omu.ac.jp

Abstract. Severe blurring of scene text images, resulting in the loss of critical strokes and textual information, has a profound impact on text readability and recognizability. Therefore, scene text image super-resolution, aiming to enhance text resolution and legibility in low-resolution images, is a crucial task. In this paper, we introduce a novel generative model for scene text super-resolution called *diffusion-conditioned-diffusion model* (DCDM). The model is designed to learn the distribution of high-resolution images via two conditions: 1) the low-resolution image and 2) the character-level text embedding generated by a latent diffusion text model. The latent diffusion text module is specifically designed to generate character-level text embedding space from the latent space of low-resolution images. Additionally, the character-level CLIP module has been used to align the high-resolution character-level text embeddings with low-resolution embeddings. This ensures visual alignment with the semantics of scene text image characters. Our experiments on the TextZoom and Real-CE datasets demonstrate the superiority of the proposed method to state-of-the-art methods. The source codes and other resources will be available through the project page: https://github.com/shreygithub/DCDM.

Keywords: Scene Text Image Super-Resolution · Diffusion-Conditioned-Diffusion Model · Character-Level CLIP

1 Introduction

Scene text understanding has remained an important area of research in computer vision for over a decade. This field encompasses various tasks, including scene text recognition [33], scene text retrieval [37], and scene text visual question answering [1]. A major challenge in these tasks is image degradation, particularly due to low resolution. Additionally, these texts are optically degraded in the form of blur and noise, which makes the reading the text difficult. Improving

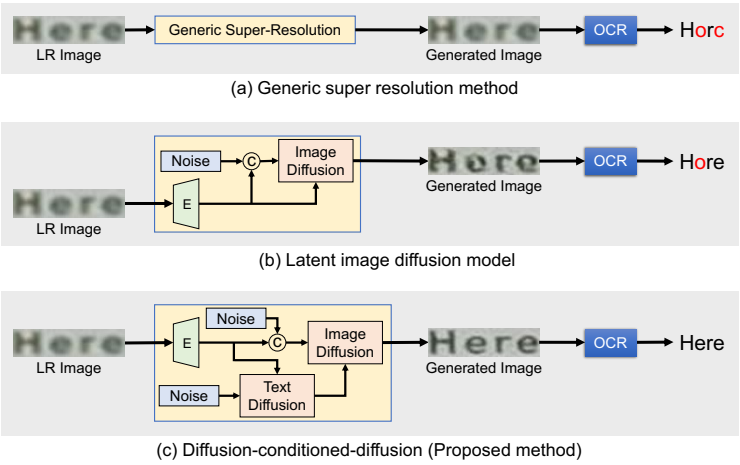

Fig. 1: Quality enhancement of low-resolution (LR) image using different methods and their optical character recognition (OCR) results. (a) Generated image by a generic super resolution method (SRCNN). (b) Generated image by a latent image diffusion model. (c) Generated image by the proposed method (diffusion-conditioned-diffusion) that combines the diffusion method based text prior generation and used as a condition for image diffusion model for scene text image super resolution.

the quality of the text strokes along with removing the noise and blur in the image is termed as scene text image super-resolution (STISR). A good solution is needed to solve various text understanding tasks in a low-resolution constraint.

In past, various attempts have been made to enhance scene text resolution along with removing blur and noise for reading the text efficiently by off-the-shelf OCR. The resolution of scene text images was improved by applying super-resolution (SR) techniques (illustrated in Fig. 1(a)), such as [42]. Afterward, textual properties were used to enhance the super-resolution results [40]. A significant limitation of general-purpose image super-resolution techniques is their inability to emphasize text strokes compared to background pixels, which is needed to improve the visual quality of scene text.

For emphasizing more on text than background pixels, various approaches have been explored. They can be broadly grouped into the following categories: text recognition loss [19,40], sequential information [3,39], text mask [49], stroke-aware loss [20,39], and text prior [16,18,48]. Their common feature is to utilize some additional information to enhance the low-resolution (LR) text images. Among those, the text-prior-based methods [16,18,48] uplifted most of the scene text super-resolution performance. In these methods, a text recognizer is applied to the LR image during inference and the recognition result is used as a noisy text prior with recognition errors. It can hinder the performance of the scene text image super-resolution.

Following the impact of diffusion models on generic image super-resolution [31, 38], an image diffusion-based method trained on synthesized data [23] has been

recently proposed, which is based on the method illustrated on Fig. 1(b). This method also relies on a text prior incorporating a text recognizer. However, a fundamental question arises; that is, *"Is a text recognizer necessary for STISR in inference?"* To seek an answer to this question, this work explores the possibility of utilizing latent text diffusion models as a generator of text prior to the succeeding image diffusion model (as illustrated in Fig. 1(c)). In other words, we use two diffusion models at once: the latent text diffusion and the image diffusion models. We call the proposed method *diffusion-conditioned-diffusion model* (DCDM). Like a text recognizer generates a text prior in the text-prior-based methods, the latent text diffusion model takes an LR image and outputs the text prior that conditions the succeeding image diffusion model. A character-level CLIP (CL-CLIP) model is used to train the latent text diffusion model. This diffusion-conditioned-diffusion method is good for super-resolution and removing blur impact due to optical degradation while grabbing the text image.

The most similar work to the proposed method is StableCascade (SC) [25], our concurrent work, which introduced two-stage latent diffusion models (LDMs) for generating images and image embeddings. The key difference between our proposed method and SC lies in introducing an image-to-text LDM, which generates a text before the succeeding text-to-image LDM.

The major contributions of the presented work are as follows:

1. A diffusion-conditioned-diffusion model has been proposed which has utilized the text characteristics for the image super-resolution for text.
2. We introduce the latent text diffusion model to generate character-level text embedding from a given low-resolution latent space. It incorporated a character-level CLIP model (called CL-CLIP) to obtain linguistic and visual connections.
3. Through detailed experiments, we demonstrate the impact of the diffusion-conditioned-diffusion model on the STISR.

2 Related Work

2.1 Single Image Super-Resolution (SISR)

The SISR is a task for estimating a high resolution (HR) image from its corresponding LR image. The ill-posed nature of the SISR problem adds more challenges to the problem. In the past, the prior information is used in the form of a distribution/energy function to aggregate the constraints of the SR image. Adaptive high-dimensional nonlocal total variation-based adaptive geometric duality prior [29] and sparse regression and statistical image priors [12] are some important works on reconstruction-based techniques. These hand-crafted-based methods work well in reducing the virtual artifacts but are still not enough to fulfill the requirements of the SISR. In recent years, convolutional neural networks (CNNs) have been frequently used and accomplish leading performance for the SISR. The SRCNN pioneers CNN to learn the mapping function between LR and HR images. In later works, the CNN architectures are designed deeper

and with more sophistication to elevate the performance of SISR, for example, Laplacian pyramid [13], dense connections based [35], residual block, and channel attention mechanism [46]. In recent work, prior information has been utilized to boost the performance of CNN architectures for SISR [9].

2.2 Scene Text Image Super-Resolution (STISR)

The general-purpose SISR focuses on natural scene images. The STISR is a special case of SISR. Unlike general-purpose SISR, the objective of STISR is not only to scale up the resolution of the text image but also to focus on improving text readability. The preliminary methods for STISR adopted CNN architectures from general-purpose SISR and directly attempted to extend it for the text images. For ICDAR 2015 competition [26], Dong *et al.* [7] extended the SRCNN [6] to text images to achieve the best result in the competition. In [24], three SR frameworks are designed to accomplish SR on binary document images. The performances of these initial methods are not good on the text images because these methods directly utilize the generic SR frameworks and ignore text-centric properties such as word or character-level layout details. PlugNet [19] utilized a pluggable SR unit for a designing multi-task framework to perform SR and recognition hand in hand.

Wang *et al.* [39] built a real-world dataset named TextZoom for STISR images. They also proposed a text super-resolution network (TSRN) to address the STISR problem on real-world text images. The sequential residual block (SRB) is the main building block of TSRN. The sequential notion of SRB is covered by using the horizontal and vertical bidirectional long short-term memory (BLSTM) blocks. Apart from BLSTM blocks (used to capture sequential information such as text), the TSRN is not doing much for text-related features. [3] proposed transformer-based super-resolution network (TBSRN) uses a self-attention module to process sequential information. The perceptual text losses as position and content awareness on a character level are applied to help the text recognition. In [16], embed the text prior (guided by HR) into the STISR model for better reconstruction of text in the HR scene text images. The methods for STISR discussed above embed text-prior information to the SR module to help reconstruct the text in HR images. The embedding of only text prior to the SR module is insufficient for the STISR. The prior low-resolution features need to be boosted with the guides of ground truth in training. We aim to design a module to boost the low-resolution features to support the reconstruction of the HR text image and achieve better recognition in this paper.

3 Proposed Method

As shown in Fig. 2, we introduce a novel Diffusion Conditional Diffusion Model (DCDM) consisting of two specialized diffusion-based modules. As shown in the top part of Fig. 2, the first module, the Latent Text Diffusion Module, is designed to learn the joint distribution between low-resolution images and text

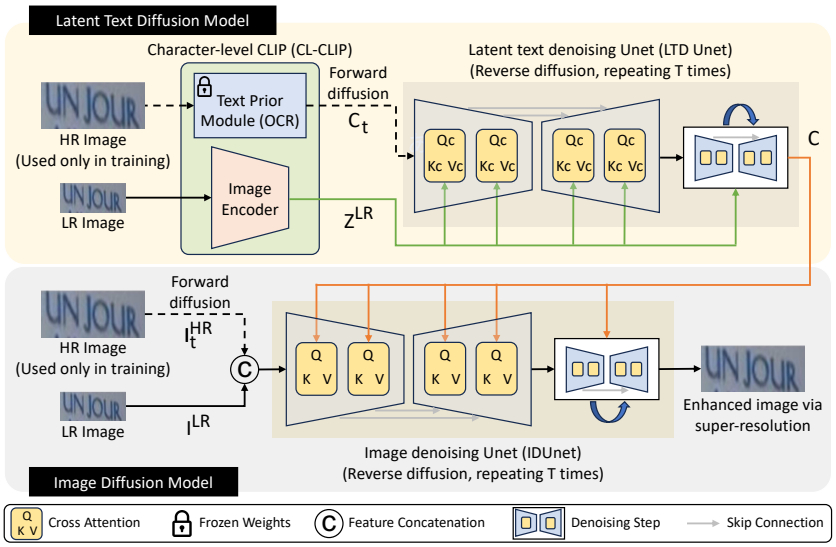

Fig. 2: The detailed diagram of the proposed method of diffusion-condition-diffusion. It consists of two forward passes, one for the latent text diffusion model and the other for the latent image diffusion model. The latent text diffusion model consists of a character-level CLIP model for alignment between characters and the structural part of an image. The latent text diffusion model acts as a conditioning module for the latent image diffusion model. The dotted line flow is only used during a training phase.

priors. This module excels in discerning complex dependencies between latent images and textual information, providing a comprehensive understanding of their interplay. Complementing this, as shown in the bottom part of Fig. 2, the second module, the Image Diffusion Model, is strategically designed for hybrid conditioning, considering both textual elements (text prior) and visual components (low-resolution images). This dual consideration allows the model to capture synergistic effects of text and images in a unified manner, enhancing its ability to discern nuanced patterns and relationships within the data.

3.1 Image Diffusion Model

In our proposed work, we introduce a novel generative model, rooted in the principles of the Diffusion Model (DM) [21], which is shown in the bottom part of Fig. 2. This model's primary objective is to acquire a high-resolution image, denoted as I^{HR}, while considering specific conditioning information denoted as I^{LR} and additional conditions represented as C. I^{HR} is used only in the training time as the ground truth. Our approach fundamentally revolves around the concept of shaping the distribution of the high-resolution images, I^{HR}. This is achieved through a gradual denoising process, effectively mirroring the reverse dynamics of a predefined fixed-length Markov Chain. Let us denote the predefined fixed-length as T. Conceptually, our model can be envisioned as an ensemble of

denoising autoencoders, each with a distinct role in the diffusion process. These denoising autoencoders, referred to as Image Denoising UNet (IDUnet), are sequentially arranged and represented as $\epsilon_\theta(I_t^{\text{HR}}, I^{\text{LR}}, C, t)$, with t iterating from 1 to T. I_t^{HR} is the intermediate denoised HR image at the t-th step. I_1^{HR} is full of noise and I_T^{HR} is expected to be close enough to I^{HR}. The primary objective function under consideration is formulated as

$$\mathcal{L}_{I^{\text{LR}} \to I^{\text{HR}}} = \mathbb{E}_{I^{\text{HR}}, I^{\text{LR}}, C, \epsilon \sim \mathcal{N}(0,1), t} \left[|\epsilon - \epsilon_\theta(I_t^{\text{HR}}, I^{\text{LR}}, C, t)|_2^2 \right], \tag{1}$$

where $t = [1, 2, \ldots, T]$ is a uniform distribution function. The central element of this objective function is the UNet denoising function, denoted as ϵ_θ. This function is intricately conditioned by multiple components, encompassing I^{HR} and I^{LR} signifying image information, C representing text embedding (see Sec. 3.2), and ϵ following a standard normal distribution, $\mathcal{N}(0, 1)$ [31]. Our objective function aims to optimize the alignment between image spaces and the denoising function while considering diverse conditioning elements, all within a fixed forward process.

3.2 Conditioning mechanisms

Our approach significantly extends the framework of DMs by adeptly accommodating and flexibly adapting to the specific conditions encapsulated within C. These conditions, represented as a hybrid combination of text embeddings and I^{LR}, actively guide the denoising process. They play a pivotal role in shaping the contours of the high-resolution image space, I^{HR}, ensuring that it aligns seamlessly with the unique contextual intricacies encompassed by this hybrid conditioning. The conditioning strategy within our model adopts a hybrid approach. Specifically, I_t^{HR} undergoes conditioning through concatenation with I^{LR}. This resultant concatenated output is further conditioned through text embeddings, denoted as $C \in \mathbb{R}^{m \times d_\tau}$. It is subsequently mapped to the intermediate layers of the UNet through the use of a cross-attention layer [36]. It is denoted by $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$, where Q, K, and V are the query, key, and value matrices, respectively. They are respectively given in the forms of $Q = W_Q^{(i)} \cdot \varphi_i([I_t^{\text{HR}}, I^{\text{LR}}])$, $K = W_K^{(i)} \cdot C$, and $V = W_V^{(i)} \cdot C$, where $W_V^{(i)} \in \mathbb{R}^{d \times d_i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$, and $W_K^{(i)} \in \mathbb{R}^{d \times d_l}$ represent learnable projection matrices [11, 30, 36], and $\varphi_i([I_t^{\text{HR}}, I^{\text{LR}}]) \in \mathbb{R}^{N \times d_i}$ signifies a (flattened) intermediate representation within the UNet, implemented through ϵ_θ. This formulation illustrates how the attention mechanism adapts to the context: Q is derived from the concatenation of I_t^{HR} and I^{LR}, K is conditioned on text embeddings C, and V is also influenced by C.

3.3 Latent Text Diffusion Model

The Latent Text Diffusion Model, shown in the top part of Fig. 2, is an LDM-based model [30] tailored for generating text embeddings while conditioned on

the low-resolution latent space Z^{LR}. Serving as a critical component within this framework, it capitalizes on the intrinsic capabilities of LDMs to synthesize text embeddings that align with and enhance the information embedded in the low-resolution latent space. This module's pivotal role is manifested in its seamless integration of textual data into the generative process. This integration empowers the model to yield outputs of heightened quality, enriched with text-related content intricately linked to the specified Z^{LR}. The denoising autoencoders, referred to as Latent text denoising Unet (LTD Unet), are sequentially arranged and represented as $\epsilon_\theta(C_t, Z^{\text{LR}}, t)$, with t iterating from 1 to T. Similar to I_t^{HR}, C_t is the intermediate denoised text prior at the t-th step. The primary objective function, denoted as $\mathcal{L}_{Z^{\text{LR}} \to C}$, is expressed as

$$\mathcal{L}_{Z^{\text{LR}} \to C} = \mathbb{E}_{C_t, Z^{\text{LR}}, \epsilon \sim \mathcal{N}(0,1), t} \left[|\epsilon - \epsilon_\theta(C_t, Z^{\text{LR}}, t)|_2^2 \right], \tag{2}$$

where $\mathcal{L}_{Z^{\text{LR}} \to C}$ serves as the primary loss function. The random variable ϵ follows a normal distribution, $\mathcal{N}(0, 1)$, while t obeys a uniform distribution from 1 to T.

Latent Text Denoising UNet: At the core of this function lies the Latent text denoising UNet (LTD UNet), denoted as $\epsilon_\theta(C_t, Z^{\text{LR}}, t)$. This objective function plays a pivotal role in training, compelling the LTD UNet to generate ϵ values closely aligned with the desired values. This enhancement, in turn, bolsters the model's capacity to denoise and refine latent text representations effectively. Additionally, low-resolution latent space $Z^{\text{LR}} \in \mathbb{R}^{N \times d}$, are incorporated into UNet's intermediate layers through a cross-attention layer. It is defined by $\text{Attention}(Q_c, K_c, V_c) = \text{softmax}\left(\frac{Q_c K_c^T}{\sqrt{d}}\right) V_c$, where $Q_c = W_{Q_c}^{(i)} \cdot \varphi_i(C_t, t)$, $K_c = W_{K_c}^{(i)} \cdot Z^{\text{LR}}$, and $V_c = W_{V_c}^{(i)} \cdot Z^{\text{LR}}$. Here, the matrices $W_{V_c}^{(i)}$, $W_{Q_c}^{(i)}$, and $W_{K_c}^{(i)}$ are learnable projection matrices, and the matrix $\varphi_i(C_t, t)$ signifies an intermediate UNet representation.

Character-level CLIP: During the training phase, addressing the challenge posed by the absence of ground truth text embedding or text tokens in the Scene Text Super-Resolution problem becomes essential, given that only pairs of HR and LR images are provided. To tackle this challenge, we leverage a text prior [16, 18] by passing the HR image through it, generating the necessary text embeddings.

To align HR text embeddings with LR images, we employ an unsupervised visual encoder Z^{LR}, guided by a contrastive loss that encourages similarity between text and images [27]. Additionally, we apply multi-view projection techniques to ensure a meaningful encoding of LR images while incorporating textual information derived from HR images through the text prior. This alignment procedure holds significance for latent text diffusion and facilitates various tasks [28]. It effectively bridges the gap between the textual information extracted from HR images and the encoded representations of LR images.

4 Experiments

4.1 Implementation Details

The complete work is based on the training of three modules namely, CL-CLIP, text denoising UNet, and image denoising UNet modules.

Character-Level CLIP (CL-CLIP) In the CL-CLIP model, the text embedding is generated by a pre-trained CRNN [32], serving as the text prior. This CRNN employs 37 tokens with a maximum sequence length of 26 tokens and has an embedding dimension of 512. For the visual encoder, a Vision Transformer (ViT) architecture, as proposed by Dosovitskiy *et al.* [8], is utilized. This ViT-based architecture employs eight heads with a patch size of 16 and has an embedding dimension of 512.

The model architecture aligns with prior works, incorporating established techniques and architectures to achieve effective performance. Furthermore, an AdamW optimizer is initialized with a learning rate of $1e-5$ and betas (0.9, 0.999) to manage gradient moment decay. AdamW combines the advantages of Adam optimization with L2 weight decay, enhancing regularization by encouraging smaller model weights and preventing overfitting. The remaining hyperparameters follow those outlined in the work by Radford *et al.* [30].

Latent text denoising UNet The structure of Latent text denoising UNet is the same as the LDM [30]. It comprises of encoder and decoder paths, featuring residual blocks with self-attention layers. Set the model channels to 128 and head channels to eight to control the network's capacity. In the encoder path, arrange a series of residual blocks with self-attention to capture complex features. Downsample the spatial resolution based on specified attention resolutions ($4\times$, $2\times$, $1\times$). The decoder path should symmetrically decode the data by upsampling. Ensure that both the encoder and decoder paths consist of two residual blocks for feature transformation. Introduce a channel multiplier that initiates with $1\times$ channels and progressively scales up by factors of one, two, and four. This ensures adaptability to feature complexity throughout the architecture. Importantly, apply residual blocks for both upscaling and downscaling operations to maintain feature consistency. For the loss function, employ Mean Squared Error (MSE) to evaluate pixel-wise differences, measuring the quality of denoised images. To optimize the model, select the Adam optimizer, which is known for its efficient convergence properties. As for the LDM, configure them in a similar fashion, adhering to specific parameters outlined in the LDM paper [30]. These parameters include setting a base learning rate of 5×10^{-06}, linear_start at 0.0015, linear_end at 0.025, and conducting training for a total of 1,000 timesteps. Additionally, use the DDIM Sampler [34] and run it for 250 timesteps.

Image denoising UNet The Image Denoising UNet structure closely resembles DDPM [10]. Similar to the Latent Denoising UNet, it incorporates downscaling, bottom, and upscaling blocks. The introduction of a channel multiplier, starting with 1x channels and progressively scaling up by factors of one, two, and four, is consistent. Cross attention is applied at multiple scales, and after residual blocks, self-attention layers are employed. In the bottom layer, the four sets of self and attention layers are applied alternatively. The loss function employed is Mean Squared Error (MSE), evaluating pixel-wise differences to measure the quality of denoised images. For optimization, the Adam optimizer is chosen for its efficient convergence properties.

Following a similar configuration to DDPM, specific parameters outlined in the DDPM paper [10] are adopted. This includes setting a base learning rate of 5×10^{-05}, linear_start at 10^{-4}, linear_end at 0.02, and conducting training for a total of $1,000$ timesteps. Additionally, the DDIM Sampler [34] is utilized, running for 250 timesteps.

4.2 Experimental Settings

Dataset Description The TextZoom dataset, as outlined in Wang *et al.*'s paper [39], comprises a substantial collection of $21,740$ LR-HR paired text images, and it includes associated text labels. This dataset originates from two prominent super-resolution datasets, RealSR [2] and SRRAW [44], specifically adapted for text image super-resolution. These LR-HR pairs are gathered under diverse real-world conditions, utilizing various cameras with varying focal lengths. This diversity aims to mimic real-world scenarios and challenges. The training set has $17,367$ samples, and the testing set is further divided into three subsets based on the focal length of the camera. The testing subsets are named *easy*, *medium*, and *hard* with $1,619$ samples, $1,411$ samples and $1,343$ samples correspondingly.

The Real-CE dataset is a real-world Chinese-English benchmark dataset [17] comprising 33,789 text line pairs. It includes 24,666 Chinese texts and 9,123 English texts. The dataset is divided into 23,547 lines for training, 3,414 lines for 4× zooming evaluation, and 6,828 lines for 2× zooming evaluation. The text line sizes vary from 16×22 to 1,156×2,883, comprising 3,755 character categories.

Evaluation Metric Three recognizers, ASTER [33], CRNN [32], and MORAN [15], are used to evaluate the accuracy metric on the TextZoom dataset. On the Real-CE dataset, the pre-trained TransOCR model [32, 43] is utilized as the accuracy metric. Additionally, for evaluating the quality of super-resolution, the PSNR and SSIM metrics are employed [41].

Training Strategy In our proposed approach, we employ HR-LR paired text images to train a diffusion-based module exclusively designed for denoising tasks within a UNet architecture. Notably, this module focuses solely on enhancing the denoising capabilities and does not contribute to the upscaling of the input image. To align the LR images with the desired HR dimensions, we opt for

Table 1: Comparison of the proposed method with SOTA methods on three subsets of the TextZoom dataset. The results are evaluated on the recognition accuracy of three text recognizers: CRNN [32], MORAN [15], and ASTER [33]. BICUBIC (HR)↓ indicates the downscaled bicubic interpolation of HR images, which can be regarded as equivalent to ground truth. BICUBIC (LR)↑ indicates the upscaled bicubic interpolation of LR images, which shows the baseline recognition results. TCDM [23] is trained without synthesized data for a fairer comparison. The best results are highlighted in bold and the second-best results are underlined, with the exception of the ground truth category.

Category	Method	Accuracy of CRNN				Accuracy of MORAN				Accuracy of ASTER			
		Easy	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*
Baseline	BICUBIC (LR)↑	36.4%	21.1%	21.1%	26.8%	60.6%	37.9%	30.8%	44.1%	67.4%	42.4%	31.2%	48.2%
Generic image super-resolution	SRCNN [6]	41.1%	22.3%	22.0%	29.2%	63.9%	40.0%	29.4%	45.6%	70.6%	44.0%	31.5%	50.0%
	SRResNet [14]	45.2%	32.6%	25.5%	35.1%	66.0%	47.1%	33.4%	49.9%	69.4%	50.5%	35.7%	53.0%
	RCAN [45]	46.8%	27.9%	26.5%	34.5%	63.1%	42.9%	33.6%	47.5%	67.3%	46.6%	35.1%	50.7%
	SAN [5]	50.1%	31.2%	28.1%	37.2%	65.6%	44.4%	35.2%	49.4%	68.1%	48.7%	36.2%	52.0%
	HAN [22]	51.6%	35.8%	29.0%	39.6%	67.4%	48.5%	35.4%	51.5%	71.1%	52.8%	39.0%	55.3%
Text-based backbone	TSRN [39]	52.5%	38.2%	31.4%	41.4%	70.1%	55.3%	37.9%	55.4%	75.1%	56.3%	40.1%	58.3%
	TBSRN [3]	59.6%	47.1%	35.3%	48.1%	74.1%	57.0%	40.8%	58.4%	75.7%	59.9%	41.6%	60.0%
	PCAN [47]	59.6%	45.4%	34.8%	47.4%	73.7%	57.6%	41.0%	58.5%	77.5%	60.7%	43.1%	61.5%
Stroke-aware	TG [4]	61.2%	47.6%	35.5%	48.9%	75.8%	57.8%	41.4%	59.4%	77.9%	60.2%	42.4%	61.3%
Text-prior	TPGSR [16]	63.1%	52.0%	38.6%	51.8%	74.9%	60.5%	44.1%	60.5%	78.9%	62.7%	44.5%	62.8%
	TATT [18]	62.6%	53.4%	39.8%	52.6%	72.5%	60.2%	43.1%	59.5%	78.9%	63.4%	45.4%	63.6%
	C3-STISR [48]	65.2%	53.6%	39.8%	53.7%	74.2%	61.0%	43.2%	60.5%	79.1%	63.3%	46.8%	64.1%
Diffusion + Text-prior + Synthsized	TCDM [23]	**67.3%**	**57.3%**	**42.7%**	**55.7%**	77.6%	62.9%	**45.9%**	62.2%	81.3%	**65.1%**	**50.1%**	65.5%
DCDM	Proposed	65.7%	**57.3%**	41.4%	55.5%	**78.4%**	**63.5%**	45.3%	**63.4%**	**81.8%**	**65.1%**	47.4%	**65.8%**
Ground truth	BICUBIC (HR) ↓	76.4%	75.1%	64.6%	72.4%	91.2%	85.3%	74.2%	84.1%	94.2%	87.7%	76.2%	86.6%

a straightforward solution—applying bicubic interpolation to the LR images. Therefore, in the context of our work, LR images essentially undergo a simple bicubic upscaling process from their original low-resolution state to match the specified HR image size.

In the training, we exclusively utilized HR images as ground truth, without incorporating any additional text tokens in contrast to TCDM [23], which used the ground truth text as their ground truth. As we show later in Tab. 1, the accuracy of HR images is no better than 82% on average. Since the ground truth text is always correct, our proposed method is overwhelmingly disadvantageous compared to TCDM.

4.3 Experimental Results

To provide a comprehensive assessment, the proposed method is compared with a wide variety of methods. It includes generic image super-resolution methods (SRCNN [6], SRResNet [14], RCAN [45], SAN [5], and HAN [22]), text-based backbone networks (TSRN [39], TBSRN [3], and PCAN [47]), a stroke-aware loss-based method (TG [4]), text-prior networks (TPGSR [16], TATT [18] and C3-STISR [48]), and diffusion networks with text prior and synthesized data (TCDM [23]) on the TextZoom dataset. Additionally, as a baseline for comparison, we evaluate the text recognition accuracy of images upscaled through simple

Table 2: Comparison of the proposed method with SOTA methods on three subsets of the TextZoom dataset. The results are evaluated on PSNR and SSMI. "-" indicates that the value is not available. TCDM [23] is trained without synthesized data for a fairer comparison. The best results are highlighted in bold and the second-best results are underlined.

Category	Method	PSNR				SSMI ($\times 10^{-2}$)			
		Easy	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*
Baseline	BICUBIC (LR)↑	22.35	18.98	19.39	20.35	78.84	62.54	65.92	69.61
Generic image	SRCNN [6]	23.48	19.06	19.34	20.78	83.79	63.23	67.91	72.27
	SRResNet [14]	24.36	18.88	19.29	21.03	86.81	64.06	69.11	74.03
	HAN [22]	23.30	19.02	20.16	20.95	86.91	65.37	73.87	75.96
Text-based backbone	TSRN [39]	25.07	18.86	19.71	19.70	88.97	66.76	73.02	71.57
	TBSRN [3]	23.46	19.17	19.68	19.10	87.29	64.55	74.52	70.66
	PCAN [47]	24.57	19.14	20.26	21.49	88.30	67.81	74.75	77.52
Stroke-aware	TG [4]	-	-	-	21.40	-	-	-	74.56
Text-prior	TPGSR [16]	24.35	18.73	19.93	19.79	88.60	67.84	75.07	72.93
	TATT [18]	24.72	19.02	20.31	21.52	90.06	**69.11**	77.03	79.30
	C3-STISR [48]	-	-	-	21.51	-	-	-	77.21
Diffusion + Text-prior + Synthsized	TCDM [23]	-	-	-	<u>22.83</u>	-	-	-	**79.58**
DCDM	Proposed	**26.47**	**20.29**	**21.25**	**22.87**	**90.80**	68.73	**77.34**	<u>79.54</u>

Table 3: Comparison of STISR methods with the proposed method on the Real-CE dataset [17] for 2x and 4x scaling, evaluated on PSNR, SSIM ($\times 10^{-2}$), and text recognition accuracy (ACC).

Method	×4						×2					
	Trained on TextZoom			Trained on Real-CE			Trained on TextZoom			Trained on Real-CE		
	PSNR	SSIM	ACC	PSNR	SSIM	ACC	PSNR	SSIM	ACC	PSNR	SSIM	ACC
TSRN [39]	17.47	48.53	17.96	18.11	48.50	23.16	18.73	56.76	24.71	18.99	52.33	28.54
TPGSR [16]	17.37	49.13	20.76	18.07	47.58	23.26	17.99	53.12	26.55	18.83	55.62	30.07
TBSRN [3]	17.59	49.19	22.46	18.33	48.26	25.27	18.41	54.56	29.05	19.01	53.66	31.81
TATT [18]	17.43	50.10	21.00	17.96	49.04	23.30	18.24	56.67	27.55	19.06	57.72	31.27
DCDM	**18.13**	**50.89**	**22.94**	**18.91**	**50.90**	**25.49**	**18.87**	**56.89**	**29.34**	**19.23**	**58.12**	**31.94**
HR image	-	-	48.07	-	-	45.14	-	-	48.07	-	-	45.14

bicubic interpolation (denoted as BICUBIC (LR)↑). The proposed method is also evaluated on the RealCE dataset [17].

Quantitative Results The proposed method is compared with other methods. The results are summarized in Tab. 1. From the table, it has been observed that the proposed method has outperformed all generic image super-resolution methods (SRCNN [6], SRResNet [14], RCAN [45], SAN [5], and HAN [22]) in all test sets. It shows the superiority of the proposed method over the traditional generic image super-resolution methods. The proposed method is also better compared to text-based backbone networks (TSRN [39], TBSRN [3], and PCAN [47]) in all three test datasets and across all three recognizers. The proposed method also achieves superior results than the stroke-level loss based method (i.e., TG [4]). We also compare our proposed method with the text prior

Table 4: Ablation study to identify the impact of latent text diffusion model in the proposed diffusion-conditioning-diffusion model. The results are evaluated on the recognition accuracy of three text recognizers: CRNN [32], MORAN [15], and ASTER [33]. TD and ID indicate the latent text diffusion model and the latent image diffusion model, respectively. The best results are highlighted in bold. The gains of the proposed method are shown in parentheses.

Method	Comp.		Accuracy of CRNN			Accuracy of MORAN			Accuracy of ASTER		
	TD	ID	*Easy*	*Medium*	*Hard*	*Easy*	*Medium*	*Hard*	*Easy*	*Medium*	*Hard*
DCDM w/o text	✗	✓	65.2%	56.5%	41.3%	78.1%	61.7%	44.5%	80.4%	64.5%	46.7%
DCDM	✓	✓	**65.7%** (↑0.5%)	**57.3%** (↑0.8%)	**41.4%** (↑0.1%)	**78.4%** (↑0.3%)	**63.5%** (↑1.8%)	**45.3%** (↑0.8%)	**81.8%** (↑1.4%)	**65.1%** (↑0.6%)	**47.4%** (↑0.7%)

approach (TPGSR [16], TATT [18] and C3-STISR [48]), which shares the use of text prior and is closely related to our approach. Our proposed method outperform outperformed on comparison with text prior methods. Finally, the diffusion networks with text prior and synthesized data (TCDM [23]) is compared with our proposed method. The table shows that both our proposed method and TCDM achieved the best accuracy in seven out of 12 columns. Therefore, our proposed method was comparable to TCDM. It should be noted that TCDM has an advantage in this comparison: as argued above, TCDM uses the text ground truth in the training, whereas our proposed method does not.

The evaluation of the proposed model on super-resolution image quality is presented in Tab. 2. Our model demonstrates significant performance improvements, nearly outperforming other methods in terms of both PSNR and SSIM metrics. The detailed results show the effectiveness of the proposed approach in enhancing the quality of super-resolution images compared to existing methods.

The performance comparison shown in Tab. 3 highlights the efficacy of various state-of-the-art STISR methods on the Real-CE dataset for 2× and 4× upscaling. The evaluation metrics used are PSNR, SSIM, and text recognition accuracy (ACC) with the pre-trained TransOCR model [32,43]. The table presents results for models trained and tested on Real-CE, as well as models trained on TextZoom and tested on Real-CE. The results demonstrate that the proposed method consistently outperforms the compared methods across both upscaling factors and all three evaluation metrics, regardless of the training dataset.

Qualitative Results In Fig. 3, we conduct a qualitative analysis of our proposed model, comparing it with representative models from different categories: SRCNN [6] (a generic image super-resolution method), TBSRN [39] (a text super-resolution backbone method), TG [4] (a stroke-level trained model), and TATT [18] (a text prior-based model). These comparisons are crucial to assess the model's performance comprehensively.

To comprehensively assess the proposed model, we conducted a comparison using diverse samples featuring variations in contrast, color, degrees of blurriness, orientation, and other factors. Fig. 3 includes a subset where our proposed model excels in correct recognition, surpassing all other models in the ability

Fig. 3: High resolution images generation by various state-of-the-art methods and the proposed method. BICUBIC (LR) and HR indicate the input to the methods and ground truth, respectively. The text recognition results by ASTER [33] are shown under the images. The characters in green and red indicate correct and incorrect recognition results, respectively.

to produce accurate results. While TATT [18] is specifically trained with a loss function designed to handle text orientation, our proposed model, interestingly, is not explicitly trained with any orientation loss. Nevertheless, our model performs comparably to TATT [18] in handling orientation-related challenges. In the accompanying figure, we present a sample that is correctly recognized by our model but is misread by other state-of-the-art Scene Text Image Super-Resolution (STISR) models. This visually illustrates the superior performance of our approach, particularly in terms of text recognition. The results underscore the robustness and effectiveness of our proposed model across a spectrum of challenging scenarios, showcasing its potential for advancing the field of STISR. In summary, our results, as presented, underscore the effectiveness of our proposed model across various facets of text-based image super-resolution, spanning accurate recognition, the importance of text prior, and high-quality image generation.

4.4 Ablation Study

In this section, an ablation study is carried out to confirm the effectiveness of the components of the proposed method. Since the proposed method consists of two diffusion models, the impact of the text diffusion block has been analyzed by removing the text-based conditioning module (denoted by "DCDM w/o text"). This ablation is mentioned in Tab. 4. It has been observed that text diffusion has a positive impact on the proposed model (DCDM). The addition of text diffusion to the image diffusion via conditioning helps to uplift the accuracy of all used text recognizers in all three subsets of datasets. The second variant of our proposed model involves the removal of the image conditioning. In this configuration, the generated text embedding from the text-based latent diffusion module is used as input for the low-resolution latent. This text embedding condition

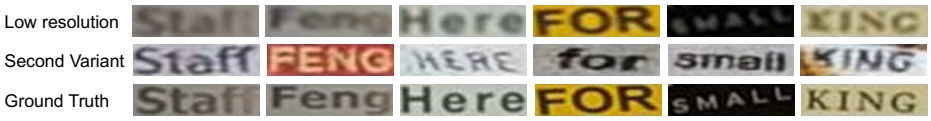

Fig. 4: Some images generated from the low-resolution images (first row) by removing the image latent conditioning in the proposed model (DCDM). Still, the model can render the correct text, but fidelity is compromised.

is then employed to guide the subsequent image diffusion model in generating high-resolution images. This variant eliminates the direct low-resolution conditioning of the model. The generated results from this second variant are visually presented in Fig. 4, showcasing the model's capability to generate high-quality images based on text embeddings and diffusion techniques.

5 Limitations

Our current implementation consume a significant amount of time [34] in the diffusion solver steps. We expect that it is mitigated by just plugging a more efficient substitute. Another limitation of this paper is that we could not compare the proposed method and TCDM under a fair condition (see Appendix A).

6 Conclusion

In this paper, we introduced a novel approach to scene text image super-resolution (STISR) through the *diffusion-conditioned-diffusion model* (DCDM). This model incorporates two distinct diffusion modules, addressing text denoising and image denoising. The latent diffusion module for text denoising plays a critical role in generating a text prior, effectively mapping noise to character embeddings with the aid of encoded low-resolution images. The image encoder, trained with high-resolution image text embeddings using the character-level CLIP (CL-CLIP) model, contributes significantly to this process. The resulting character embedding prior serves as a conditioning criterion for the diffusion module dedicated to image denoising. Additionally, a conditioning mechanism involving low-resolution images and a normal distribution further refines the model.

In contrast to the proposed method, conventional STISR methods rely on a text recognizer. Hence, we posed the question, *"Is a text recognizer necessary for STISR during inference?"* To answer the question, quantitative and qualitative evaluations on the TextZoom and Real-CE datasets were performed. The quantitative results revealed an improvement over state-of-the-art methods, emphasizing the model's effectiveness. Qualitatively, the generated images not only demonstrated realistic characteristics but also maintained fidelity under the proposed DCDM framework. The experimental results allow us to conclude *"No, a latent text diffusion model can effectively replace a text recognizer."*

Acknowledgments

This work was partially supported by JSPS Kakenhi Grant Numbers 22H00540 and 24K03020.

References

1. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4291–4301 (2019) 1
2. 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 (2019) 9
3. 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 and Pattern Recognition. pp. 12026–12035 (2021) 2, 4, 10, 11
4. 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 (2022) 10, 11, 12
5. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019) 10, 11
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015) 4, 10, 11, 12
7. Dong, C., Zhu, X., Deng, Y., Loy, C.C., Qiao, Y.: Boosting optical character recognition: A super-resolution approach. arXiv preprint arXiv:1506.02211 (2015) 4
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8
9. Fu, B., Li, Y., Wang, X.h., Ren, Y.g.: Image super-resolution using tv priori guided convolutional network. Pattern Recognition Letters **125**, 780–784 (2019) 4
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 9
11. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021) 6
12. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. IEEE transactions on pattern analysis and machine intelligence **32**(6), 1127–1133 (2010) 3
13. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017) 4
14. 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 (2017) 10, 11
15. Long, S., He, X., Yao, C.: Scene text detection and recognition: The deep learning era. International Journal of Computer Vision **129**, 161–184 (2021) 9, 10, 12
16. Ma, J., Guo, S., Zhang, L.: Text prior guided scene text image super-resolution. IEEE Transactions on Image Processing **32**, 1341–1353 (2023) 2, 4, 7, 10, 11, 12
17. Ma, J., Liang, Z., Xiang, W., Yang, X., Zhang, L.: A benchmark for chinese-english scene text image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19452–19461 (2023) 9, 11
18. Ma, J., Liang, Z., Zhang, L.: A text attention network for spatial deformation robust scene text image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5911–5920 (2022) 2, 7, 10, 11, 12, 13
19. 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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 158–174. Springer (2020) 2, 4
20. Nakaune, S., Iizuka, S., Fukui, K.: Skeleton-aware text image super-resolution. University of Tsukuba: Tsukuba, Japan (2021) 2
21. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021) 5
22. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 191–207. Springer (2020) 10, 11
23. Noguchi, C., Fukuda, S., Yamanaka, M.: Scene text image super-resolution based on text-conditional diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1485–1495 (2024) 2, 10, 11, 12, 1
24. Pandey, R.K., Vignesh, K., Ramakrishnan, A., et al.: Binary document image super resolution for improved readability and OCR performance. arXiv preprint arXiv:1812.02475 (2018) 4
25. Pernias, P., Rampas, D., Richter, M.L., Pal, C., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=gU58d5QeGv 3
26. Peyrard, C., Baccouche, M., Mamalet, F., Garcia, C.: Icdar2015 competition on text image super-resolution. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1201–1205. IEEE (2015) 4
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 7
28. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 7
29. Ren, C., He, X., Nguyen, T.Q.: Single image super-resolution via adaptive high-dimensional non-local total variation and adaptive geometric feature. IEEE Transactions on Image Processing **26**(1), 90–106 (2016) 3

30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022) 6, 8
31. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022) 2, 6
32. 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 transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016) 8, 9, 10, 12
33. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2035–2048 (2018) 1, 9, 10, 12, 13, 3
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 8, 9, 14
35. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of the IEEE international conference on computer vision. pp. 4799–4807 (2017) 4
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6
37. Wang, H., Bai, X., Yang, M., Zhu, S., Wang, J., Liu, W.: Scene text retrieval via joint text detection and similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4558–4567 (2021) 1
38. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023) 2
39. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 650–666. Springer (2020) 2, 4, 9, 10, 11, 12
40. 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) 2
41. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9
42. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 251–260 (2017) 2
43. Yu, H., Chen, J., Li, B., Ma, J., Guan, M., Xu, X., Wang, X., Qu, S., Xue, X.: Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. arXiv preprint arXiv:2112.15093 (2021) 9, 12
44. 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 (2019) 9
45. 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 (2018) 10, 11

46. 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 (2018) 4
47. 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 (2021) 10, 11
48. Zhao, M., Wang, M., Bai, F., Li, B., Wang, J., Zhou, S.: C3-STISR: Scene text image super-resolution with triple clues. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 1707–1713 (2022). https://doi.org/10.24963/ijcai.2022/238 2, 10, 11, 12
49. Zhu, S., Zhao, Z., Fang, P., Xue, H.: Improving scene text image super-resolution via dual prior modulation network. Proceedings of the AAAI Conference on Artificial Intelligence **37**(3), 3843–3851 (2023). https://doi.org/10.1609/aaai.v37i3.25497 2