

RESTORE TEXT FIRST, ENHANCE IMAGE LATER: TWO-STAGE SCENE TEXT IMAGE SUPER-RESOLUTION WITH GLYPH STRUCTURE GUIDANCE

Anonymous authors

Paper under double-blind review

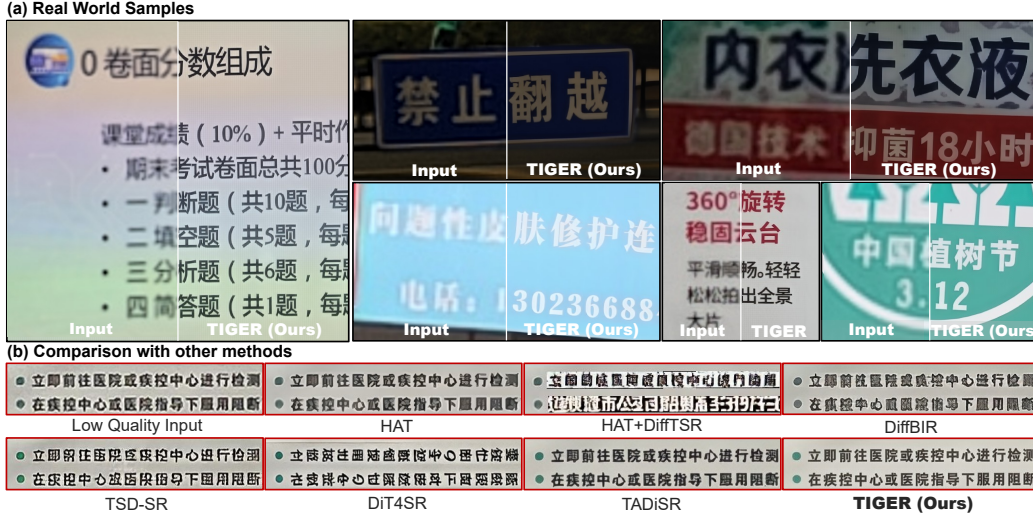


Figure 1: We present **TIGER** (Text–Image Guided supEr-Resolution), a novel framework for scene text super-resolution. Its ‘text-first, image-later’ paradigm ensures accurate glyph restoration and consistently high overall image fidelity and visual quality.

ABSTRACT

Current generative super-resolution methods show strong performance on natural images but distort text, creating a fundamental trade-off between image quality and textual readability. To address this, we introduce **TIGER** (Text–Image Guided supEr-Resolution), a novel two-stage framework that breaks this trade-off through a “text-first, image-later” paradigm. TIGER explicitly decouples glyph restoration from image enhancement: it first reconstructs precise text structures and then uses them to guide subsequent full-image super-resolution. This glyph-to-image guidance ensures both high fidelity and visual consistency. To support comprehensive training and evaluation, we also contribute the **UltraZoom-ST** (UltraZoom-Scene Text), the first scene text dataset with extreme zoom ($\times 14.29$). Extensive experiments show that TIGER achieves **state-of-the-art** performance, enhancing readability while preserving overall image quality.

1 INTRODUCTION

Scene Text Image Super-Resolution is a critical problem in computer vision for its vital role in car navigation (Li & Cui, 2025), scene understanding (Kil et al., 2023; Deshmukh et al., 2024), and document enhancement (Souibgui et al., 2023). It seeks to restore a high-quality super-resolution (SR) image from a degraded low-resolution (LR) input while preserving the correct text meaning. Unlike general image super-resolution for natural scenes, where plausible texture synthesis or detail hallucination is acceptable and does not alter the semantic meaning of the image, visual texts exhibit extremely low tolerance to structural errors. This is especially evident in Chinese, where a minor stroke distortion or omission may entirely alter the meaning (Yu et al., 2023; Li et al., 2023a).

Table 1: Dataset comparison. ‘Non-Latin characters’ indicates inclusion of languages such as Chinese, ‘Multi-lines’ denotes images with multiple text lines, ‘Real-world degradation ($> \times 4$)’ refers to captured LR with zoom beyond $\times 4$. TextZoom (Wang et al., 2020) lacks non-Latin text, CTR (Yu et al., 2021) lacks multi-line text, and Real-CE (Ma et al., 2023) has only mild degradation. UltraZoom-ST covers all, providing a more challenging benchmark.

Content	TextZoom	CTR	Real-CE	UltraZoom-ST (Ours)
Non-Latin characters	✗	✓	✓	✓
Multi-lines	✗	✗	✓	✓
Real-World Degradation ($> \times 4$)	✗	✗	✗	✓

In recent years, with the rapid development of generative models such as diffusion (Ho et al., 2020; Dhariwal & Nichol, 2021; Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Chang et al., 2023; Zhang et al., 2024b), image super-resolution has increasingly relied on their strong generative priors to recover missing details from low-resolution (LR) inputs (Wu et al., 2024b; Yu et al., 2024; Lin et al., 2024; Wu et al., 2024a; Dong et al., 2025a; Duan et al., 2025; Hu et al., 2025). Although these approaches can effectively restore fine-grained natural textures (e.g., grass, leaves), they often distort text regions, turning them into gibberish as shown in Fig.1 (b). Some researchers attempt to address this by processing only text regions (Zhang et al., 2024a; Li et al., 2023a). While these methods improve readability notably, the absence of global background constraints often introduces style inconsistencies and block artifacts between text and background. The unsatisfactory performance of the present methods can be attributed to several reasons. Firstly, there is a deficiency in scene text paired data. Current scene text super-resolution datasets (Wang et al., 2020; Yu et al., 2021; Ma et al., 2023) are limited in degradation and focused on textline annotations, making it difficult for models to learn the mapping from low- to high-resolution text. Moreover, current methods primarily enhance overall image quality but fail to capture fine-grained glyph structures. This limitation is amplified for Chinese characters, due to their complex glyph designs and relatively low saliency in images. Consequently, models tend to under-represent and collapse glyphs into oversimplified, averaged forms, producing overlapping or distorted characters. Ultimately, current approaches face a persistent trade-off between maintaining readability and ensuring high image quality. *Our key observation is that these goals need not be antagonistic if text and non-text are explicitly handled differently. Text structures can be reconstructed with dedicated mechanisms and then used to guide full-image restoration, ensuring coherent style without artifacts.*

Building on this insight, we introduce **TIGER** (Text–Image Guided supEr-Resolution), a progressive two-stage paradigm for scene text super-resolution built on the principle of “*restoring text structure first, enhancing the whole image later.*” Unlike methods that rely on a single generative prior, TIGER explicitly separates the treatment of text and non-text regions: a diffusion-based local text refiner focuses on reconstructing fine-grained stroke geometry in text regions, ensuring glyph fidelity and structural consistency. The recovered text structures are then injected as conditional guidance into the subsequent full-image restoration stage, steering global super-resolution to harmonize text and background while suppressing artifacts and preserving overall visual quality.

Moreover, to address the issue of data scarcity and enable comprehensive evaluation (Wang et al., 2020; Ma et al., 2023), we introduce the **UltraZoom-ST** (UltraZoom-Sence Text) benchmark dataset, the first scene text dataset that contains extreme zooming mode ($\times 14.29$), providing extra challenging scenarios for the field. Content differences are outlined in Table 1. It includes high-quality 5,036 LR–HR pairs captured at multiple focal lengths (ranging from 14mm to 200mm), with 49,675 text lines in total. Each pair comes with detailed annotations, including detection boxes and text transcripts, to support both training and evaluation. The dataset includes diverse scenarios such as shop signs, posters, and documents. It also covers varying lighting conditions, including daylight, indoor lighting, and nighttime, offering a challenging yet practical benchmark for the field. To ensure reliable alignment across extreme focal lengths, we adopt a coarse-to-fine cascade alignment strategy: coarse alignment is performed via global geometric transformation, and residual misalignments are corrected through local refinement. This process achieves accurate LR–HR alignment, thereby ensuring both the quality of the dataset and the reliability of our experiments. The main contributions are summarized as follows:

1. We propose TIGER, the first two-stage scene text super-resolution framework that introduces a novel ‘*text-first, image-later*’ paradigm to decouple glyph restoration from image enhancement, improving both readability and visual quality.

2. We introduce UltraZoom-ST, the first scene text benchmark with extreme zooming ($\times 14.29$), offering well-aligned, richly annotated LR–HR pairs for comprehensive evaluation under challenging and diverse real-world conditions.
3. Extensive experiments on both Real-CE and UltraZoom-ST show the proposed method outperforms prior state-of-the-art models, particularly in preserving text structure fidelity.

2 RELATED WORKS

Real-World Image Super-Resolution. Real-world image super-resolution (Real-SR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs degraded under uncontrolled real-world conditions. Early GAN-based methods, such as BSRGAN (Zhang et al., 2021) and Real-ESRGAN (Wang et al., 2021), synthesize degradations through random combinations of known distortions and use adversarial training for restoration. While they generate natural-looking images, their instability and insensitivity to fine details limit their ability to recover structural elements, especially text. Recent work introduces diffusion models into Real-SR (Yue et al., 2023; Wang et al., 2024b), improving perceptual quality but still struggling with complex structures like scene text. StableSR (Wang et al., 2024a) and DiffBIR (Lin et al., 2024) apply ControlNet (Zhang et al., 2023) to condition generation on LR inputs, while PASD (Yang et al., 2024) and SeeSR (Wu et al., 2024b) incorporate high-level semantics to enhance fidelity. SUPIR (Yu et al., 2024) scales training with large image-text pairs and introduces degradation-robust encoders. Other approaches, including OSEDiff (Wu et al., 2024a) and TSD-SR (Dong et al., 2025b), directly apply the diffusion process on LR images and distill models for one-step sampling. DiT-SR (Cheng et al., 2025), Dream-Clear (Ai et al., 2024), and DiT4SR (Duan et al., 2025) adopt diffusion transformers (DiT) for Real-SR. Despite enhancing perceptual fidelity, these methods often overlook accurate text structures. TADiSR (Hu et al., 2025) addresses Chinese scene text super-resolution by using Kolrs as the base model and aggregating cross-attention maps for text structure supervision. However, constrained by the resolution of cross-attention, it struggles to restore small or severely degraded text.

Text Image Super-Resolution. Text image super-resolution (Text-SR) restores textual content from cropped images containing isolated words or text lines. Early methods (Dong et al., 2015) apply general SR architectures such as SRCNN (Dong et al., 2014) to enhance OCR performance on LR. TextSR (Wang et al., 2019) introduces GANs with text recognition loss, while PlugNet (Mou et al., 2020) and STT (Chen et al., 2021) jointly train SR and recognition modules for more discriminative features. TSRN (Wang et al., 2020) introduces the TextZoom dataset and incorporates an edge-aware module to preserve text details, while TATT (Ma et al., 2022) uses a global attention module to handle irregular text layouts. Recent methods leverage stronger generative priors. MARCONet (Li et al., 2023a) employs StyleGAN priors and a glyph structure codebook for realistic text reconstruction. DiffTSR (Zhang et al., 2024a) employs latent diffusion to separately denoise text and text-image components. Despite progress in text structure restoration, existing methods lack global background constraints, causing style inconsistencies and block artifacts between text and background.

3 METHODOLOGY

3.1 ARCHITECTURE OVERVIEW

Current image super-resolution methods focus on enhancing overall image quality but often fail to accurately preserve glyph structures, leading to distorted text in super-resolved images, as shown in Fig. 1 (b). Conversely, text image super-resolution methods improve text readability but do not retain the global semantic information of the background, causing incoherence between text and non-text regions. To leverage the strengths of both approaches, we introduce the TIGER framework. As illustrated in Fig. 2, the TIGER framework is composed of 2 stages, the **Text Restoration stage** and the **Text Enhancement stage**. In the text restoration stage, we extract the text regions from the LR input $x_L \in \mathbb{R}^{H \times W \times C}$ and feed them into the glyph structure restoration model to restore the text structure based on the text region of the LR input and the predicted text. We then reassemble the text structures to their original positions to obtain a text mask $\hat{x}_m \in \mathbb{R}^{H \times W \times C}$. In the image enhancement stage, the text mask and LR input are then processed by a ControlNet-like network to obtain the enhanced SR output $\hat{x}_H \in \mathbb{R}^{H \times W \times C}$. The following sections detail each stage.

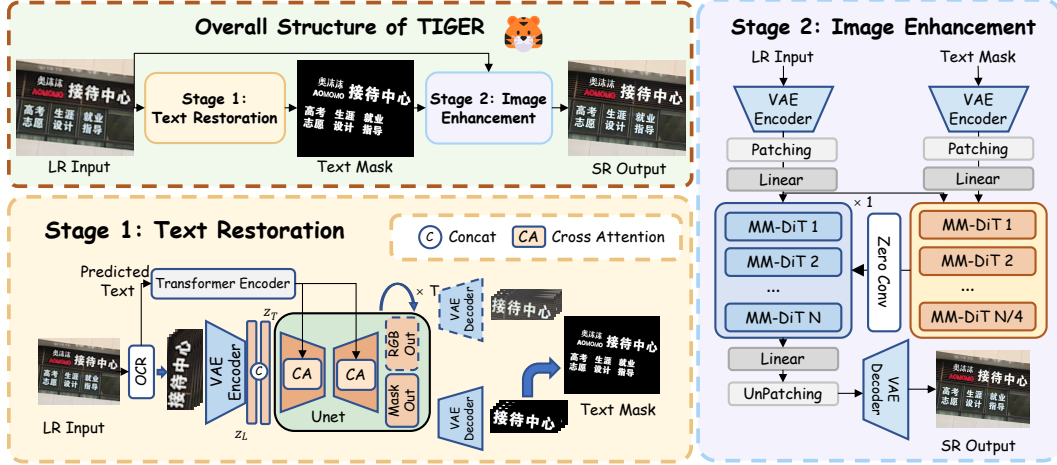


Figure 2: The framework of TIGER, which includes the Text Restoration stage (stage 1) and the Image Enhancement stage (stage 2). Stage 1 refines text regions to recover accurate glyph structures. Stage 2 uses these structures to guide full-image restoration for coherent text and background.

3.2 TEXT RESTORATION STAGE

Text Restoration Pipeline. Existing methods like Hi-SAM (Ye et al., 2024) work well on clean high-resolution text but fail on the incomplete, distorted text in low-resolution images. To address this, we propose a region-level text restoration pipeline shown in Fig. 2 built on Rombach et al. (2022) to recover glyph structures $x_m \in \mathbb{R}^{H \times W \times C}$. An OCR detector first localizes text regions $\{x_L^0, \dots, x_L^{N-1}\}$ and extracts their contents $\{y^0, \dots, y^{N-1}\}$ as semantic conditions. Each region $\tilde{x}_L$ is encoded by VAE (Kingma & Welling, 2013) into $\tilde{z}_L \in \mathbb{R}^{h \times w \times c}$, concatenated with noise $z_T \in \mathbb{R}^{h \times w \times 2c}$, and iteratively denoised by a UNet ϵ_θ into two branches: $z_{t-1}^{RGB} \in \mathbb{R}^{h \times w \times c}$ for appearance and $z_{t-1}^m \in \mathbb{R}^{h \times w \times c}$ for structure, which are merged as z_{t-1} . The text content y is embedded into c_{te} and fused into z_t via cross-attention to guide structure recovery. After T denoising steps, the mask branch output z_0^m is decoded into $\tilde{x}_m$, and all restored regions are assembled into the final text mask $\hat{x}_m$. Focusing on real text reconstruction enables the model to capture glyph structures without non-text interference and reduces sensitivity to text saliency.

Training Strategy. The scarcity of real-world segmentation masks for degraded text forces reliance on synthetic data, yet its artificial degradation limits generalization. Annotating real degraded samples, however, is labor-intensive and costly. To marry synthetic precision with real-world degradation, we propose a two-phase training strategy. In Phase 1, both synthetic and real data are included in training. This allows the model to generate text masks from the LR image and capture real-world degradation patterns. However, noisy masks in real data degrade mask quality. To address this, we freeze the RGB out block and the mask out block and train the UNet with only synthetic data to refine the quality of output text masks in Phase 2. The training objective is described as follows:

$$\mathcal{L} = \lambda_{td} \mathcal{L}_{td} + \lambda_{Seg} \mathcal{L}_{Seg}, \quad (1)$$

where $\mathcal{L}_{td}$, $\mathcal{L}_{Seg}$, λ_{td} , and λ_{Seg} denote the text-control diffusion loss, segmentation-oriented loss, and their corresponding hyperparameters. The text-control diffusion loss is formulated as follows:

$$\mathcal{L}_\epsilon = \mathbb{E}_{z_0, \tilde{z}_L, c_{te}, t, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(z_t, \tilde{z}_L, c_{te}, t)\|_2^2]. \quad (2)$$

To improve the quality of the text mask, we propose a segmentation-oriented loss. Let ϵ_t denote the noise predicted by the denoiser network ϵ_θ . Following Ho et al. (2020), z_0^m can be estimated by combining the time step t with the noisy latent image z_t^m . This estimate is subsequently passed through the VAE decoder to obtain an approximate reconstruction of the original input text mask, denoted as x_0^m . In this way, text mask generation can be supervised at the pixel level. We combine Mean Squared Error (MSE), Focal, and Dice Losses, commonly used in segmentation tasks (Ye et al., 2024), to compare x_0^m with the original image x_0^m , as shown in the following equation:

$$\mathcal{L}_{Seg} = \|x_0^m - x_0^m\|_2^2 + \lambda_{Focal} \text{FocalLoss}(x_0^m, x_0^m) + \lambda_{Dice} \text{DiceLoss}(x_0^m, x_0^m), \quad (3)$$

where λ_{Focal} and λ_{Dice} are their respective balancing coefficients. This combination forms a novel text-first paradigm that equips the model with both structural accuracy and real-world generalization, laying the foundation for the subsequent full-image super-resolution stage.

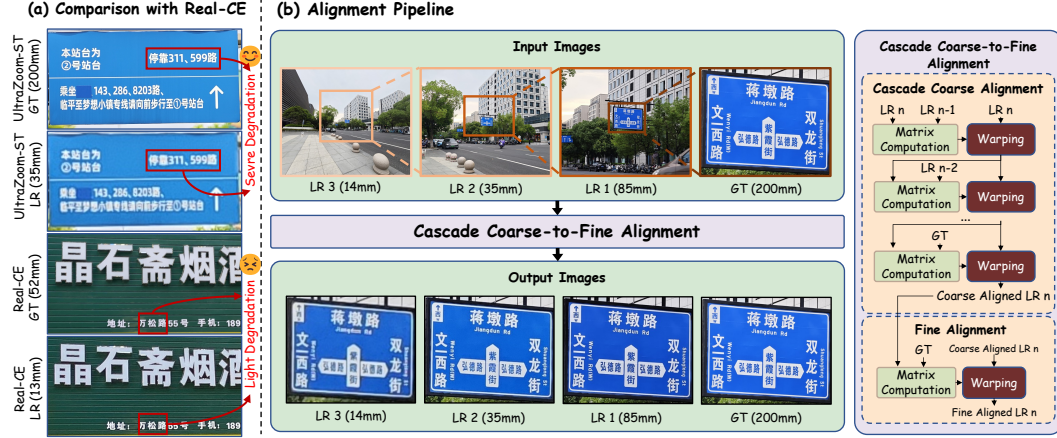


Figure 3: Overview of UltraZoom-ST. (a) Real-CE LRs show only mild degradation (red box), while UltraZoom-ST LRs exhibit stronger degradation (red box), enabling a more comprehensive evaluation. (b) Coarse-to-fine alignment: images are sorted by focal length, each warped to the next higher-focal neighbor using an estimated homography matrix, then refined to the 200 mm GT.

3.3 IMAGE ENHANCEMENT STAGE

Image Enhancement pipeline. To effectively take advantage of the generated glyph structure $\hat{x}_m$ and enhance the quality of LR x_L , we adopt a ControlNet (Zhang et al., 2023) network ϵ_ϕ . As illustrated in Fig. 2, after getting their latent representation $\hat{z}_m$ and z_L , we use the network to denoise the z_L at t timestep, using the null-text embedding c_{Null} . The output super-resolved latent representation $\hat{z}_H$ will be denoted as:

$$\hat{z}_H = z_L - \sigma_t \epsilon_\phi(z_L, \hat{z}_m, t, c_{Null}), \quad (4)$$

where σ_t is a scalar determined by the predefined diffusion time step. Together with the text restoration stage, this image enhancement pipeline forms a progressive ‘text-first, image after’ paradigm; our design injects structure-aware control into the generative process, allowing the network to enhance global image quality without eroding the recovered text structures.

Training Strategy. For the quality of image super-resolution, we constrain the reconstruction loss between the predicted high-resolution image $\hat{x}_H$ decoded from $\hat{z}_H$ by ϵ and the ground truth high-resolution image x_H using a weighted sum of MSE and LPIPS losses:

$$\mathcal{L}_{img} = \lambda_{l2} \|x_H - \hat{x}_H\|_2^2 + \lambda_{LPIPS} LPIPS(x_H, \hat{x}_H), \quad (5)$$

where λ_{l2} and λ_{LPIPS} are balancing coefficients for different loss terms. To enhance glyph control and emphasize the glyph structure, we extract the text boundaries using Sobel operators (Roberts & Mullis, 1987). This edge loss is expressed as:

$$\mathcal{L}_{edge} = \|\text{Sobel}(x_H) - \text{Sobel}(\hat{x}_H)\|_2^2. \quad (6)$$

We combine the two loss terms and use λ_{edge} as the balancing coefficient in the complete loss function for stage 2:

$$\mathcal{L} = \mathcal{L}_{img} + \lambda_{edge} \mathcal{L}_{edge}. \quad (7)$$

The combination of glyph-aware ControlNet conditioning and edge-constrained training objectives enables our model to preserve stroke integrity while harmonizing text and background appearance.

4 DATASET AND BENCHMARK

Existing datasets like Ma et al. (2023) offer only subtle degradation, as shown in Fig. 3 (a), making them insufficient for evaluating model robustness. To address the lack of both challenging and publicly available datasets tailored for scene text image super-resolution, especially for Chinese text, we introduce **UltraZoom-ST**, a challenging real-world benchmark. It is collected using ViVO X200 Ultra equipped with four fixed focal lengths (14 mm, 35 mm, 85 mm, 200 mm), enabling image

Table 2: Evaluation results on image quality. Numbers in **bold** indicate the best performance, and underscored numbers indicate the second best. TIGER (Ours) performs best on image quality.

Methods	Real-CE					UltraZoom-ST				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$
Real-ESRGAN	22.30	0.787	0.239	0.188	53.60	23.99	0.790	0.248	0.194	30.60
HAT	23.61	<u>0.830</u>	0.214	0.176	51.16	<u>25.17</u>	0.815	0.249	0.198	30.12
MARCONet	21.89	0.785	0.238	0.150	52.97	22.13	0.768	0.306	0.205	34.28
SeeSR	23.59	0.822	<u>0.195</u>	0.169	43.75	23.64	0.788	0.219	0.184	26.73
SupIR	21.78	0.723	<u>0.310</u>	0.198	44.94	23.62	0.754	0.308	0.207	27.91
DiffTSR	22.10	0.768	0.278	0.168	44.91	22.41	0.767	0.300	0.189	31.25
DiffBIR	22.44	0.747	0.260	0.201	46.44	23.67	0.724	0.262	0.197	23.10
OSDiff	21.86	0.771	0.197	<u>0.127</u>	41.00	25.07	<u>0.819</u>	<u>0.201</u>	0.169	<u>20.53</u>
DreamClear	22.47	0.772	0.216	<u>0.157</u>	<u>38.97</u>	24.10	<u>0.773</u>	0.238	0.191	21.75
TSD-SR	21.43	0.754	0.220	0.175	47.21	22.79	0.757	0.207	0.194	24.08
DiT4SR	20.54	0.764	0.268	0.186	49.79	23.16	0.767	0.215	<u>0.159</u>	20.58
TADiSR	23.83	0.790	0.286	0.154	44.42	24.61	0.796	0.203	<u>0.160</u>	36.61
TIGER (Ours)	24.12	0.839	0.164	0.125	38.72	25.48	0.830	0.196	0.156	20.01

Table 3: Evaluation of image quality and text accuracy. Metrics with cr are computed only on text regions. TIGER (ours) achieves the best performance in both.

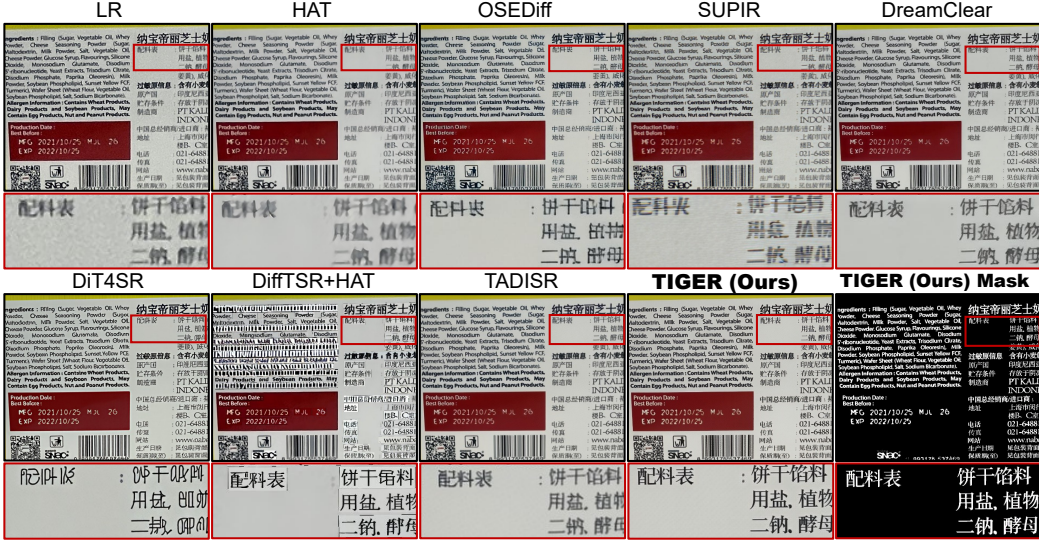
Methods	Real-CE					UltraZoom-ST				
	PSNR $_{cr}$	SSIM $_{cr}$	LPIPS $_{cr}$	DISTS $_{cr}$	OCR-A	PSNR $_{cr}$	SSIM $_{cr}$	LPIPS $_{cr}$	DISTS $_{cr}$	OCR-A
Real-ESRGAN	21.71	0.824	0.257	0.224	0.560	21.25	0.786	0.302	0.266	0.371
HAT	23.03	0.862	0.249	0.239	0.566	<u>22.18</u>	0.813	0.291	0.276	0.379
MARCONet	20.68	<u>0.786</u>	0.267	0.223	0.550	<u>17.48</u>	<u>0.690</u>	0.529	0.366	0.334
SeeSR	22.74	0.844	0.237	0.212	0.374	20.45	0.767	0.297	0.241	0.191
SupIR	19.96	0.778	0.326	0.263	0.278	20.10	0.756	0.344	0.287	0.236
DiffTSR	20.46	0.818	0.275	0.239	0.441	17.00	0.665	0.452	0.313	0.317
DiffBIR	21.20	0.792	0.278	0.224	0.376	20.70	0.757	0.302	0.245	0.266
OSDiff	19.38	0.768	0.269	0.208	0.268	21.43	0.798	0.276	0.263	0.300
DreamClear	22.74	0.845	<u>0.190</u>	<u>0.168</u>	0.502	20.50	0.768	0.317	0.276	0.214
TSD-SR	19.33	0.763	<u>0.279</u>	0.234	0.349	19.42	0.748	0.296	0.249	0.266
DiT4SR	17.95	0.738	0.317	0.244	0.292	19.73	0.760	0.273	0.216	0.237
TADiSR	23.39	0.855	0.253	0.258	0.647	21.59	0.799	0.360	<u>0.336</u>	0.384
TIGER (Ours)	23.43	0.864	0.173	0.167	0.673	22.22	0.814	0.228	0.212	0.430

pairs with extreme $\times 14.29$ zoom. It includes diverse scenes—street views, book covers, advertisements, menus, and posters—captured under varied lighting, with all text lines manually annotated. However, such extreme zoom introduces severe misalignment that breaks pixel-wise optimization methods like Cai et al. (2019), originally designed for moderate $\times 2$ to $\times 4$ zoom. To overcome this, we design a **Cascade Coarse-to-fine Alignment** pipeline shown in Fig. 3 (b): images are first sorted by focal length, and each low-quality image is sequentially aligned to its next higher-focal neighbor using either optimization-based registration or feature-based extractors (e.g., Lowe (1999); Künzel et al. (2025)), then refined against the 200 mm ground truth for precise alignment.

Through meticulous annotating and aligning, we obtained a total of 5,036 image pairs, with 49,675 lines of text. We set images under 200mm focal lengths as GT, and obtain 1,439, 1,798, and 1,799 pairs for $\times 14.29$, $\times 5.71$, and $\times 2.35$ zooming modes, respectively. Among them, we randomly select 470, 589, and 581 pairs for evaluation under each zooming mode. Each image pair contains one or more text lines. These evaluation sets enable us to evaluate the performance of models under more complex and challenging scenarios.

For the UltraZoom-ST-benchmark, we utilize 5 evaluation metrics to examine the quality of image super-resolution on the scale of the full image. Firstly, we adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al., 2019), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), Deep Image Structure and Texture Similarity (DISTS) (Ding et al., 2020), and Fréchet Inception Distance (FID) (Heusel et al., 2017) for evaluation of image quality. Specifically, PSNR and SSIM are computed in the pixel space to quantify low-level reconstruction fidelity, LPIPS and DISTS are computed in the feature space to assess perceptual similarity, and FID is employed to evaluate the distributional discrepancy between generated and real images. However, these metrics do not directly reflect the quality or accuracy of

(a) Visual comparison on Real-CE. Our decoupling strategy yields superior performance on small text.



(b) Visual comparison on UltraZoom-ST, showing robustness to severe degradation with 2-phase training.

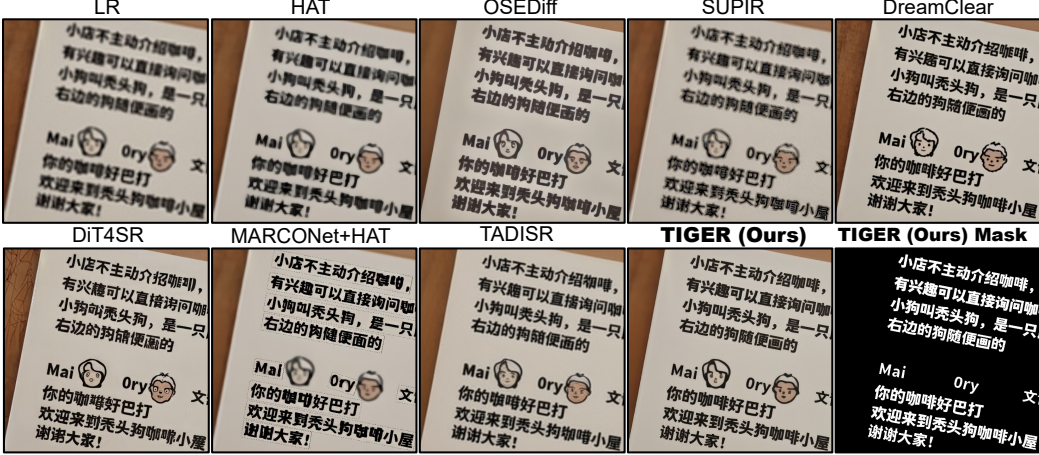


Figure 4: Qualitative Evaluation on Real-CE and UltraZoom-ST.

super-resolved text. To address this, we crop annotated text regions from the images and compute PSNR, SSIM, LPIPS, and DISTS on these cropped regions, denoted as $PSNR_{cr}$, $SSIM_{cr}$, $LPIPS_{cr}$, and $DISTS_{cr}$. For text accuracy, we apply an OCR model to recognize the text and compare its outputs with the ground-truth annotations using the Levenshtein ratio (Yujian & Bo, 2007):

$$OCR-A = (\text{Len}(s_{pred}) + \text{Len}(s_{gt}) - \text{Dist}(s_{pred}, s_{gt})) / (\text{Len}(s_{pred}) + \text{Len}(s_{gt})), \quad (8)$$

where s_{pred} stands for the predicted text sequence, s_{gt} denotes the annotated ground-truth text sequence, and $\text{Dist}(\cdot, \cdot)$ is the Levenshtein distance between the text sequences.

5 EXPERIMENT

5.1 IMPLEMENTATION DETAILS

We combine synthetic data (built upon LSDIR (Li et al., 2023b) with text rendered via LBTS (Tang et al., 2023) and Real-ESRGAN (Wang et al., 2021) degradation) and real paired data from Real-CE (Ma et al., 2023) and UltraZoom-ST. Misaligned pairs in Real-CE are filtered and reannotated, yielding 337 training and 188 testing pairs. Stage 1 uses cropped text regions and is trained using an IDM-based architecture (Zhang et al., 2024a) with combined segmentation and reconstruction losses. Stage 2 is based on Stable Diffusion 3.5 (Esser et al., 2024) and employs a tile-based inference strategy (Yu et al., 2024; Hu et al., 2025); it is pretrained on synthetic data and fine-tuned on real data. More detailed information can be found in the Appendix A.3.

Table 4: Validation on UltraZoom-ST (UZST). Finetuning improves OCR-A, proving effectiveness.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	OCR-A $\uparrow$
OSDiff w/o UZST	23.50	0.791	0.197	0.162	24.30	0.228
OSDiff w/ UZST	25.07	0.819	0.201	0.169	20.53	0.300
DiT4SR w/o UZST	22.58	0.754	0.252	0.196	26.93	0.193
DiT4SR w/ UZST	23.17	0.767	0.215	0.159	20.58	0.237
Ours w/o UZST	22.96	0.782	0.341	0.253	33.78	0.400
Ours w/ UZST	25.48	0.830	0.196	0.156	20.01	0.430

5.2 COMPARISON RESULTS

5.2.1 QUANTITATIVE RESULTS

We evaluate existing competing methods, including GAN-based image super-resolution approaches such as Real-ESRGAN (Wang et al., 2021) and HAT (Chen et al., 2023); diffusion-based approaches such as SeeSR (Wu et al., 2024b), SupIR (Yu et al., 2024), DiffBIR (Lin et al., 2024), OSDiff (Wu et al., 2024a), DreamClear (Ai et al., 2024), TSD-SR (Dong et al., 2025b), and DiT4SR (Duan et al., 2025); as well as text-focused reconstruction methods such as MARCONet (Li et al., 2023a), DiffTSR (Zhang et al., 2024a), and TADiSR (Hu et al., 2025). Evaluations are performed on the Real-CE Benchmark (Ma et al., 2023) and the benchmark described in Sec. 4. To ensure fairness, we fine-tune the released pre-trained models on the training sets of both benchmarks using the official code when available. Following Hu et al. (2025), we integrate the outputs of MARCONet and DiffTSR with HAT-generated results to simulate real-world application scenarios and enable comprehensive full-image evaluation. We evaluate Real-CE at the hardest difficulty level ($\times 4$), while UltraZoom-ST is evaluated across all difficulty levels—with average results reported here. Detailed per-level evaluations appear in the Appendix A.4.

As shown in Table 2 and Table 3, TIGER outperforms competing methods on both the Real-CE and UltraZoom-ST benchmarks in terms of image quality and text accuracy. TADiSR, limited by the resolution of its cross-attention mechanism, performs poorly under severe degradation and with small text. DiffTSR and MARCONet fail to effectively handle text backgrounds and struggle with text regions that have a large width-to-height aspect ratio. In contrast, our method restores images with high quality and fine-grained text, owing to the decoupling strategy. It achieves OCR-A scores above 0.67 and 0.43, and demonstrates overall superiority in both pixel-level and perceptual accuracy.

5.2.2 QUALITATIVE RESULTS

Fig. 4 shows visual comparisons on Real-CE and UltraZoom-ST. Real-world degradations cause noise, blurred strokes, and distortions in the input images. GAN-based methods reduce noise but fail to fix text structure, while diffusion-based SR methods, though powerful, lack text-structure guidance and often distort strokes, sometimes making text unreadable.

Methods tailored for text image super resolution (e.g., MARCONet and DiffTSR) exhibit limitations in processing long text sequences, often generating distorted or semantically meaningless results. They also lack global semantic guidance, making it hard to blend text regions with backgrounds. As seen in Fig. 4 (a), DiffTSR+HAT (red box) produces text colors inconsistent with the original LR and SR output of HAT. In Fig. 4 (b), it even distorts the face near the word “Ory”. TADiSR works well on lightly degraded images but fails under severe degradation and performs poorly on small text due to cross-attention mask resolution limits. Our method overcomes these issues, restoring high-quality glyph structures even in challenging cases.

5.3 ABLATION STUDY

Effectiveness of the training set. Table 4 reports consistent OCR-A improvements across different architectures, including the UNet-based OSDiff and the DiT-based DiT4SR. This demonstrates that even models without explicit mechanisms for text structure modeling can still learn the mapping from low- to high-resolution text using our dataset, further validating its effectiveness. Due to their lack of a dedicated mechanism for text structure modeling, they cannot outperform our methods.

Table 5: Ablation study on Real-CE with stage 2 fixed as the baseline. From top to bottom, we compare the performance of using text masks rendered with a **standard font**, extracted using **SAM-TS**, reconstructed with **latent diffusion model** conditioned on LR, and reconstructed with our **text restoration pipeline**. Our pipeline faithfully restores glyph structures, yielding the highest accuracy.

Text Mask Settings	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	OCR-A $\uparrow$
Std Font Guidance	23.31	0.786	0.249	0.164	41.83	0.553
SAM-TS Extraction	23.91	0.835	0.212	0.147	41.22	0.579
LDM Guidance	22.42	0.798	0.222	0.147	43.08	0.601
TIGER (Ours)	24.12	0.839	0.164	0.125	38.72	0.673

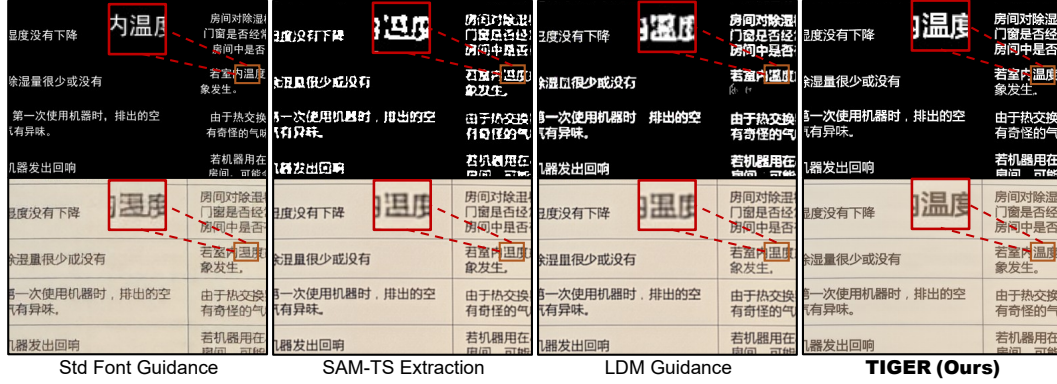


Figure 5: Qualitative Results of Ablation Study with stage 2 fixed as the baseline.

Effectiveness of the TIGER component. We set stage 2 as the fixed baseline for image restoration and validate the effectiveness of our framework through 3 ablated variants:

- **Standard Font Guidance.** As shown in Fig. 5 (first column), the standard font, while structurally correct, provides weak guidance due to style and position mismatches. As a result, the restored text exhibits poor accuracy compared to TIGER, with the OCR-A dropping from 0.671 to 0.553, as shown in Table 5.
- **SAM-TS Extraction Hu et al. (2025); Ye et al. (2024).** While this improves visual quality by better aligning output masks with the LR image, SAM-TS can only extract distorted structures and can’t recover degraded text (Fig. 5, second column), leading to suboptimal performance compared with TIGER, with OCR-A dropping from 0.671 to 0.579.
- **LDM Guidance.** This variant uses a latent diffusion model conditioned on LR to restore text structure, achieving better accuracy by reconstructing or compensating for lost text. However, without separate learning from synthetic and real-world data, it struggles with high-quality text structures in real-world scenarios. The masks remain noisy under real-world degradations (Fig. 5, third column), leading to poorer accuracy compared to TIGER, with OCR-A dropping from 0.671 to 0.601 as shown in Table 5.

Our method leverages RGB output to jointly learn from real-world degradations and accurate synthetic masks, producing high-quality masks and achieving the highest accuracy overall.

6 CONCLUSION

In this paper, we delve into the extensively researched problem of scene text super-resolution. To address this challenge, we propose a novel approach called TIGER, which decouples the glyph structure restoration and image enhancement. For the restoration of text structure, we propose the text restoration pipeline, which enables a 2-phase training strategy to fully take advantage of both synthetic and real-world data. For the image enhancement, we propose the image enhancement pipeline, which effectively utilizes the glyph structure to restore both text fidelity and image details coherently in full-image super-resolution. In terms of training data and benchmark, we present the UltraLens-ST, the first real-world scene text dataset that contains extreme zooming mode, offering more challenging scenarios for the field. Extensive experiments on Real-CE and UltraLens-ST demonstrate the superiority of TIGER over existing methods.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have made the following efforts: (1) We will release our code and dataset. (2) We provide implementation details in Sec. 5.1 and Appendix A.3, including the training process and selection of hyper-parameters. (3) We provide details on evaluation metrics and dataset preparation in Sec. 4 and Appendix A.2, and the code and data will be made available along with it.

ETHICS STATEMENT

This work focuses on improving scene text image super-resolution to support beneficial applications such as enhancing accessibility, document restoration, and navigation assistance. However, we acknowledge potential risks, including misuse in privacy-sensitive contexts (e.g., recovering text from personal or social media images) or unintended deployment in surveillance. To mitigate such risks, users are encouraged to combine our methods with privacy-preserving techniques such as watermarking or selective inpainting. Our UltraZoom-ST dataset was collected from public, non-sensitive scenes under ethical guidelines, without including personally identifiable or private information. We believe the societal benefits of improved text image restoration outweigh potential risks, provided that the technology is applied responsibly.

REFERENCES

- Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.
- 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3086–3095, 2019.
- Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4055–4075. PMLR, 2023. URL <https://proceedings.mlr.press/v202/chang23b.html>.
- Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, pp. 12026–12035, 2021.
- Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pp. 22367–22377, 2023.
- Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2455–2463, 2025.
- Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. URL <https://arxiv.org/abs/2507.05595>.
- Gayatri Deshmukh, Onkar Susladkar, Dhruv Makwana, Sparsh Mittal, et al. Textual alchemy: Co-former for scene text understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2931–2941, 2024.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

- Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 44(5):2567–2581, 2020.
- Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pp. 184–199, 2014.
- Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. Boosting optical character recognition: A super-resolution approach. *arXiv preprint*, 2015.
- Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: one-step diffusion with target score distillation for real-world image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 23174–23184. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.02158. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dong_TSD-SR_One-Step_Diffusion_with_Target_Score_Distillation_for_Real-World_Image_CVPR_2025_paper.html.
- Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23174–23184, 2025b.
- Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- 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. *NeurIPS*, 30, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Qiming Hu, Linlong Fan, Yiyan Luo, Yuhang Yu, Xiaojie Guo, and Qingnan Fan. Text-aware real-world image super-resolution via diffusion model with joint segmentation decoders. *arXiv preprint arXiv:2506.04641*, 2025.
- Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. Prestu: Pre-training for scene-text understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15270–15280, 2023.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Johannes Künzel, Anna Hilsmann, and Peter Eisert. RIPE: Reinforcement Learning on Unlabeled Image Pairs for Robust Keypoint Extraction. *arXiv*, 2025.
- Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight OCR system. *CoRR*, abs/2206.03001, 2022. doi: 10.48550/ARXIV.2206.03001. URL <https://doi.org/10.48550/arXiv.2206.03001>.
- Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=Vv76fCYffN>.

- Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *CVPR*, pp. 10103–10113, 2023a.
- Yawei Li, Kai Zhang, Jingyun Liang, Jiezhong Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdrr: A large scale dataset for image restoration. In *CVPRW*, pp. 1775–1787, 2023b.
- Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pp. 430–448, 2024.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999.
- Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *CVPR*, pp. 5911–5920, 2022.
- Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for chinese-english scene text image super-resolution. In *ICCV*, pp. 19452–19461, 2023.
- 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*, pp. 158–174, 2020.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- Richard A Roberts and Clifford T Mullis. *Digital signal processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.
- 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. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- Mohamed Ali Souibgui, Sanket Biswas, Andres Mafla, Ali Furkan Biten, Alicia Fornés, Yousri Kessentini, Josep Lladós, Lluís Gomez, and Dimosthenis Karatzas. Text-diae: a self-supervised degradation invariant autoencoder for text recognition and document enhancement. In *proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2330–2338, 2023.
- Zhengmi Tang, Tomo Miyazaki, and Shinichiro Omachi. A scene-text synthesis engine achieved through learning from decomposed real-world data. *IEEE Transactions on Image Processing*, 32: 5837–5851, 2023.
- Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=ezBH9WE9s2>.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 132(12):5929–5949, 2024a.
- 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 preprint*, 2019.
- 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*, pp. 650–666, 2020.

- Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pp. 1905–1914, 2021.
- Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, pp. 25796–25805, 2024b.
- Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 37:92529–92553, 2024a.
- Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pp. 25456–25467, 2024b.
- Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pp. 74–91, 2024.
- Maoyuan Ye, Jing Zhang, Juhua Liu, Chenyu Liu, Baocai Yin, Cong Liu, Bo Du, and Dacheng Tao. Hi-sam: Marrying segment anything model for hierarchical text segmentation. *TPAMI*, 2024.
- Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, pp. 25669–25680, 2024.
- Haiyang Yu, Jingye Chen, Bin Li, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *arXiv preprint arXiv:2112.15093*, 2021.
- Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11943–11952, 2023.
- Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 36:13294–13307, 2023.
- Li Yujian and Liu Bo. A normalized levenshtein distance metric. *TPAMI*, 29(6):1091–1095, 2007.
- Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pp. 4791–4800, 2021.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3836–3847, 2023.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.
- Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *CVPR*, pp. 25827–25836, 2024a.
- Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7396–7404, 2024b.

A APPENDIX

A.1 LLM ACKNOWLEDGMENTS

We thank ChatGPT (GPT-5) and DeepSeek for the help in refining the language and improving clarity. All wording and factual content were reviewed and approved by the authors.

Table 6: Evaluation Results on Both Image Quality and Text Accuracy on UltraZoom-ST.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	PSNR $_{cr}\uparrow$	SSIM $_{cr}\uparrow$	LPIPS $_{cr}\downarrow$	DISTS $_{cr}\downarrow$	OCR-A $\uparrow$
Real-ESRGAN	25.84	0.854	0.139	0.130	28.25	23.59	0.843	0.170	0.168	0.632
HAT	27.33	0.873	0.128	0.133	27.21	24.76	0.870	0.139	0.146	0.643
MACRONet	22.90	0.804	0.234	0.173	40.19	17.92	0.693	0.515	0.360	0.542
SeeSR	25.02	0.826	0.167	0.159	32.31	21.62	0.796	0.245	0.207	0.374
SupIR	25.58	0.823	0.202	0.160	29.33	21.91	0.808	0.239	0.214	0.351
DiffTSR	23.48	0.805	0.230	0.157	35.77	17.92	0.688	0.392	0.285	0.482
DiffBIR	25.25	0.793	0.175	0.152	24.57	22.26	0.803	0.218	0.190	0.391
OSDiff	26.38	0.852	0.153	0.137	23.92	22.84	0.833	0.200	0.204	0.433
DreamClear	25.31	0.811	0.169	0.151	24.19	21.44	0.790	0.253	0.221	0.291
TSD-SR	23.99	0.800	0.161	0.165	30.40	20.62	0.786	0.232	0.218	0.401
DiT4SR	24.71	0.810	0.173	0.143	29.15	21.25	0.800	0.221	0.193	0.482
TADiSR	26.31	0.858	0.177	0.149	30.47	23.55	0.856	0.175	0.178	0.581
Ours	<u>27.00</u>	<u>0.871</u>	0.123	0.120	22.34	24.31	<u>0.860</u>	0.135	0.142	0.658
Real-ESRGAN	23.70	0.785	0.218	0.184	42.62	20.86	0.782	0.272	0.235	0.348
HAT	25.05	0.819	0.207	0.182	41.32	21.98	<u>0.816</u>	<u>0.250</u>	0.241	0.354
MACRONet	22.28	0.774	0.284	0.199	49.14	17.43	0.694	0.523	0.361	0.316
SeeSR	23.76	0.795	0.198	0.175	37.92	20.55	0.778	0.271	0.222	0.317
SupIR	23.49	0.746	0.302	0.204	41.80	20.00	0.757	0.335	0.282	0.248
DiffTSR	22.57	0.773	0.281	0.184	45.78	16.99	0.671	0.443	0.312	0.312
DiffBIR	23.76	0.731	0.245	0.196	35.99	20.82	0.770	0.275	0.234	0.293
OSDiff	<u>25.35</u>	<u>0.827</u>	<u>0.192</u>	0.166	29.06	21.62	0.809	0.263	0.263	0.307
DreamClear	24.20	0.778	0.209	0.177	32.08	20.66	0.779	0.283	0.251	0.247
TSD-SR	22.93	0.757	0.200	0.199	37.61	19.64	0.760	0.282	0.246	0.281
DiT4SR	23.45	0.774	0.204	<u>0.158</u>	31.73	20.12	0.779	<u>0.250</u>	<u>0.208</u>	0.266
TADiSR	24.68	0.795	0.362	0.227	52.13	21.57	0.798	0.366	0.347	0.375
Ours	25.63	0.834	0.185	0.150	28.82	22.14	0.823	0.204	0.194	0.456
Real-ESRGAN	22.09	0.718	0.423	0.284	89.22	18.87	0.721	0.503	0.427	0.116
HAT	22.66	0.738	0.451	0.299	89.58	<u>19.27</u>	<u>0.738</u>	0.533	0.484	0.123
MACRONet	21.00	0.716	0.422	0.254	79.67	17.01	0.680	0.553	0.382	0.123
SeeSR	21.79	0.731	0.310	0.228	64.31	18.87	0.717	0.396	0.308	<u>0.157</u>
SupIR	21.38	0.681	0.447	0.268	70.09	17.99	0.690	0.487	0.383	0.110
DiffTSR	20.88	0.713	0.412	0.233	74.07	15.87	0.630	0.538	0.348	0.120
DiffBIR	21.61	0.630	0.392	0.255	69.64	18.61	0.685	0.438	0.328	0.114
OSDiff	<u>23.11</u>	<u>0.768</u>	<u>0.274</u>	0.214	49.76	19.44	0.740	0.385	0.334	0.127
DreamClear	22.45	0.719	0.364	0.260	62.95	19.14	0.725	0.440	0.377	0.124
TSD-SR	21.14	0.705	0.273	0.225	57.07	17.67	0.686	0.393	<u>0.293</u>	0.119
DiT4SR	20.88	0.707	0.282	0.181	47.45	17.36	0.687	0.367	0.253	0.117
TADiSR	22.43	0.721	0.555	0.320	98.44	19.20	0.730	0.585	0.518	0.152
Ours	23.41	0.774	0.301	<u>0.209</u>	<u>54.74</u>	19.73	0.748	<u>0.374</u>	0.322	0.160
RealEsrGAN	23.99	0.790	0.248	0.194	30.60	21.25	0.786	0.302	0.266	0.371
HAT	<u>25.17</u>	0.815	0.249	0.198	30.12	<u>22.18</u>	<u>0.813</u>	0.291	0.276	0.379
MARCONet	<u>22.13</u>	0.768	0.306	0.205	34.28	<u>17.48</u>	<u>0.690</u>	0.529	0.366	0.334
SeeSR	23.64	0.788	0.219	0.184	26.73	20.45	0.767	0.297	0.241	0.191
SupIR	23.62	0.754	0.308	0.207	27.91	20.10	0.756	0.344	0.287	0.236
DiffTSR	22.41	0.767	0.300	0.189	31.25	17.00	0.665	0.452	0.313	0.317
DiffBIR	23.67	0.724	0.262	0.197	23.10	20.70	0.757	0.302	0.245	0.266
OSDiff	25.07	0.819	0.201	0.169	20.53	21.43	0.798	0.276	0.263	0.300
DreamClear	24.10	<u>0.773</u>	<u>0.238</u>	0.191	<u>21.75</u>	20.50	0.768	0.317	0.276	0.214
TSD-SR	22.79	0.757	0.207	0.194	24.08	19.42	0.748	0.296	0.249	0.266
DiT4SR	23.16	0.767	0.215	<u>0.159</u>	20.58	19.73	0.760	<u>0.273</u>	<u>0.216</u>	0.237
TADiSR	24.61	0.796	0.203	<u>0.160</u>	36.61	21.59	0.799	<u>0.360</u>	<u>0.336</u>	0.384
Ours	25.48	0.830	0.196	0.156	20.01	22.22	0.814	0.228	0.212	0.430

Table 7: Statistics of dataset size and line count in subsets of UltraZoom-ST.

Subset	image count	line count	mean lines/img	img > 5 lines
14 mm	1,439	15,073	10.47	754
35 mm	1,798	17,263	9.60	867
85 mm	1,799	17,339	9.64	869

Table 8: Efficiency analysis.

Methods	Flops (GFLOPs)	Speed (ms)	OCR-A
HAT	6670.32	1086.68	0.379
DiffTSR	58502.14	8610.59	0.317
DiT4SR w/o llava	160787.14	17385.14	0.237
DreamClear w/o llava	412843.67	83193.81	0.214
TADiSR	4497.96	342.32	0.384
TIGER (ours) Stage 1	6215.85	663.5	0.430
TIGER (ours) Stage 2	3734.01	360.06	

A.2 DATASET COLLECTION

We use VIVO X200 Ultra to collect images for 4 separate focal lengths (14 mm, 35 mm, 84 mm, and 200 mm). We first use PP-OCRV5 Cui et al. (2025) for the rough annotation, then we manually filter the images and annotations. During the filtering and annotation, each image undergoes the following rules:

- Width or height of the image should be no less than 256.
- Height of the text should not be less than 32 pixels.
- Score of OCR recognition of the text should not be lower than 0.9.
- Content of the text should not be empty or consist solely of whitespace.

A.3 MORE IMPLEMENTATION DETAILS

Dataset Settings. For training, we combine a synthetic dataset with real paired datasets (Real-CE (Ma et al., 2023) and UltraZoom-ST). Our synthetic dataset builds upon LSDIR (Li et al., 2023b), containing 27,000 triplets (x_H, x_L, x_m) . We render text on the GT of LSDIR and the corresponding text mask using the LBTS (Tang et al., 2023), then apply Real-ESRGAN degradation Wang et al. (2021) to generate LR images. Since there are misaligned image pairs in Real-CE (Hu et al., 2025; Zhang et al., 2024a) and text lines not annotated, we filter out misaligned image pairs and reannotate the images manually. In the end, we obtain 337 training pairs and 188 testing pairs from the Real-CE dataset. Images under 13mm and 52mm focal lengths are considered as LR x_L and GT x_H , respectively. Following Hu et al. (2025), we use SAM-TS (Ye et al., 2024) to obtain the text mask x_m from x_H . UltraZoom-ST triplets (x_H, x_L, x_m) are processed identically. For stage 1 training, we use cropped text regions from the dataset, while for stage 2, we use full images. Following Tuo et al. (2024), we use PP-OCRV3 Li et al. (2022) to extract strings from the text region of predicted SR for evaluation of OCR-A.

Training Details. For stage 1, we build our model based on the IDM baseline of DiffTSR (Zhang et al., 2024a). We set λ_{td} , λ_{Seg} , λ_{Focal} , and λ_{Dice} as 1, 0.1, 20, and 1 respectively. In stage 2, our model is built upon Stable Diffusion 3.5 medium (Esser et al., 2024) and follows a similar tile-based inference strategy to TADiSR (Hu et al., 2025) and SUPIR (Yu et al., 2024). The timestep t is set to 150. λ_{l2} , λ_{LPIPS} , and λ_{edge} are set to 1, 5, 100 when using the synthetic dataset for pretraining. Then, λ_{edge} is set to 0 and the remaining coefficients are kept unchanged when training on real paired datasets. We train both stages of the model using the AdamW (Loshchilov & Hutter, 2017) optimizer and set the learning rate to 5×10^{-5} and 5×10^{-6} for stages 1 and 2 separately. All experiments are conducted on NVIDIA H20 GPUs. For stage 1, we train the model with synthetic and real datasets for 8 epochs, and then use synthetic data to finetune the model for 2 epochs. For

stage 2, we first pretrain the model on the synthetic dataset for 50 epochs. Then we train on the real paired datasets for 50 epochs.

A.4 ADDITIONAL RESULTS

We provide detailed quantitative results of UltraZoom-ST in Table 6. Additional qualitative results are provided in Fig. 6.

A.5 STATIC AND EXAMPLES OF ULTRAZOOM-ST

In Table 7, we provide detailed statistics on the composition of the UltraZoom-ST dataset. Additionally, in Fig. 7, we present some example images from the dataset.

A.6 EFFICIENCY ANALYSIS

As shown in Table 8, stage 1 is a standard diffusion process that takes multiple steps in inference, the efficiency of our model may be suboptimal compared to one-step methods. However, our method achieves state-of-the-art performance in OCR-A, which cannot be easily obtained by extending inference time.

A.7 DISCUSSIONS AND LIMITATIONS

Our method leverages an OCR model to localize and interpret text in low-resolution images. In cases of severe degradation, however, OCR fails to recognize the text, which leads to failures in our method as well. Human-provided annotations could potentially mitigate this limitation.

A further challenge arises in real-world deployment. Because our text restoration model is diffusion-based, it requires multiple inference steps, which hinders real-time performance. Although our model reconstructs text structures more plausibly than prior approaches, it does not yet satisfy the speed demands of practical applications. Strategies such as model distillation and quantization offer promising directions to address this bottleneck.

This work is presented as an academic study of scene text image super-resolution in real-world conditions. Still, given its relevance to many applications, it carries both potential benefits (e.g., enhancing image quality in consumer devices) and risks (e.g., exposing private information from photos on social media). On balance, the societal benefits far outweigh the risks, particularly since complementary techniques such as inpainting and watermarking can be used to safeguard sensitive information.



Figure 6: Additional qualitative results of TIGER.



Figure 7: Detailed examples of UltraZoom-ST.