Diffusion-based Blind Text Image Super-Resolution

Yuzhe Zhang[1]   Jiawei Zhang[2]   Hao Li[2]   Zhouxia Wang[3]   Luwei Hou[2]
Dongqing Zou[2,4]   Liheng Bian[1,*]

[1] Beijing Institute of Technology, [2] SenseTime Research, [3] The University of Hong Kong, [4] PBVR
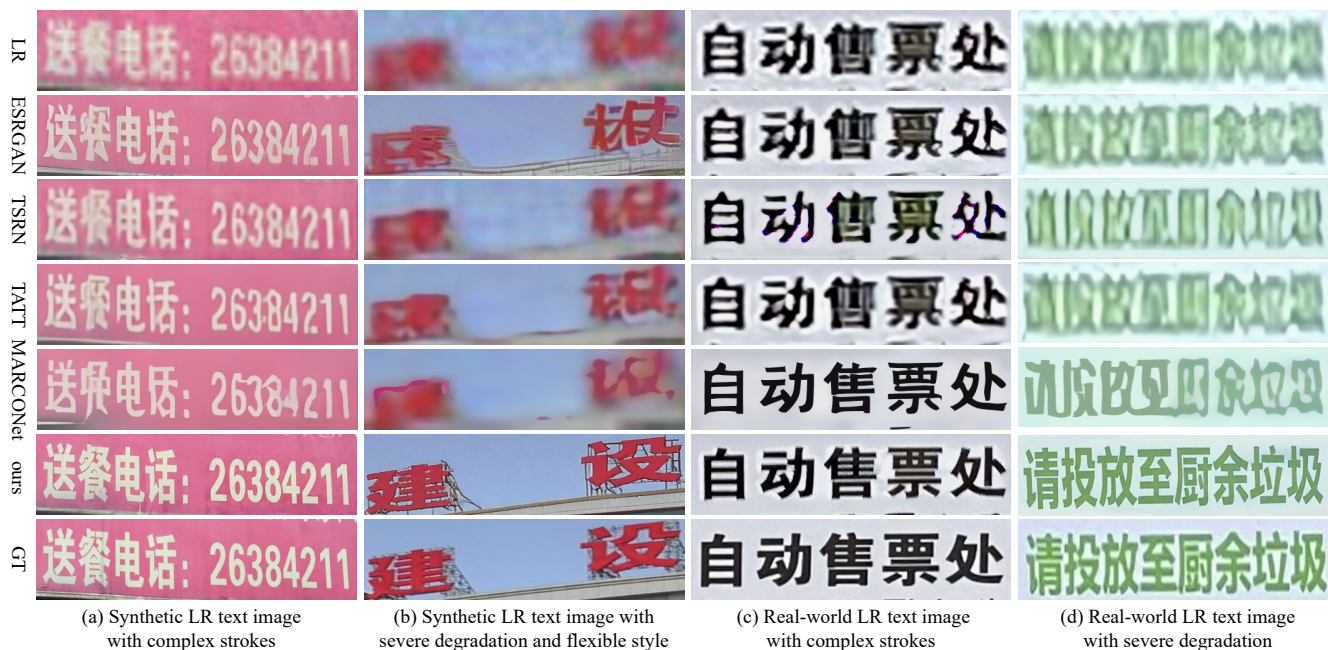
Figure 1. Blind text image super-resolution results between different methods on synthetic and real-world text images. Our method can restore text images with high text fidelity and style realness under complex strokes, severe degradation, and various text styles.

(a) Synthetic LR text image with complex strokes
(b) Synthetic LR text image with severe degradation and flexible style
(c) Real-world LR text image with complex strokes
(d) Real-world LR text image with severe degradation

Abstract

Recovering degraded low-resolution text images is challenging, especially for Chinese text images with complex strokes and severe degradation in real-world scenarios. Ensuring both text fidelity and style realness is crucial for high-quality text image super-resolution. Recently, diffusion models have achieved great success in natural image synthesis and restoration due to their powerful data distribution modeling abilities and data generation capabilities. In this work, we propose an Image Diffusion Model (IDM) to restore text images with realistic styles. For diffusion models, they are not only suitable for modeling realistic image distribution but also appropriate for learning text distribution. Since text prior is important to guarantee the correctness of the restored text structure according to existing arts, we also propose a Text Diffusion Model (TDM) for text recognition which can guide IDM to generate text images with correct structures. We further propose a Mixture of Multi-modality module (MoM) to make these two diffusion models cooperate with each other in all the diffusion steps. Extensive experiments on synthetic and real-world datasets demonstrate that our Diffusion-based Blind Text Image Super-Resolution (DiffTSR) can restore text images with more accurate text structures as well as more realistic appearances simultaneously. Code is available at https://github.com/YuzheZhang-1999/DiffTSR.

1. Introduction

Blind text image super-resolution (SR) focuses on recovering high-resolution (HR) images from low-resolution (LR) ones corrupted by various unknown degradations. Unlike natural image super-resolution tasks which pay more attention to enriching and enhancing the image details, text fidelity and style realness should also be guaranteed in the restored text images. Mistakenly estimated text structures, such as distorted, missing, additional, or overlapping

---

[1]This work was done when Yuzhe Zhang was an intern at SenseTime.
* Corresponding author.

strokes, will lead to inaccurate character semantics which is unacceptable in the restored text images. Similarly, incorrectly generated text styles, such as changes in fonts, glyphs, colors, and poses, will make the restored text images visually unpleasant and unreal.

In order to reconstruct text images with the correct structures, existing methods [3, 25–27, 32, 42] introduce to utilize low-level and high-level text priors to guide the restoration process by considering text structure-related losses or incorporating additional text recognition modules. Although these methods enhance the visual appearance of characters in reconstructed images, they are difficult to restore accurate text structure when encountering text images with complex strokes or severe degradations. To alleviate the above issues, MARCONet [23] employs a codebook to store the discrete code of each character which can be used to generate high-resolution structural details with high text fidelity. In addition, StyleGAN [20] is exploited in MARCONet to generate visually pleasant text styles. Even though MARCONet can handle complex strokes and severe degradation to a certain extent, the predefined font styles during training limit its ability to deal with unseen and diverse text styles in the real world, leading to the unrealness and infidelity in some restored images.

Recently, diffusion models [15, 37] have exhibited great success in natural image synthesis [6, 31, 33, 38] and restoration [8, 21, 24, 36] due to their powerful data distribution modeling and data generation capabilities. In this paper, we argue that diffusion model should also be suitable to model diverse text styles, which include fonts, glyphs, colors, and poses, to restore visually more pleasant and realistic text images. As a result, we propose an Image Diffusion Model (IDM) based on stable diffusion [33] to effectively model the text styles. To keep the text character fidelity, IDM is conditioned on the input low-resolution image and text prediction priors. However, accurately recognizing text from severely degraded images is challenging, and inaccurate text recognition will lead to incorrect text structures in the restoration results. According to the analysis of [17], diffusion model is also appropriate to model the discrete variable distribution like text. On this basis, we introduce to use a Text Diffusion Model (TDM) to correctly recognize texts conditioned on low-resolution input and provide text prior to help IDM restore text images with high fidelity. It is worth emphasizing that TDM can benefit IDM and vice versa. Therefore, we further propose a Mixture of Multimodality module (MoM) so that these two diffusion models can cooperate with each other in all the diffusion steps.

Extensive experiments demonstrate that our Diffusion-based Blind Text Image Super-Resolution (DiffTSR) can restore text images, especially for Chinese text images with complex strokes, from degraded ones with satisfactory text fidelity and style realness simultaneously. In summary, our work has the following main contributions:

- We propose to use IDM and TDM to model text image distribution and text distribution in order to restore text images with high text fidelity and style realness.
- We propose a MoM module to make IDM and TDM closely cooperate with each other in all the diffusion steps.
- Extensive experiments demonstrate that the proposed DiffTSR performs better than existing methods on both synthetic and real-world datasets.

2. Related Work

Blind Image Super-Resolution. Blind image super-resolution (SR) aims to enhance the resolution and quality of images with complex unknown degradation in real-world scenarios. Recent works have made efforts to achieve more effective blind SR from two aspects: degradation model estimation [1, 12, 18, 28] and real-world data synthesis [2, 9, 19, 46]. The former learns the degradation model from low-resolution (LR) images in an unsupervised manner [40] and then applies non-blind SR methods. The latter involves synthesizing LR-HR image training pairs through a complex degradation strategy that imitates real-world degradation. Specifically, BSRGAN [49] uses a random shuffling strategy to achieve more generalized degradation data synthesis, while Real-ESRGAN [44] further enhances the complexity of image degradation through a high-order degradation modeling process. Although the above methods have achieved great success in blind SR of natural images, we observe that it is insufficient to effectively enhance the quality of text images without considering the specific character structures and text style.

Text Image Super-Resolution. Text image super-resolution aims to enhance the details of the image meanwhile improving the readability of the text, *i.e.* the accuracy of text recognition. Recent research mainly focuses on exploring the guidance of text recognition priors and character structure priors to improve the performance of text image SR. Specifically, based on the characteristics of text images, existing works mainly exploit text-related prior information from three aspects to constrain text image super-resolution: text aware loss [3, 41, 42], text recognition prior [25, 30, 47, 52], and text structure prior [23]. Previous research demonstrates that the text priors play an important role in text structure enhancement. However, most of them do not fully utilize text prior information and cannot restore text images with diverse text styles, severe degradation, or complex strokes.

Diffusion Model. Diffusion model [15] has attracted great attention, due to its impressive performance in image synthesis [6, 13, 33], and controllable image generation [16, 29, 34, 35, 50]. Benefiting from the powerful data distribution modeling capability of diffusion models, recent research [21, 22, 36, 39, 45] also achieves impres-
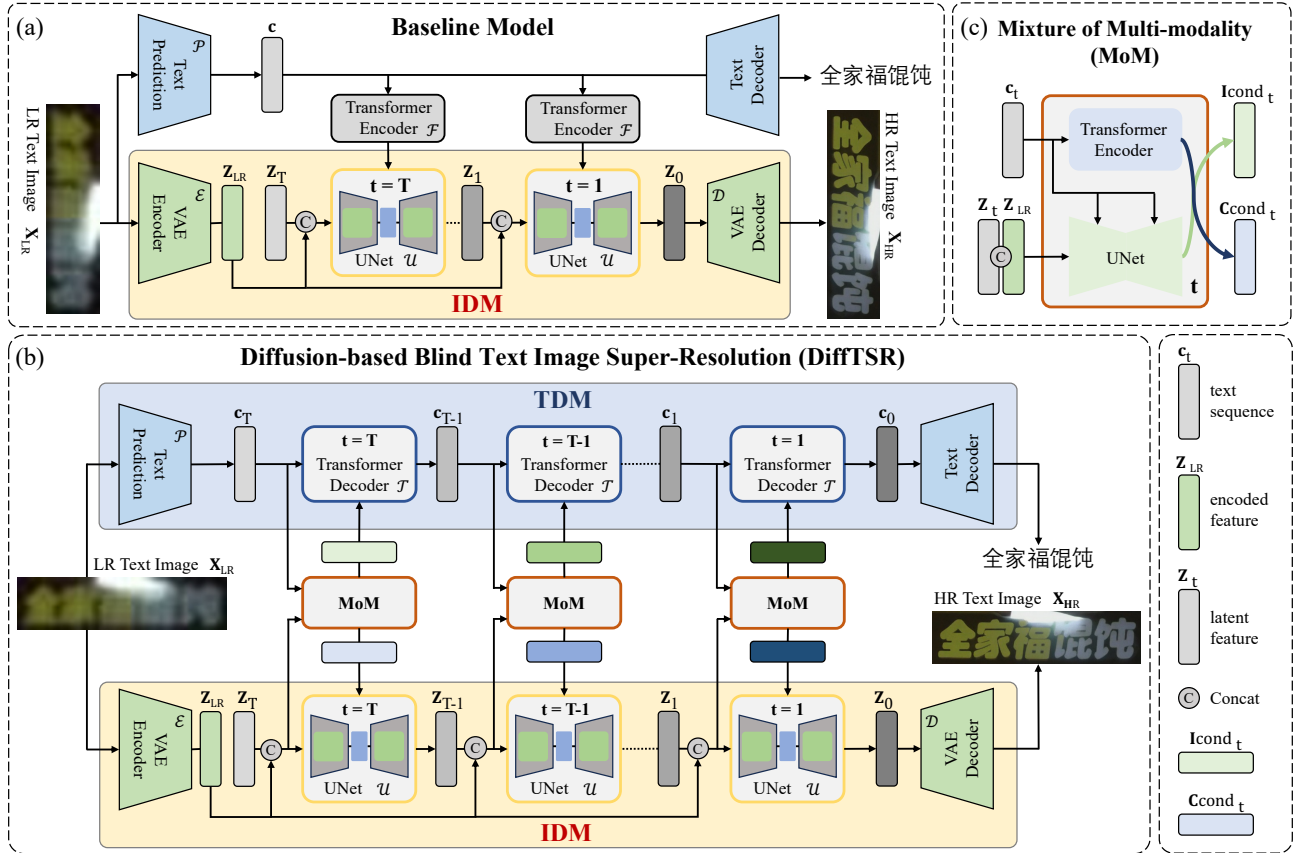
Figure 2. Overview of Diffusion-based Blind Text Image Super-Resolution (DiffTSR) along with the baseline. (a) Our baseline model. It contains an Image Diffusion Model (IDM) and a text recognition model. The IDM performs the diffusion-based text image super-resolution conditioned on the latent feature $\mathbf{Z}_{LR}$ from the LR image and text prior $\mathbf{c}$ which is extracted by the text recognition model from the LR image. (b) DiffTSR architecture. It mainly consists of three parts: i) IDM performs the image diffusion conditioned on $\mathbf{Z}_{LR}$ and $\mathbf{C}cond_t$ to achieve the high-realness image generation, ii) TDM conducts the text diffusion conditioned on $\mathbf{I}cond_t$, which starts the reverse process from the initial text prior $\mathbf{c}_T$, to achieve more accurate text prior prediction and correction, iii) MoM module fuses and encodes the intermediate features of IDM and TDM at the previous step, and outputs the conditions $\mathbf{C}cond_t$ and $\mathbf{I}cond_t$ for the current time step. IDM and TDM cooperate with each other through MoM to finally achieve text image super-resolution with high fidelity and realness. (c) Details of MoM. It fuses $\mathbf{Z}_{LR}$, $\mathbf{Z}_t$, and $\mathbf{c}_t$ at step t, and encodes them into $\mathbf{I}cond_t$ and $\mathbf{C}cond_t$ for TDM and IDM respectively.

sive performance in image super-resolution by utilizing the diffusion prior. In addition, existing research has shown that diffusion models are also suitable for modeling discrete data [17], such as text [11], segmentation map [48], *etc*. In this work, we aim to explore the collaboration between image diffusion models and text diffusion models, and achieve high-quality text image super-resolution with high text fidelity and style realness.

3. Methodology

3.1. Overview

In this paper, we propose to use the diffusion model to restore degraded text images by considering text prior. We first propose a baseline model which is shown in Figure 2 (a) and described in Section 3.2. It uses a text recog-

nition model to provide text prior. Then, the proposed Image Diffusion Model (IDM) is used to restore the text images conditioned on the text prior. Even though the above baseline model can restore degraded text images with relatively acceptable fidelity, it will produce distorted text structures when encountered with severe degradation. Besides IDM, the proposed Diffusion-based Blind Text Image Super-Resolution (DiffTSR) also contains a Text Diffusion Model (TDM) and a Mixture of Multi-modality module (MoM) based on the assumption that more accurate text recognition information can be beneficial for IDM to generate a more realistic image; meanwhile, a higher-quality text image can benefit for better recognition. In DiffTSR, TDM is a diffusion model that gradually recognizes text sequence with given image information. As to MoM, it is like a bridge to connect IDM with TDM. It provides updated

text prior to IDM and image information to TDM during the diffusion process. In this way, the proposed DiffTSR can restore text images with high style realness and text fidelity simultaneously. The overall architecture of DiffTSR is shown in Figure 2 (b) and described in Section 3.3.

3.2. Baseline

Our baseline model is illustrated in Figure 2 (a). It mainly consists of two parts, 1) an Image Diffusion Model (IDM), 2) a text recognition model. The text recognition model estimates text sequence $\mathbf{c}$ from the low-resolution text images $\mathbf{X}_{LR}$ as text prior in every diffusion step, and IDM implements the image super-resolution through the diffusion reverse process conditioned on $\mathbf{c}$ and $\mathbf{X}_{LR}$.

To model the distribution of real-world text images and achieve realistic image generation, the proposed IDM is based on Stable Diffusion [33]. IDM performs the diffusion forward and reverse process in the latent space through a VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$. After encoding HR image $\mathbf{X}$ into latent space by $\mathcal{E}$ as $\mathbf{Z} = \mathcal{E}(\mathbf{X})$, IDM sequentially adds noises into $\mathbf{Z}$ at time step t as $\mathbf{Z}_t$ and a sequence of noise prediction network $\mathcal{U}_\theta$ is used to gradually remove the noises in the reverse process. In order to make the restoration results consistent with the input LR image, encoded feature $\mathbf{Z}_{LR} = \mathcal{E}(\mathbf{X}_{LR})$ is considered as a condition in $\mathcal{U}_\theta$ by concatenation with $\mathbf{Z}_t$. Meanwhile, text prior is also considered as another condition in $\mathcal{U}_\theta$. Specifically, we encode $\mathbf{c}$, which is estimated from the text recognition model $\mathcal{P}$ in [3], through a transformer encoder $\mathcal{F}_\psi$, and fuse the encoded feature $\mathcal{F}_\psi(\mathbf{c})$ into the intermediate layers of $\mathcal{U}_\theta$ by the cross-attention mechanism.

The details of the sampling process of our baseline model can be described as follows. We first utilize the text recognition model $\mathcal{P}$ to predict the text sequence $\mathbf{c}$ from the LR text image $\mathbf{X}_{LR}$. After that, IDM starts the reverse process and repeats the denoising step $\mathcal{U}_\theta$ conditioned on the latent feature $\mathbf{Z}_{LR}$ extracted from LR image $\mathbf{X}_{LR}$ by VAE encoder $\mathcal{E}$ and text prior $\mathcal{F}_\psi(\mathbf{c})$ extracted from $\mathbf{c}$ by a transformer encoder $\mathcal{F}_\psi$ until obtaining $\mathbf{Z}_0$. Then the restored text image can be reconstructed through VAE decoder as $\mathcal{D}(\mathbf{Z}_0)$. Our baseline model can restore text images with high realness, which benefits from the ability of IDM to generate realistic details.

3.3. Diffusion-based Blind Text Image Super-Resolution

Even though the baseline model above can effectively restore low-resolution text images, it will still generate unpleasant results when encountered with severe degradation which is shown in Figure 3 (a). This is because the text recognition model cannot work well with highly distorted text images and IDM cannot restore text images with high text fidelity under inaccurate text prior. In this subsection, we propose Diffusion-based Blind Text Image
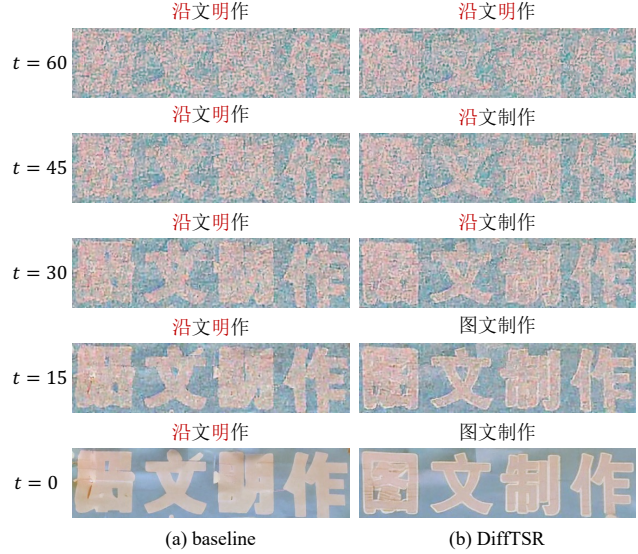

Figure 3. Motivation. To provide text prior for text image restoration, the baseline model recognizes text from degraded images which is inaccurate when the degradation is severe. With inaccurate text prior, the baseline model cannot restore text image with high text fidelity which is shown in (a). The proposed TDM and IDM can benefit from each other through MoM in DiffTSR and gradually recognizes more accurate text sequence and restore higher-quality text image through the reverse diffusion process which is shown in (b). The text sequences above each super-resolution result at different time steps are the recognized text characters used for blind image super-resolution and the characters in red are the mistakenly estimated ones.

Super-Resolution (DiffTSR) to restore text images with high text fidelity and style realness by jointly optimizing image restoration and text recognition in every diffusion step. Besides IDM the same as the baseline model, DiffTSR also contains a Text Diffusion Model (TDM) and a Mixture of Multi-modality module (MoM). The details of TDM, MoM, and the whole DiffTSR are described as follows.

Existing works [10, 17] indicate that the diffusion model can not only model image distribution but also model discrete data such as text. To model the distribution of text sequence $\mathbf{c}$, TDM also follows the Markov chain of the diffusion process that slowly adds random noises to the text sequence in the forward process and then learns the reverse process to reconstruct the text sequence from the noisy data. Unlike IDM which is a continuous diffusion model and the added noises satisfy Gaussian distribution, TDM is a discrete one. Similar to [17, 48], TDM assumes the transition distribution $q(\mathbf{c}_t \mid \mathbf{c}_{t-1})$ follows a categorical distribution in the forward process. With this assumption, TDM proposes to use a Transformer Decoder $\mathcal{T}_\eta$ to remove noises from $\mathbf{c}_t$ and generate $\mathbf{c}_{pred}$. To make the text sequence modeling more consistent with the context of the input image, $\mathbf{I}cond_t$, which contains image information estimated

by MoM, is mapped to the intermediate layer of $\mathcal{T}_\eta$ through the cross-attention mechanism. TDM benefits from text modeling capability as well as image conditions guidance and produces more reasonable and accurate text sequences consistent with the LR image.

As text recognition can benefit text image super-resolution and vice versa, we propose a Mixture of Multi-modality module (MoM) for joint optimization as shown in Figure 2 (c). MoM consists of two time-aware modules, a UNet and a Transformer encoder. The UNet of MoM at time step t first extracts image information from the concatenated $\mathbf{Z}_t$ and $\mathbf{Z}_{LR}$. Then, $\mathbf{c}_t$ is mapped into the intermediate layer of UNet through cross-attention mechanism. UNet fuses and encodes the multi-modality information into the image condition $\mathbf{I}cond_t$ for TDM at each time step, thereby adaptively generating image conditions that are more suitable for TDM to achieve higher recognition accuracy. At the same time, the Transformer encoder of MoM receives corrected characters $\mathbf{c}_t$ from the previous step of TDM, and encodes $\mathbf{c}_t$ into the characters embedding space of IDM as the text condition $\mathbf{C}cond_t$. In summary,

$$[\mathbf{I}cond_t, \mathbf{C}cond_t] = MoM_\phi([\mathbf{Z}_{LR}, \mathbf{Z}_t], \mathbf{c}_t, t), \quad (1)$$

where $\mathbf{C}cond_t$ and $\mathbf{I}cond_t$ serve as the condition for IDM and TDM at time step t, respectively.

After introducing IDM, TDM and MoM, the sampling process of DiffTSR is shown in Algorithm 1 and Figure 2 (b). To begin with, we extract features $\mathbf{Z}_{LR}$ from LR image $\mathbf{X}_{LR}$ through VAE encoder $\mathcal{E}$. At the same time, we randomly sample $\mathbf{Z}_T$ with Gaussian distribution, and get the initially estimated text $\mathbf{c}_T = \mathcal{P}(\mathbf{X}_{LR})$ from the LR image $\mathbf{X}_{LR}$. Note that TDM starts the reverse process from the initial estimated text rather than the random sampled ones. After the initialization process, IDM uses the UNet $\mathcal{U}$ to remove noises from the latent features with given $\mathbf{Z}_{LR}$, the denoised feature from the previous step $\mathbf{Z}_t$ as well as text prior $\mathbf{C}cond_t$ from MoM at every time step. At the same time, TDM uses Transformer Decoder $\mathcal{T}$ to estimate the latent state of the text sequence with given the state $\mathbf{c}_t$ from the previous time step and the image condition $\mathbf{I}cond_t$ from MoM. With T collaborative diffusion steps, $\mathbf{Z}_0$ can be estimated and VAE Decoder $\mathcal{D}$ is used to reconstruct HR text image $\mathbf{X}_0$ with high fidelity and realness. Thanks to the joint optimization strategy through MoM, the proposed DiffTSR can gradually restore HR text image by IDM and more accurate text sequence in TDM which is shown in Figure 3 (b). For more details about the sampling and training strategy of DiffTSR, please see the supplementary material.

4. Experiments

4.1. Experimental Settings

Training Datasets. In this work, we mainly focus on blind text image super-resolution for Chinese characters in the

---

Algorithm 1 DiffTSR Sampling

    ▷ **input :** LR Text Image $\mathbf{X}_{LR}$
    ▷ **output :** HR Text Image $\mathbf{X}_{HR}$
1: $\mathbf{Z}_{LR} = \mathcal{E}(\mathbf{X}_{LR})$
2: $\mathbf{Z}_T \sim \mathcal{N}(0, I)$
3: $\mathbf{c}_T = \mathcal{P}(\mathbf{X}_{LR})$
4: **for** $t = T, \ldots, 1$ **do**
5:     $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
6:     $[\mathbf{I}cond_t, \mathbf{C}cond_t] = MoM_\phi([\mathbf{Z}_{LQ}, \mathbf{Z}_t], \mathbf{c}_t, t)$
7:     $\epsilon_{pred,t} = \mathcal{U}_\theta([\mathbf{Z}_t, \mathbf{Z}_{LR}], \mathbf{C}cond_t, t)$
    // IDM sampling based on stable diffusion [33]
8:     $\mathbf{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t^{IDM}}}\left(\mathbf{Z}_t - \frac{1-\alpha_t^{IDM}}{\sqrt{1-\bar{\alpha}_t^{IDM}}}\epsilon_{pred,t}\right) + \sigma_t \mathbf{z}$
9:     $\mathbf{c}_{pred,t} = \mathcal{T}_\eta(\mathbf{c}_t, \mathbf{I}cond_t, t)$
10:   $\tilde{\boldsymbol{\pi}} = \left[\alpha_t^{TDM}\mathbf{c}_t + \frac{1-\alpha_t^{TDM}}{K}\right] \odot \left[\bar{\alpha}_{t-1}^{TDM}\mathbf{c}_{pred,t} + \frac{1-\bar{\alpha}_{t-1}^{TDM}}{K}\right]$
11:   $\boldsymbol{\pi}_{\text{post}}(\mathbf{c}_t, \mathbf{c}_{pred,t}) = \frac{\tilde{\boldsymbol{\pi}}}{\sum_{k=1}^{K}\tilde{\pi}_k}$
    // TDM sampling based on multinomial diffusion [17]
12:   $\mathbf{c}_{t-1} \sim \mathcal{C}(\mathbf{c}_{t-1} \mid \boldsymbol{\pi}_{\text{post}}(\mathbf{c}_t, \mathbf{c}_{pred,t}))$ if $t > 1$ else $\mathbf{c}_{t-1} \sim \mathcal{C}(\mathbf{c}_0 \mid \mathbf{c}_{pred,t})$
13: **end for**
14: $\mathbf{X}_0 = \mathcal{D}(\mathbf{Z}_0)$
15: **return** $\mathbf{X}_0$
    // $\mathcal{C}$ denotes the categorical distribution with probability parameters after $\mid$.
    // $1 - \alpha_{t-1}^{IDM}$ and $1 - \alpha_{t-1}^{TDM}$ are the noise schedule for IDM and TDM.
    // The processing details from Ln. 9 to Ln. 11 are described in [17].

---

real world. In order to obtain amount of HR Chinese text images along with text annotations, we use the large-scale real-world Chinese text images dataset CTR [4]. To select the images as the ground truth in the training process, we preprocess the CTR training set by the following steps: i) remove the images with a resolution smaller than 64 pixels, ii) only retain images with a width-to-height ratio greater than 2, iii) only retain images with the length of text annotations not larger than 24, iv) resize the image to 128×512. Then, there are 63,644 HR text images $\mathbf{X}_{HR}$ remaining with text annotations $\mathbf{c}$, and we refer to this dataset as the CTR-TSR-Train. The degradation pipeline proposed in BSRGAN [49] and Real-ESRGAN [44] is used to generate LR text images $\mathbf{X}_{LR}$.

Testing Datasets. We evaluate our method on both synthetic and real-world datasets for $\times 2$ and $\times 4$ blind super-resolution. For the synthetic testing set, we use the same preprocessing and degradation strategy as in CTR-TSR-Train to generate CTR-TSR-Test. The images are selected from the testing set of CTR and there are 8,089 samples in total. For real-world dataset, we use the RealCE [27] testing set, which is a recently proposed real-world Chinese-English benchmark dataset. We remove the images with

| method | ×2 | | | | | ×4 | | | | |
	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑
SRCNN	23.73	0.338	54.47	0.7856	0.7991	20.74	0.501	116.5	0.6031	0.6160
ESRGAN	24.75	0.191	9.308	0.8112	0.8239	20.90	0.310	21.86	0.6179	0.6272
NAFNet	25.04	0.286	37.42	0.8083	0.8212	21.82	0.447	87.93	0.6451	0.6573
TSRN	20.86	0.392	70.75	0.7805	0.7937	19.41	0.535	137.3	0.6149	0.6267
TBSRN	24.43	0.282	57.61	0.8018	0.8156	21.56	0.442	132.6	0.6360	0.6486
TATT	24.87	0.291	58.73	0.7911	0.8041	21.84	0.453	107.6	0.6273	0.6403
MARCONet	20.77	0.374	94.60	0.6934	0.7068	19.33	0.436	108.5	0.5123	0.5241
ours	**25.08**	**0.156**	**5.906**	**0.8594**	**0.8718**	**21.85**	**0.231**	**8.482**	**0.8350**	**0.8471**

Table 1. Quantitative comparison for the synthetic dataset CTR-TSR-Test with different methods including SRCNN [7], ESRGAN [43], NAFNet [5], TSRN [42], TBSRN [3], TATT [25], MARCONet [23] and our method for ×2 and ×4 blind text image super-resolution.

| method | ×2 | | | | | ×4 | | | | |
	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑
SRCNN	17.87	0.224	54.44	0.7922	0.8936	16.63	0.364	128.1	0.7101	0.8018
ESRGAN	18.19	0.231	28.70	0.7929	0.8945	16.84	0.407	83.22	0.7121	0.8047
NAFNet	17.86	0.216	50.42	0.7916	0.8925	16.76	0.359	118.1	0.7122	0.8023
TSRN	15.54	0.327	106.0	0.7892	0.8902	15.22	0.418	148.5	0.6963	0.7873
TBSRN	17.34	0.236	71.29	0.7915	0.8932	16.51	0.367	130.8	0.7050	0.7960
TATT	17.76	0.247	59.62	0.7953	0.8951	16.79	0.422	118.3	0.7214	0.8135
MARCONet	16.72	0.363	92.11	0.7743	0.8738	16.04	0.397	103.1	0.6638	0.7411
ours	**18.88**	**0.211**	**25.08**	**0.9085**	**0.9247**	**17.49**	**0.336**	**70.59**	**0.8475**	**0.8747**

Table 2. Quantitative comparison for the real-world dataset RealCE [27] with different methods including SRCNN [7], ESRGAN [43], NAFNet [5], TSRN [42], TBSRN [3], TATT [25], MARCONet [23] and our method for ×2 and ×4 blind text image super-resolution.

more than 24 characters or images with severe LR-HR misalignment. Finally, we obtain 1531 LR-HR pairs for real-world testing set.

Compared Methods and Evaluation Metrics. In order to validate the effectiveness of our method, we compare our DiffTIR with the natural image super-resolution methods (*i.e.*, SRCNN [7], ESRGAN [43], and NAFNet [5]) and text image super-resolution methods (*i.e.*, TSRN [42], TBSRN [3], TATT [25], and MARCONet [23]) respectively. For a fair comparison, we revise their implementation to handle ×2 and ×4 image upsampling, and finetune them with CTR-TIR-train dataset. Moreover, we employ 5 metrics to evaluate the performance of the above methods on text image restoration. We adopt the peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS) [51] to evaluate the distance between the restored image and reference image in the image space and feature space, respectively. To further evaluate the realness of the restored image, we employ the Fréchet Inception Distance (FID) [14]. To better evaluate the text fidelity of the restored text image, we employ the word accuracy (ACC), and normalized edit distance (NED) [27]. Particularly, we adopt pre-trained TransOCR [3, 4] as the text recognition model for ACC and NED.

4.2. Quantitative Comparison

We show the quantitative comparison on the synthetic test dataset CTR-TSR-Test and the real-world test dataset RealCE. As shown in Table 1, DiffTSR performs better than

the compared methods in all metrics. It achieves the best PSNR which demonstrates it can accurately reconstruct the HR images. Benefiting from the powerful text image modeling ability, our method shows better performance in LPIPS and FID which indicates higher realness in the restored images. Our method also performs better in terms of ACC and NED which demonstrates it can effectively keep the text fidelity with the text prior provided by TDM. Also please note that our method still shows the best performance on RealCE without any fine-tuning on RealCE training set, as shown in Table 2, indicating its strong generalization performance and powerful modeling ability for real-world text images.

4.3. Qualitative Comparison

The qualitative results of the synthetic dataset CTR-TSR-Test are shown in Figure 4. Most of the super-resolution methods and text super-resolution ones, e.g. SRCNN [7], NAFNet [5], TSRN [42], TBSRN [3], and TATT [25], restore LR images worse than our method which demonstrates the ability of the proposed IDM to generate text images with high realness. With the strong generation ability of GAN, ESRGAN [43] can restore more realistic images (the first result). However, it will also generate artifacts when the degradation is too severe (the second and third results). Even though MARCONet [23] has the capability to restore more visually pleasant text structures because of the codebook, it will generate artifacts (the first result) when the text style is not considered during its training pro-

Figure 4. Qualitative comparison for the synthetic dataset CTR-TSR-Test with different methods including SRCNN [7], ESRGAN [43], NAFNet [5], TSRN [42], TBSRN [3], TATT [25], MARCONet [23] and our method for ×4 super-resolution.
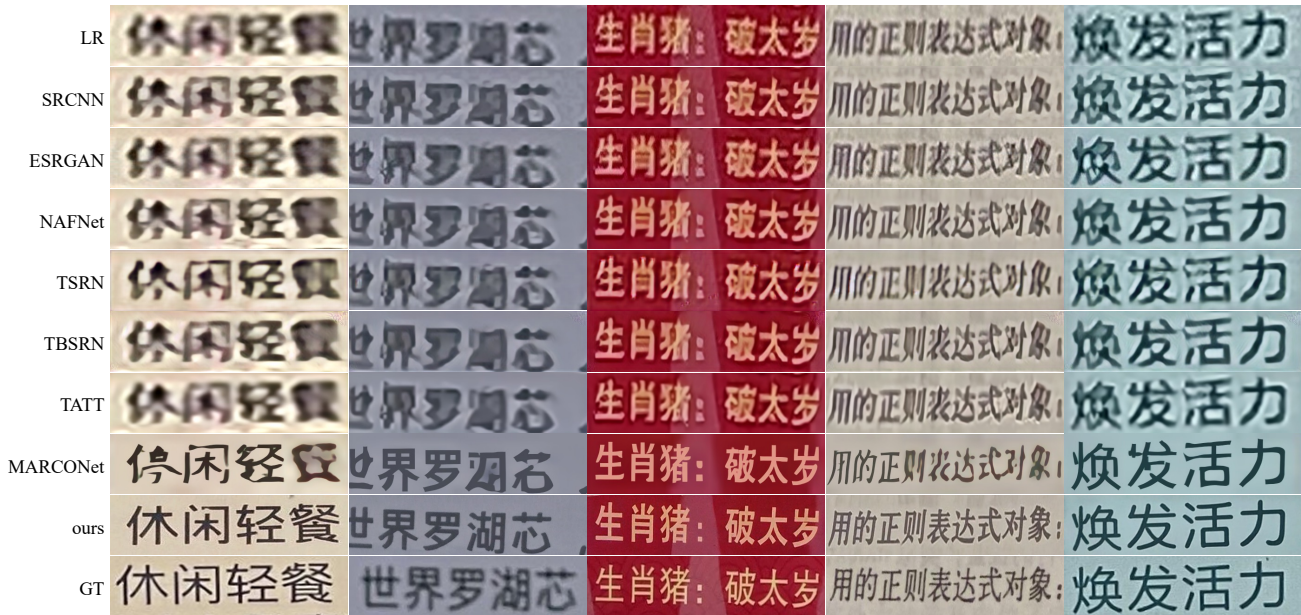

Figure 5. Qualitative comparison for the real-world dataset RealCE [27] with different methods including SRCNN [7], ESRGAN [43], NAFNet [5], TSRN [42], TBSRN [3], TATT [25], MARCONet [23] and our method for ×4 super-resolution.

cess. In addition, the text prior of MARCONet is inaccurate when the degradation is severe (the second example) or encountered with occlusion (the third example). These will make the restoration results of MARCONet with less text fidelity. With the help of the proposed TDM to model the text sequence and MoM to simultaneously optimize IDM and TDM, our method can generate HR images with higher text fidelity.

We also compare different methods based on the real-world dataset RealCE [27]. Note that all the methods are not trained on the training set of RealCE to evaluate the generalization ability when encountered with unknown styles and degradation. The results shown in Figure 5 demonstrate that most of the methods, such as SRCNN [7], ESRGAN [43], NAFNet [5], TSRN [42], TBSRN [3], and TATT [25], can hardly remove degradation in this real-world dataset. Al-

| method | Settings | | | | CTR-TSR-Test / RealCE [27] | | | | |
	IDM	TR	TDM	MoM	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑
exp1	✓	✗	✗	✗	20.61 / 16.86	0.289 / 0.375	13.91 / 74.63	0.6342 / 0.6953	0.6450 / 0.7801
exp2	✓	✓	✗	✗	21.58 / 17.05	0.253 / 0.398	9.968 / 76.24	0.6811 / 0.7241	0.6903 / 0.8183
exp3	✓	✓	✓	✗	21.63 / 17.22	0.233 / 0.340	8.925 / 73.85	0.7412 / 0.7936	0.7530 / 0.8417
ours	✓	✓	✓	✓	**21.85 / 17.49**	**0.231 / 0.336**	**8.482 / 70.59**	**0.8350 / 0.8475**	**0.8471 / 0.8747**

Table 3. Ablation study based on ×4 super-resolution to validate the effectiveness of initial text recognition (TR), TDM, and MoM. For the detailed settings of different methods, please see Sec. 4.4.

though MARCONet [23] can restore HR text image to some extent, it will still generate some inaccurate and unpleasant strokes in the results. With the strong distribution modeling abilities of IDM, the proposed methods can generalize well on real-world scenarios and restore text images with high style realness as well as text fidelity.

4.4. Ablation Study

In this subsection, we validate the effectiveness of different components in the proposed method and the comparison is shown in Table 3 and Figure 6. For 'exp1', it only contains IDM conditioned on the LR image. As shown in the second row of Figure 6, the results look like Chinese characters. However, the generated characters actually do not exist in the Chinese alphabet because the text prior is not considered during the diffusion process. In 'exp2' which is the baseline model in Sec. 3.2, it uses the text recognition (TR) method [3] to predict text sequence and provide text prior to IDM. As a result, 'exp2' can keep the text fidelity better than 'exp1' according to the third row of Figure 6. Whereas, the results are still unsatisfactory when the degradation is too severe and the text character recognition is inaccurate as seen in the orange bounding boxes. 'exp3' uses TDM, whose initial state is provided by TR [3], to predict text sequence and provide text prior to IDM. With the strong text sequence distribution modeling ability from TDM by diffusion, 'exp3' can recognize text more accurately than 'exp2' as shown in the fourth row of Figure 6. But the text characters in the green bounding boxes are still incorrect. This is because TDM in 'exp3' does not utilize the higher-quality image information provided by IDM to recognize more accurate text sequence during the diffusion process. Our method contains MoM module which can provide better text prior for IDM and better image prior for TDM in the diffusion steps. In this way, TDM in our method can correct the mistakenly estimated text sequence with higher-quality image information from IDM. At the same time, IDM can restore text image with higher fidelity which is shown in the fifth row of Figure 6. Similarly, Table 3 shows that the proposed method can achieve consistently better performance with more components considered which demonstrates the effectiveness of TR, TDM, and MoM.
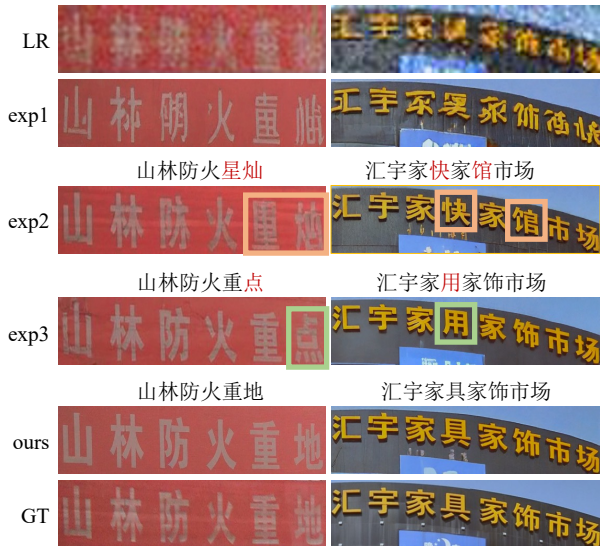

Figure 6. Ablation study to validate the effectiveness of initial text recognition (TR), TDM, and MoM. The text sequences above each image super-resolution result are the recognized text characters used for image super-resolution and the characters in red are the mistakenly estimated ones which will lead the text restoration inaccurate in the orange and green bounding boxes. For the detailed settings of different methods, please see Sec. 4.4.

5. Conclusion

In this paper, we propose to use the diffusion model to solve the blind text image super-resolution problem. As diffusion has a strong ability to model distribution and generate data, the proposed IDM can restore realistic HR text images. At the same time, we also apply another diffusion model (TDM) to model the distribution of text sequence and provide text prior to IDM. In this way, IDM can also generate text images with high text fidelity. At last, we propose MoM to make these two diffusion models appropriately cooperate with each other during the diffusion process. Extensive experiments on synthetic and real-world datasets demonstrate our method can perform better than existing arts based on style realness and text fidelity simultaneously.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *NIPS*, 2019. 2

[2] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 2

[3] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, 2021. 2, 4, 6, 7, 8

[4] Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *arXiv:2112.15093*, 2021. 5, 6

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 6, 7

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NIPS*, 2021. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015. 6, 7

[8] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023. 2

[9] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*. IEEE, 2019. 2

[10] Masato Fujitake. Diffusionstr: Diffusion model for scene text recognition. In *ICIP*, 2023. 4

[11] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv:2210.08933*, 2022. 3

[12] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 2

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 2017. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 2020. 2

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 2

[17] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *NIPS*, 2021. 2, 3, 4, 5

[18] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *NIPS*, 2020. 2

[19] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPR*, 2020. 2

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[21] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NIPS*, 2022. 2

[22] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022. 2

[23] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *CVPR*, 2023. 2, 6, 7, 8

[24] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *CVPR*, 2023. 2

[25] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *CVPR*, 2022. 2, 6, 7

[26] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *TIP*, 2023.

[27] Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for chinese-english scene text image super-resolution. In *ICCV*, 2023. 2, 5, 6, 7, 8

[28] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, 2020. 2

[29] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 2

[30] Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *ECCV*, 2020. 2

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021. 2

[32] Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. Icdar2015 competition on text image super-resolution. In *ICDAR*, 2015. 2

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 5

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour,

Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NIPS*, 2022. 2

[36] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, (4), 2022. 2

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020. 2

[39] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv:2305.07015*, 2023. 2

[40] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 2

[41] Wenjia Wang, Enze Xie, Peize Sun, Wenhai Wang, Lixun Tian, Chunhua Shen, and Ping Luo. Textsr: Content-aware text super-resolution guided by recognition. *arXiv:1909.07113*, 2019. 2

[42] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *ECCV*, 2020. 2, 6, 7

[43] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 6, 7

[44] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 2, 5

[45] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv:2212.00490*, 2022. 2

[46] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020. 2

[47] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2

[48] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *ICCV*, 2023. 3, 4

[49] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 2, 5

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[52] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-stisr: Scene text image super-resolution with triple clues. *IJCAI*, 2022. 2