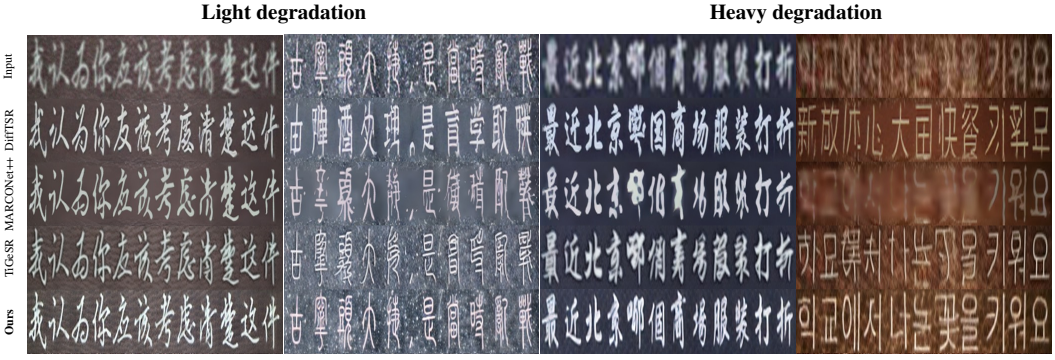
PERCEPTION OR PRIOR? LEGIBILITY-GUIDED PATH SELECTION FOR TEXT IMAGE RESTORATION IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Text image restoration requires preserving character identity and stroke geometry. Recognition-guided methods exploit character priors but are limited by vocabulary coverage. However, visible stroke patterns can guide reconstruction even for out-of-vocabulary characters, while reliable priors can complement visual evidence under severe degradation. Motivated by this complementarity, we propose **GRASP** (**G**lyph **R**econstruction via **A**daptive **S**election of **P**aths), a framework for legibility-guided selection between visual reconstruction and character-prior completion. A geometry-aware extractor aligns character features in a shared glyph space, where a visual path reconstructs observed structures and a prior path generates character-conditioned glyphs with input-dependent style modulation. A selector uses visual evidence and recognition confidence to choose complete glyphs, with visual fallback for predictions labeled as unknown. The selected glyphs form a structure mask for full-image restoration. We develop a synthetic training pipeline spanning 900 fonts and introduce **MText500**, a multilingual real-world benchmark. Experiments demonstrate the effectiveness of GRASP across character subsets and degradation levels, including languages absent from full-image restoration training.



(a) Known characters **(b)** Unknown characters **(c)** Known characters **(d)** Unknown characters

Figure 1: **Motivation for legibility-guided path selection.** Light degradation (a,b) calls for preserving visible strokes. Heavy degradation (c,d) benefits from reliable priors for known characters and visual recovery for unknowns. GRASP selects these paths per character. We compare with DiffTSR (Zhang et al., 2024), MARCONet++ (Li et al., 2025), and TiGeSR (Luo et al., 2026). Known/Unknown denotes characters inside/outside the recognition vocabulary.

1 INTRODUCTION

Text image restoration aims to recover both textual content and visual appearance from degraded observations. This task is particularly challenging because character identity depends on highly structured stroke patterns. Even minor deviations in stroke topology or geometry can change one character into another. Although general restoration models (Yu et al., 2024; Lin et al., 2024; 2025; Wu et al., 2024a;b) leverage powerful diffusion priors to recover realistic visual details, these models may synthesize plausible strokes that fail to preserve the original character identity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To improve textual fidelity, existing methods introduce recognition constrained supervision (Wang et al., 2019; Chen et al., 2021; Shi et al., 2023; Qin & Wang, 2024) or explicit recognition and generative priors (Ma et al., 2023; 2022; Zhao et al., 2022; Zhang et al., 2024; Min et al., 2026; Kim et al., 2025; Li et al., 2023; 2025; Luo et al., 2026). However, recognition losses constrain stroke geometry indirectly, making the restoration insufficient under severe degradation. Furthermore, explicit priors can provide structural guidance, but vocabulary-bound priors cannot provide identity-specific representations for out-of-vocabulary characters. Unfamiliar fonts also cause structural mismatches even for supported characters. These limitations are particularly relevant to text restoration in the wild, where multilingual content, complex typeface structure, and scenes containing misspellings are common. Such factors frequently introduce character identities and glyph appearances that are beyond the coverage of the training data. As shown in Fig. 1, mismatched priors can override visible stroke patterns and produce incorrect glyph structures.

Our key observation is that visible stroke patterns can guide the restoration of characters outside the recognition vocabulary. Characters across writing systems often share local stroke patterns, such as line segments, curves, and junctions. These shared patterns provide a basis for reconstructing out-of-vocabulary glyphs from visual evidence without assigning them to a known character category. With mild degradation, stroke layout and connectivity usually remain, so restoration can preserve the original structure, even for out-of-vocabulary characters and complex typefaces. With severe degradation, restoration should clarify recoverable strokes using shared visual patterns and remaining structural evidence. For supported characters, reliable OCR can reduce ambiguity and guide completion of missing strokes. However, conflicts exist between preserving visual fidelity and reconstructing characters using learned priors because of complex degradation. Coupling these two capabilities during training creates a trade-off: greater reliance on learned priors risks distorting out-of-vocabulary characters, while stronger dependence on visual evidence limits restoration under severe degradation.

Motivated by these observations, we introduce **GRASP** (**G**lyph **R**econstruction via **A**daptive **S**election of **P**aths), a framework for legibility-guided selection between visual recovery and character-prior completion. GRASP decouples these two capabilities through a visual branch and a prior branch operating in a 128×128 character space, and adaptively selects their reconstructions at the character level. A Geometry-Aware Character Feature Extractor constructs this shared space through two modules. An Attention-based Character Geometry Predictor estimates character transforms from multi-scale cross-attention maps. An Adaptive Affine RoI Extractor normalizes the corresponding features into a shared space for both reconstruction branches. The visual branch combines attention to stroke regions with multi-scale residual connections to preserve fine details. The prior branch generates glyph structures from predicted token embeddings, incorporating font style and affine geometry through adaptive instance normalization. Local and whole-character discriminators reinforce their respective specialization in stroke detail and global structure. Our Legibility-Guided Path Selection Module uses recoverable visual structure and recognition reliability as complementary cues. It chooses between candidate glyphs using aligned visual features and OCR confidence, supervised by an oracle comparing structure and character identity. Selection operates at the character level, avoiding the pixel-wise stroke inconsistencies. For characters predicted as unknown, GRASP falls back to the visual branch, preserving recovery beyond the recognition vocabulary.

Following TiGeSR (Luo et al., 2026), the selected glyphs form a spatial mask for full-image restoration. However, visual plausibility alone does not ensure textual fidelity. A generative restoration model may replace the reconstructed strokes with plausible alternatives. We apply layout augmentation that introduces discrepancies between the low-quality input and the mutually aligned ground-truth image and mask, encouraging the restored text to follow the mask.

To train GRASP, we develop a synthetic training pipeline with 900 fonts and introduce **MText500**, a multilingual real-world benchmark covering diverse typography and degradations. We evaluate glyph and full-image restoration on unfamiliar glyphs with recoverable strokes and severely degraded characters requiring reliable priors. Trained only on Chinese characters, English letters, and digits, the full-image restoration network is further evaluated on multilingual text with different degradations.

Our contributions are summarized as follows:

- We propose **GRASP**, a geometry-aware glyph reconstruction framework that decouples visual recovery from character-prior completion in a shared character space, preserving a visual recovery path for characters beyond the recognition vocabulary.

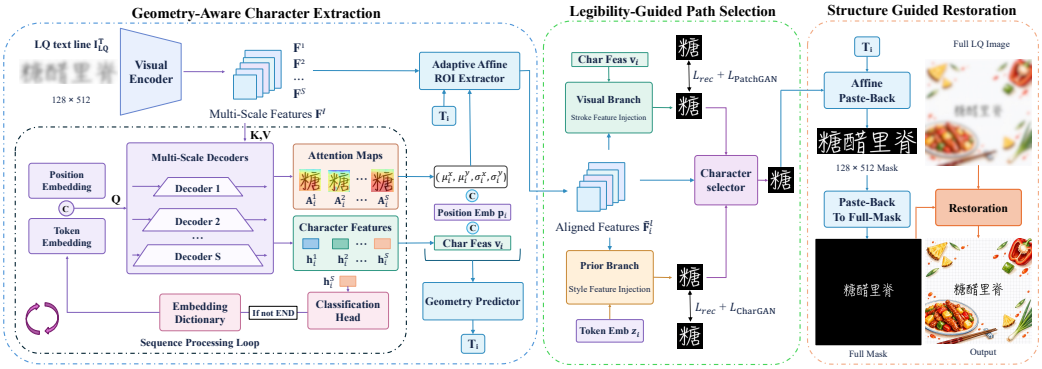


Figure 2: **Overview of GRASP. Left:** Geometry-aware extraction aligns character features in a shared glyph space. **Middle:** Complementary visual and prior branches reconstruct glyphs, followed by legibility-guided character-level selection. **Right:** Selected glyphs form a full-image structure mask that guides RGB restoration together with the low-quality input.

- We introduce legibility-guided character-level selection that uses glyph structure, aligned visual features, and OCR confidence to choose between complementary reconstruction paths. Whole-glyph selection preserves structural coherence, with an explicit visual fallback for characters predicted as unknown.
- We develop a synthetic training pipeline spanning 900 fonts and introduce **MText500**, a multilingual real-world benchmark covering diverse character sets, typography, and degradations, supporting evaluation within and beyond the recognition vocabulary.

2 Related Work

2.1 Real World Image Restoration

Real-world image restoration recovers high-quality images from low-quality images affected by complex and often unknown degradations. CNN-based approaches learn mappings from degraded inputs to clean targets, with methods such as RealSR (Cai et al., 2019). GAN-based approaches Zhang et al. (2021); Wang et al. (2021) introduce adversarial supervision to recover perceptually realistic details. More recently, diffusion-based methods have advanced real-world image restoration. StableSR (Wang et al., 2024) adapts pretrained diffusion priors to super-resolution, PASD (Yang et al., 2024) introduces pixel-aware guidance, SeeSR (Wu et al., 2024b) incorporates semantic-aware guidance, DiffBIR (Lin et al., 2024) separates degradation removal from detail generation, and SUPIR (Yu et al., 2024) scales up diffusion-based restoration. OSEDiff (Wu et al., 2024a) enables efficient one-step restoration, while HYPIR (Lin et al., 2025) combines pretrained diffusion priors with adversarial fine-tuning to achieve high-quality restoration in a single forward pass. DiT-SR (Cheng et al., 2025) and DiT4SR (Duan et al., 2025) explore diffusion Transformer architectures for super-resolution. Despite their success in recovering natural image details, these methods do not explicitly enforce character identity or stroke fidelity. This discrepancy motivates text-specific restoration methods that account for both visual appearance and textual content.

2.2 Text Image Restoration

Scene Text Image Restoration. Methods for cropped words and text lines improve readability through recognition supervision (Wang et al., 2019; Chen et al., 2021), gradient profile losses (Wang et al., 2020), or recognition-derived priors (Ma et al., 2023; 2022). Beyond recognition features, MARCONet (Li et al., 2023) and MARCONet++ (Li et al., 2025) introduce generative structure priors to provide character-specific shape and appearance information. DiffTSR (Zhang et al., 2024) couples textual and visual diffusion to improve blind text super-resolution. Such priors help resolve missing content, but their usefulness depends on their compatibility with the observed glyph; an incorrect character prediction or a mismatched shape prior can instead bias the reconstruction.

Text-Aware Full-Image Restoration. Recent approaches extend text restoration to full scene images. TeReDiff (Min et al., 2026) jointly learns image restoration and text spotting, using intermediate

recognition predictions to guide diffusion denoising. UniT (Kim et al., 2025) further integrates a vision-language model to iteratively refine textual guidance. Such semantic guidance depends on recognition reliability and may introduce incorrect character priors when unfamiliar text is incorrectly recognized. TiGeSR (Luo et al., 2026) adopts a text-first, image-later pipeline, with a text structure reconstruction module similar to DiffTSR (Zhang et al., 2024). Although it performs well on many images, its own ablation experiments indicate limited reliance on OCR guidance. Its apparent preference for visual features over character priors limits restoration under severe degradation. Together, these approaches motivate adapting the use of character priors to recoverable visual evidence and recognition reliability.

3 METHODOLOGY

3.1 GEOMETRY-AWARE CHARACTER EXTRACTION

To better disentangle visual evidence and character priors for individual glyphs, GRASP reformulates text reconstruction in a character space, where each character is represented by a 128×128 output. This formulation requires accurately extracting individual character features from a text line and mapping the reconstructed glyphs back to their original positions. Therefore, we introduce a Geometry-Aware Character Extractor that localizes each character from its cross-attention response and maps the corresponding multi-scale features into a shared space.

Attention-Based Character Geometry Prediction As shown in Fig. 2, the visual encoder extracts multi-scale features $\{\mathbf{F}^l\}_{l=1}^S$ from $\mathbf{I}_{\mathrm{LQ}}^{\mathrm{T}}$. Multi-scale recognition decoders produce character visual features $\{\mathbf{h}_i^l\}_{l=1}^S$ and character-specific cross-attention maps $\{\mathbf{A}_i^l\}_{l=1}^S$, which provide localization cues without an additional character detector. For each character, we summarize the highest-resolution attention map by its normalized centroid and spatial extent, $\mathbf{m}_i = [\mu_i^x, \mu_i^y, \sigma_i^x, \sigma_i^y]$, and aggregate multi-scale visual features into $\mathbf{v}_i$ using attention-weighted pooling. A residual MLP predicts character scales and translation offsets:

$$[s_i^x, s_i^y, \Delta t_i^x, \Delta t_i^y] = \mathcal{P}_{\mathrm{geo}}\left([\mathbf{m}_i; \mathbf{v}_i; \mathbf{p}_i]\right), \tag{1}$$

where $\mathbf{p}_i$ encodes the character index.

Rather than predicting character locations from scratch, we construct a translation prior $\bar{\mathbf{t}}_i$ that centers a character patch of predefined initial scale at the attention centroid. The predicted offsets refine this prior, yielding the character-to-line scale-and-translation transform

$$\mathbf{T}_i = \begin{bmatrix} s_i^x & 0 & W(\bar{t}_i^x + \Delta t_i^x) \\ 0 & s_i^y & H(\bar{t}_i^y + \Delta t_i^y) \end{bmatrix}, \tag{2}$$

where H and W denote the text-line height and width. This parameterization combines attention-based localization with learned geometric refinement. Attention construction, feature aggregation, and the translation prior are detailed in Supp. Sec. A.

Adaptive Affine RoI Extraction Using $\mathbf{T}_i$, we map a regular character grid $\mathcal{G}$ to the text line and extract aligned features through differentiable bilinear sampling:

$$\widetilde{\mathbf{F}}_i^l = \mathrm{GridSample}\left(\mathbf{F}^l, \mathbf{T}_i(\mathcal{G})\right), \tag{3}$$

with sampling coordinates converted to the convention of each feature map. The shared transform adapts the sampling location and extent to each character while preserving correspondence across scales. The extracted features are aligned with the 128×128 character space and supplied to both glyph reconstruction paths. We retain $\mathbf{T}_i$ to place the selected glyph back into the text line.

3.2 DUAL-BRANCH GLYPH RECONSTRUCTION

The visual and prior paths reconstruct glyphs from perception and character conditioning, respectively:

$$\hat{\mathbf{G}}_i^v = \mathcal{R}_v(\mathbf{v}_i, \{\widetilde{\mathbf{F}}_i^l\}_{l=1}^S), \qquad \hat{\mathbf{G}}_i^p = \mathcal{R}_p(z_i, \{\widetilde{\mathbf{F}}_i^l\}_{l=1}^S). \tag{4}$$

The recognition classifier and prior embeddings share vocabulary $\mathcal{V}$, excluding special tokens. During prior training, z_i uses ground-truth identities, mapping out-of-vocabulary labels to $\langle \mathrm{unk} \rangle$; selector training and inference use OCR predictions.

Visual-Evidence Reconstruction Path A projection of $\mathbf{v}_i$ initializes an upsampling decoder without character-class embeddings. Aligned spatial features preserve stroke details through

$$\mathbf{X}_i^{v,l} = \mathcal{B}_l^v([\mathrm{Up}(\mathbf{X}_i^{v,l-1}); \mathcal{C}_l^v(\widetilde{\mathbf{F}}_i^l)]), \tag{5}$$

where $\mathcal{C}_l^v$ projects features and $\mathcal{B}_l^v$ is a residual block at scale l. All valid characters, including unknowns, use the ground-truth glyph $\mathbf{G}_i^{\mathrm{GT}}$ as the reconstruction target.

Character-Prior Reconstruction Path A jointly learned embedding indexed by z_i initializes the prior decoder. Pooled aligned features predict stage-specific style parameters for channel-wise modulation:

$$\widetilde{\mathbf{X}}_i^{p,l} = (1 + \boldsymbol{\gamma}_i^l) \odot \mathrm{IN}(\mathbf{X}_i^{p,l}) + \boldsymbol{\beta}_i^l, \tag{6}$$

where IN is instance normalization. Thus, visual conditioning uses modulation rather than spatial concatenation. The prior target is $\mathbf{G}_i^{p,\mathrm{tgt}} = \mathbf{G}_i^{\mathrm{GT}}$ for $y_i \in \mathcal{V}$ and $\mathbf{0}$ otherwise.

Reconstruction objectives. For $b \in \{v, p\}$, $\mathcal{L}_{\mathrm{rec}}^b$ combines pixel, perceptual, foreground segmentation, and structural similarity terms. Local and whole-character discriminators D_v, D_p encourage stroke detail and global structure, respectively, with generator loss $\mathcal{L}_{\mathrm{adv}}^b = -\mathbb{E}_{i \in \Omega_b}[D_b(\hat{\mathbf{G}}_i^b)]$. Each path uses $\mathcal{L}_b = \mathcal{L}_{\mathrm{rec}}^b + \lambda_{\mathrm{adv}}^b \mathcal{L}_{\mathrm{adv}}^b$, giving $\mathcal{L}_{\mathrm{glyph}} = \mathcal{L}_v + \mathcal{L}_p$. More details appear in Supp. Sec. B.

3.3 Legibility-Guided Path Selection

Visual features and recognition confidence provide complementary legibility cues. The selector predicts the probability of choosing the visual path:

$$p_i^v = \mathcal{S}_{\mathrm{sel}}(\hat{\mathbf{G}}_i^v, \hat{\mathbf{G}}_i^p, \mathbf{V}_i, c_i), \tag{7}$$

where $\mathbf{V}_i$ aggregates aligned multi-scale features and c_i is the maximum recognition probability.

Quality-based supervision. An oracle compares both candidates against the original ground-truth glyph using pixel and foreground segmentation errors together with an identity loss from a frozen OCR recognizer $\mathcal{O}$:

$$E_i^b = \lambda_{\mathrm{pix}} E_{\mathrm{pix},i}^b + \lambda_{\mathrm{seg}} E_{\mathrm{seg},i}^b + \lambda_{\mathrm{ocr}} E_{\mathrm{ocr},i}^b, \qquad b \in \{v, p\}, \tag{8}$$

where $E_{\mathrm{ocr},i}^b = \ell_{\mathrm{ocr}}(\mathcal{O}(\hat{\mathbf{G}}_i^b), y_i)$ for $y_i \in \mathcal{V}$ and zero otherwise. The detached soft target is

$$o_i = \mathrm{sg}\left[\mathrm{Sigmoid}\left(\frac{E_i^p - E_i^v}{\tau}\right)\right], \tag{9}$$

where $\tau > 0$ controls sharpness, larger o_i favors the visual path, and sg stops gradients. Detached weights ρ_i downweight candidates of similar quality. We optimize

$$\mathcal{L}_{\mathrm{sel}} = \frac{\sum_{i \in \Omega_{\mathrm{sel}}} \rho_i \, \mathrm{BCE}(p_i^v, o_i)}{\sum_{i \in \Omega_{\mathrm{sel}}} \rho_i + \epsilon} + \frac{\lambda_{\mathrm{bin}}}{|\Omega_{\mathrm{sel}}|} \sum_{i \in \Omega_{\mathrm{sel}}} p_i^v (1 - p_i^v), \tag{10}$$

where Ω_{sel} includes all valid characters, including unknowns, and $\epsilon > 0$ stabilizes normalization. The second term encourages decisive selection.

Inference. We select the complete visual glyph when OCR predicts $\langle \mathrm{unk} \rangle$ or $p_i^v \geq 0.5$, and the prior glyph otherwise. Selector and oracle details appear in Supp. Sec. C.

3.4 Training Objectives

Stage I: Visual and geometry learning. We jointly train the encoder, recognition module, geometry predictor, visual path, and its discriminator on all valid characters. Recognition and geometry use token cross-entropy $\mathcal{L}_{\mathrm{text}}$ and Smooth L1 loss $\mathcal{L}_{\mathrm{geo}}$. Visual glyphs form a differentiable line mask $\hat{\mathbf{M}}^{\mathrm{T}} = \mathrm{Compose}(\{(\hat{\mathbf{G}}_i^v, \mathbf{T}_i)\}_{i=1}^L)$. Against the ground-truth mask, $\mathcal{L}_{\mathrm{paste}}$ combines BCE, Dice, and ℓ_1 terms to supervise reconstruction and placement:

$$\mathcal{L}_{\mathrm{stage1}} = \lambda_{\mathrm{text}} \mathcal{L}_{\mathrm{text}} + \lambda_{\mathrm{geo}} \mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{paste}} \mathcal{L}_{\mathrm{paste}} + \mathcal{L}_v. \tag{11}$$

Stage II: Prior learning. We freeze Stage I modules and optimize the prior decoder, embeddings, and discriminator with $\mathcal{L}_p$, using ground-truth identities and the targets in Section 3.2.

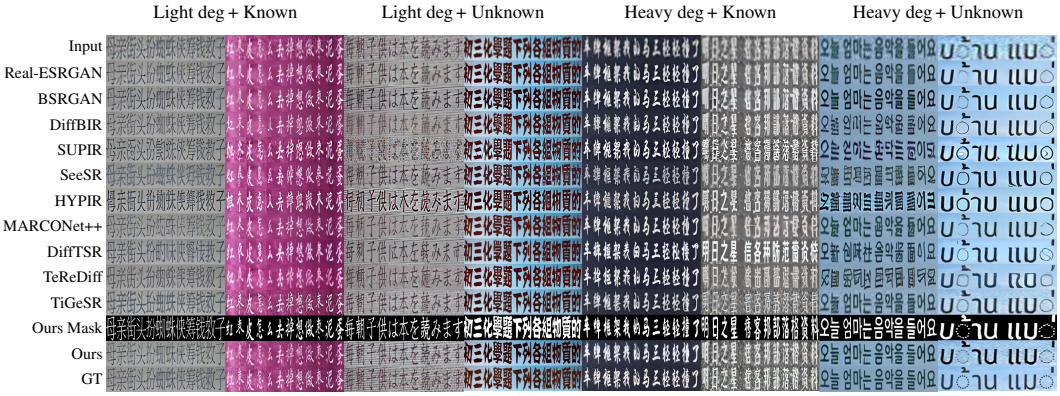


Figure 3: Qualitative comparison on synthetic degradations under light and heavy degradation settings, with known and unknown text categories. Each setting contains two examples.

Stage III: Selector learning. We freeze all preceding modules and train only the selector with $\mathcal{L}_{\text{sel}}$. OCR-conditioned prior candidates expose it to inference-time recognition errors. Loss and discriminator details appear in Supp. Sec. D.

3.5 Layout Augmentation for Full-Image Restoration

Selected glyphs form a spatial mask that guides restoration alongside the low-quality input. During restoration training, layout augmentation keeps the ground-truth image and its mask aligned while introducing slight text-placement discrepancies with the low-quality input. Supervision against the corresponding ground-truth image encourages the network to follow the mask-specified layout, strengthening mask control over restored text.

4 Experiments

Training Data. We synthesize paired images and text masks by rendering text with randomized typography and geometry using 900 fonts on patches from DF2K, comprising DIV2K (Agustsson & Timofte, 2017) and Flickr2K (Timofte et al., 2017). Degraded inputs are generated online using the two-stage Real-ESRGAN pipeline (Wang et al., 2021). Recognition and prior modules share a vocabulary of simplified Chinese characters, English letters, and digits; traditional Chinese characters outside this vocabulary are used as unknown-class examples for glyph reconstruction. Thus, out-of-vocabulary indicates vocabulary exclusion, not necessarily absence from reconstruction training. Additional Japanese and Korean samples, labeled $\langle\text{unk}\rangle$, train only the recognition classifier with both encoder and decoder features detached. They do not expand the vocabulary for train glyph reconstruction or full-image restoration.

Synthetic Test Data. Our paired synthetic benchmark, **SynData**, comprises 600 clean images with 1,440 text lines, split equally into Known (724 lines with simplified Chinese, English, and digits) and Unknown (716 lines with Japanese, Korean, French, Greek, and Thai) subsets. In Unknown, 72% of characters fall outside the shared recognition and prior vocabulary; thus, this designation applies to the subset rather than every character. This mixture better reflects real-world scenarios. Text is placed using SynthText's segmentation and depth guidance (Gupta et al., 2016) or randomly on high-quality background crops. We apply light and heavy degradations from the training degradation family to the same clean images for controlled comparisons.

Real-World Test Data. We further evaluate on **MText500**, a curated real-world test set containing 500 low-quality scene-text images, each with a long side of 1,024 pixels and corresponding ground-truth text annotations. The set is balanced across four subsets: known/light, known/heavy, unknown/light, and unknown/heavy with 125 images each and 236, 660, 453, and 788 annotated text lines, respectively. This yields 250 light samples with 689 text lines and 250 heavy samples with 1,448 text lines. MText500 is constructed from ReCTS Liu et al. (2019), ICDAR2015 Karatzas et al. (2015), and MLT2019 Nayef et al. (2019), contributing 67, 62, and 371 images, respectively. At the image-stratum level, it contains 90 Chinese, 160 English-or-digit, 40 Arabic, 37 Bangla, 34 Hindi, 48 Japanese, and 91 Korean samples. In total, the dataset contains 2,137 annotated text lines, including 848 Latin, 406 Chinese, 255 Korean, 224 Arabic, 134 Bangla, 131 Hindi, 111 Japanese, 24 symbol,

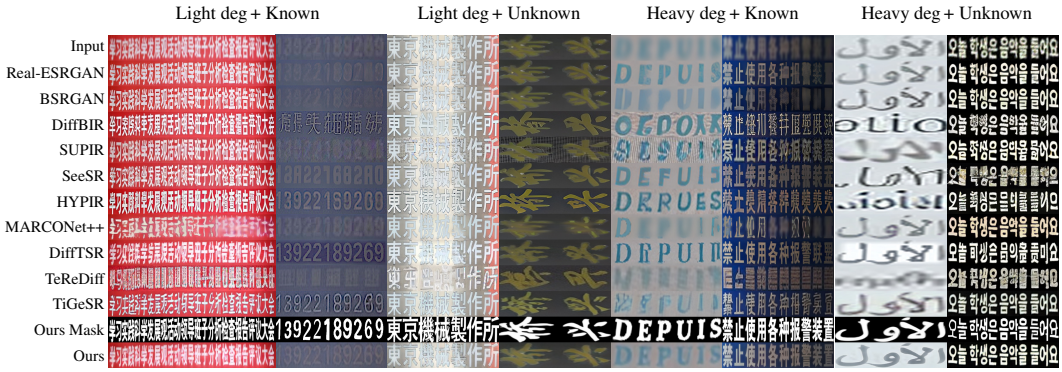


Figure 4: Qualitative comparison on the real-world MText500 dataset under light and heavy degradations, with Known and Unknown subsets shown separately. Each setting contains two examples.

and 4 mixed-script lines. We report OCR-A and FID; the FID reference set contains 984 collected high-quality scene-text images with 6,508 text lines, covering the same languages as MText500.

Implementation Details. We implement GRASP in PyTorch using 1024×1024 training images, 512×128 normalized text lines, and 128×128 glyphs. Text-line outputs are mapped back via projective transformations to preserve spatial layout. Across training stages, all components use AdamW with a learning rate of 5×10^{-5}, 1,000 warm-up iterations, cosine decay, and a per-GPU batch size of 4. We adopt the TiGeSR (Luo et al., 2026) restoration architecture, trained with paired supervision on 512×512 crops and layout augmentation (Section 3.5) to align outputs with text masks. Further details are provided in Supp. Sec. E.

4.1 Comparison with Other Methods

We compare GRASP with ten restoration methods from three groups: (i) GAN-based general image restoration methods, including BSRGAN (Zhang et al., 2021) and Real-ESRGAN (Wang et al., 2021); (ii) general image restoration methods that exploit pretrained diffusion priors, including DiffBIR (Lin et al., 2024), SUPIR (Yu et al., 2024), SeeSR (Wu et al., 2024b), and HYPIR (Lin et al., 2025); and (iii) methods designed specifically for text restoration, including MARCONet++ (Li et al., 2025), DiffTSR (Zhang et al., 2024), TeReDiff (Min et al., 2026), and TiGeSR (Luo et al., 2026).

Quantitative Comparison. We evaluate restoration using six complementary metrics: PSNR and SSIM (Wang et al., 2004) measure reconstruction fidelity; LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) assess perceptual similarity; FID (Heusel et al., 2017) measures distributional image quality; and OCR recognition accuracy (OCR-A) (Luo et al., 2026), evaluated with PP-OCRv6 (Zhang et al., 2026), measures textual correctness. Tables 1 and 2 report results under light and heavy degradations, respectively, with Known and Unknown evaluated separately for both SynData and MText500. SynData is evaluated using all six metrics, while MText500 is evaluated using FID and OCR-A. All metrics are computed on text-line crops extracted from full-image restoration outputs; full-image metrics are provided in the supplementary material. To control for differences in text detection, all methods use the same bounding boxes for text-line extraction and pasting.

Under light degradation on SynData, GRASP achieves the best LPIPS, DISTS, and FID on both subsets. It also ranks first in OCR-A on Known and PSNR on Unknown, second in PSNR on Known and OCR-A on Unknown, and third in SSIM on both subsets. These results demonstrate strong perceptual restoration alongside competitive reconstruction fidelity and textual correctness. BSRGAN and Real-ESRGAN also achieve competitive OCR-A. This is consistent with the importance of preserving observable character structures when degradation is mild. The lower OCR-A of several generative and text-specific baselines suggests that strong restoration priors do not necessarily improve readability when sufficient visual evidence remains. Under heavy degradation on SynData, GRASP ranks first in PSNR, SSIM, LPIPS, DISTS, and OCR-A on both subsets. In particular, it achieves OCR-A scores of 69.88% on Known and 69.87% on Unknown, exceeding the strongest competing methods by 7.35 and 2.60 percentage points, respectively. These results are consistent with the benefits of combining learned character priors with transferable stroke-level cues when the observed glyph structure is incomplete or ambiguous.

Table 1: **Evaluation on SynData and MText500 with light degradation.** Metrics are computed on text-line crops extracted from full-image restoration outputs. Full-image metrics are provided in the supplementary material. OCR-A is reported in %. Best and second-best results are shown in **bold** and underlined, respectively.

	SynData												MText500			
	Known						Unknown						Known		Unknown	
Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	FID↓	OCR-A↑	FID↓	OCR-A↑
BSRGAN	**21.72**	**.7503**	.2533	.1829	51.67	80.05	22.54	**.7523**	.2502	.1838	55.18	**77.17**	78.58	96.11	54.68	83.49
Real-ESRGAN	21.12	.7486	.2473	.1826	49.13	80.60	21.78	.7506	.2415	.1813	51.87	76.80	78.94	96.04	54.72	83.53
DiffBIR	18.13	.5467	.3322	.2073	53.30	68.89	19.74	.5979	.2954	.1926	50.50	69.66	79.46	92.38	54.61	81.39
SUPIR	16.43	.5019	.3736	.2292	51.08	66.53	18.05	.5481	.3432	.2187	54.02	65.34	79.32	95.37	54.98	82.77
SeeSR	17.36	.5044	.3793	.2320	58.97	60.74	18.91	.5609	.3382	.2151	55.02	64.00	78.59	94.62	54.57	82.59
HYPIR	14.94	.4167	.4013	.2516	62.84	64.71	16.22	.4550	.3692	.2389	58.75	64.98	79.45	94.72	55.00	82.67
MARCONet++	20.15	.6473	.3275	.2091	62.41	73.67	21.70	.6615	.3272	.2141	75.41	74.66	**77.32**	95.33	**53.90**	82.55
DiffTSR	18.43	.5928	.3389	.2084	56.03	71.46	19.97	.6162	.3367	.2126	63.97	68.98	77.80	95.98	54.34	83.25
TeReDiff	16.25	.4099	.4450	.2661	69.68	51.08	17.62	.4626	.4102	.2485	66.55	57.88	81.01	83.49	55.10	75.38
TiGeSR	20.38	.6615	.3085	.2045	54.30	76.58	21.79	.6782	.2961	.2020	54.95	75.73	79.80	95.88	54.71	83.45
Ours	21.70	.7405	**.2413**	**.1618**	**42.41**	**81.10**	**22.96**	.7481	**.2273**	**.1569**	**41.58**	76.87	77.38	**97.05**	54.24	**86.15**

Table 2: **Evaluation on SynData and MText500 with heavy degradation.** Settings follow Table 1.

	SynData												MText500			
	Known						Unknown						Known		Unknown	
Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	FID↓	OCR-A↑	FID↓	OCR-A↑
BSRGAN	17.85	.5136	.4347	.2845	95.57	60.47	18.64	.5479	.4175	.2760	92.47	65.49	42.20	89.77	35.96	54.93
Real-ESRGAN	17.74	.5221	.4250	.2764	86.76	62.35	18.50	.5547	.4066	.2691	85.75	67.27	42.83	89.96	36.23	54.92
DiffBIR	16.64	.4432	.4208	.2538	64.99	57.35	17.66	.4894	.3895	.2407	61.96	63.52	42.58	60.03	36.32	49.42
SUPIR	15.19	.3936	.4423	.2608	**60.52**	56.10	16.61	.4487	.4093	.2473	62.06	60.42	43.49	72.68	36.42	53.86
SeeSR	16.45	.4389	.4257	.2539	64.43	54.90	17.61	.4895	.3877	.2382	60.60	61.68	42.30	71.06	**35.75**	53.34
HYPIR	14.72	.3896	.4302	.2569	61.56	57.64	15.83	.4313	.3947	.2422	**57.22**	61.61	42.64	75.79	36.67	54.95
MARCONet++	17.65	.4929	.4403	.2686	80.65	60.45	18.83	.5308	.4268	.2658	91.55	64.41	42.19	89.95	36.35	54.87
DiffTSR	16.10	.4510	.4245	.2468	66.91	61.57	17.30	.4851	.4215	.2500	74.34	58.59	**42.03**	89.54	36.11	54.94
TeReDiff	15.71	.3799	.4575	.2706	71.29	50.49	16.78	.4282	.4243	.2526	65.61	56.65	45.09	38.52	37.69	40.70
TiGeSR	17.61	.4748	.4226	.2667	73.43	62.53	18.83	.5140	.4001	.2594	73.01	66.19	43.12	88.45	36.72	54.94
Ours	**17.92**	**.5311**	**.3715**	**.2305**	61.45	**69.88**	**19.37**	**.5717**	**.3474**	**.2248**	59.50	**69.87**	42.07	**90.78**	35.96	**60.31**

On MText500, GRASP achieves the highest OCR-A across all four subset–degradation combinations. Under light degradation, it obtains 97.05% on Known and 86.15% on Unknown; under heavy degradation, it achieves 90.78% and 60.31%, respectively. The improvements over the strongest competing methods are particularly pronounced on Unknown, reaching 2.62 percentage points under light degradation and 5.36 percentage points under heavy degradation. GRASP also ranks second in FID in all four settings. Together, these results demonstrate consistent gains in textual correctness on real-world images while maintaining competitive distributional image quality.

Qualitative Comparison. Figure 3 and Figure 4 compare GRASP with other methods across both degradation levels and character subsets. In the lightly degraded examples, GRASP preserves distinctive glyph styles and visible stroke layouts, whereas competing outputs sometimes regularize these details or introduce structural changes. Under heavy degradation, GRASP produces more coherent stroke patterns in the illustrated Unknown example, while some competing outputs contain incomplete glyphs or substantially altered character shapes. These examples illustrate differences in structural preservation and readability that complement the quantitative evaluation.

4.2 Ablation Studies

Tables 3 and 4 examine the roles of the reconstruction branches, adaptive selector, text masks, and restoration model. Unless explicitly stated otherwise, all ablations use the same trained GRASP model and the same fixed full-image restorer, changing only the specified inference configuration. In the Prior Branch Only ablation, characters predicted as $\langle \mathrm{unk} \rangle$ contribute empty glyph masks, without falling back to the visual branch. The full model instead selects the visual reconstruction for these predictions. Both GT OCR variants replace predicted character labels with ground-truth labels; Full w/ GT OCR retains the learned selector and fixes recognition confidence at 0.95.

Complementarity of reconstruction branches. The single-branch ablations reveal complementary strengths of visual evidence and character priors. Under light degradation, visual-only outperforms both prior-only variants across all reported metrics on both subsets, even when the prior branch receives ground-truth character labels. Under heavy degradation, prior-only achieves higher OCR-A

Table 3: **Ablations on SynData with light degradation.** Known and Unknown are evaluated on text-line crops. Overall OCR-A uses weights $724 : 716$. OCR-A is in %. Rankings use displayed values within each block. The upper block marks best/second-best in **bold**/underlined, and the lower block marks only best in **bold**. The lower block uses the same fixed TiGeSR restoration model (T-R) with TiGeSR and GRASP masks, respectively.

	Known						Unknown						Overall
Setting	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	OCR-A↑
Full (GRASP)	21.70	.7405	.2413	.1618	42.41	81.10	22.96	.7481	.2273	.1569	41.58	76.87	79.00
Visual Branch Only	**21.85**	**.7445**	**.2384**	**.1611**	**42.18**	81.33	**23.02**	**.7495**	**.2260**	**.1564**	**41.38**	76.99	79.17
Prior Branch Only	21.41	.7341	.2467	.1649	44.08	81.27	22.88	.7450	.2361	.1643	43.81	76.43	78.86
Prior Only w/ GT OCR	21.40	.7338	.2469	.1649	44.16	81.29	22.90	.7449	.2366	.1649	44.02	76.45	78.88
Full w/ GT OCR	21.70	.7403	.2414	.1618	42.45	**81.38**	22.97	.7481	.2273	.1568	41.54	**77.08**	**79.24**
TiGeSR	**20.38**	**.6615**	**.3085**	**.2045**	**54.30**	76.58	**21.79**	**.6782**	**.2961**	**.2020**	54.95	75.73	76.16
w/ T-R	19.21	.6185	.3174	.2070	54.70	**79.69**	20.77	.6520	.2990	.2026	**53.82**	**75.84**	**77.78**

Table 4: **Ablations on SynData with heavy degradation.** Settings follow Table 3.

	Known						Unknown						Overall
Setting	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑	OCR-A↑
Full (GRASP)	17.92	.5311	.3715	**.2305**	61.45	69.88	19.37	.5717	.3474	.2248	59.50	69.87	69.88
Visual Branch Only	**18.24**	**.5421**	**.3700**	.2322	63.06	68.59	19.45	**.5744**	**.3459**	**.2247**	**59.49**	69.74	69.16
Prior Branch Only	17.77	.5270	.3728	.2317	61.98	69.78	19.46	.5741	.3578	.2362	62.33	67.21	68.50
Prior Only w/ GT OCR	17.76	.5268	.3727	.2315	61.95	69.91	**19.49**	.5742	.3589	.2373	62.63	68.28	69.10
Full w/ GT OCR	17.92	.5310	.3713	**.2305**	**61.31**	**70.05**	19.38	.5720	.3473	.2248	59.52	**70.11**	**70.08**
TiGeSR	**17.61**	**.4748**	**.4226**	**.2667**	**73.43**	62.53	**18.83**	**.5140**	**.4001**	**.2594**	**73.01**	66.19	64.35
w/ T-R	16.61	.4282	.4336	.2684	76.40	**68.54**	18.22	.4951	.4058	.2643	74.91	**67.28**	**67.91**

on Known, whereas visual-only performs better on Unknown, with the same pattern holding after correcting the prior branch's character labels. These results support complementary strengths across degradation levels and character subsets.

Effectiveness of legibility-guided path selection. With predicted OCR, adaptive selection outperforms both fixed branches in OCR-A on Known and Unknown under heavy degradation, but falls below visual-only under light degradation. With ground-truth labels and confidence fixed at 0.95, selection surpasses visual-only on both light-degradation subsets, achieving 79.24% overall OCR-A versus 79.17%. In contrast, prior-only with ground-truth labels reaches only 78.88%. Thus, improved prior reconstructions alone do not explain the gain; the results support the selector's ability to exploit complementary candidates under reliable recognition conditioning. The contrast with predicted-OCR selection suggests that recognition conditioning limits its benefits under light degradation.

Transferability of reconstructed text masks. Keeping the TiGeSR restoration weights fixed, substituting GRASP masks improves overall recognition at both degradation levels, with gains on both subsets. These results show that our masks provide more useful recognition cues. However, these gains do not consistently extend to image-quality metrics, suggesting that more informative text masks without a matched restoration do not guarantee better restoration across all metrics.

5 Conclusion

We presented GRASP, a framework that combines geometry-aligned visual reconstruction and character-prior completion through legibility-guided selection to guide full-image text restoration. Experiments on SynData and MText500 demonstrate strong textual fidelity across character subsets and degradation levels. Ablations show that visual reconstruction is more effective on Unknown under both degradation levels and on Known under light degradation, while character priors improve recognition on Known under heavy degradation. Adaptive selection improves overall OCR accuracy over either path under heavy degradation, but does not consistently surpass visual-only reconstruction under light degradation because of OCR accuracy. Future work will improve selection under uncertain recognition and geometry handling for complex scripts.

AI USE STATEMENT

We used AI-assisted tools to help identify and fix bugs in our implementation, organize intermediate files generated during experiments, and polish the language and clarity of the manuscript. The authors take full responsibility for the correctness of the implementation, the accuracy of the reported results, and the final content of the paper.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

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.

Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2021.

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. doi: 10.1609/aaai.v39i3.32247.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chunle 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*, pp. 18948–18958, 2025.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315–2324, 2016.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay R. Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *International Conference on Document Analysis and Recognition*, pp. 1156–1160, 2015. doi: 10.1109/ICDAR.2015.7333942.

Jin Hyeon Kim, Paul Hyunbin Cho, Claire Kim, Jaewon Min, Jaeeun Lee, Jihye Park, Yeji Choi, and Seungryong Kim. Unified diffusion transformer for high-fidelity text-aware image restoration. *arXiv preprint arXiv:2512.08922*, 2025.

Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10103–10113, 2023.

Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Enhanced generative structure prior for chinese text image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (12):11366–11381, 2025. doi: 10.1109/TPAMI.2025.3596329.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision*, pp. 430–448, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics*, 44(6):1–21, 2025. doi: 10.1145/3763346.

Xi Liu, Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, Xiang Bai, Baoguang Shi, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR 2019 robust reading challenge on reading Chinese text on signboard. In *International Conference on Document Analysis and Recognition*, pp. 1577–1581, 2019. doi: 10.1109/ICDAR.2019.00253.

Minxing Luo, Linlong Fan, Qiushi Wang, Ge Wu, Yiyan Luo, Yuhang Yu, Jinwei Chen, Yaxing Wang, Qingnan Fan, and Jian Yang. Restore text first, enhance image later: Two-stage scene text image super-resolution with glyph structure guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30553–30563, 2026.

Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911–5920, 2022.

Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023. doi: 10.1109/TIP.2023.3237002.

Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2026.

Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-Lin Liu, and Jean-Marc Ogier. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In *International Conference on Document Analysis and Recognition*, pp. 1582–1587, 2019. doi: 10.1109/ICDAR.2019.00254.

Rui Qin and Bin Wang. Scene text image super-resolution via content perceptual loss and criss-cross transformer blocks. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–10, 2024. doi: 10.1109/IJCNN60899.2024.10650570.

Qin Shi, Yu Zhu, Chuantao Fang, and Dawei Yang. TLWSR: Weakly supervised real-world scene text image super-resolution using text label. *IET Image Processing*, 17(9):2780–2790, 2023. doi: 10.1049/ipr2.12827.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 114–125, 2017.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7.

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 arXiv:1909.07113*, 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 *Proceedings of the European Conference on Computer Vision*, pp. 650–666. Springer, 2020. doi: 10.1007/978-3-030-58607-2_38.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-2938.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the European Conference on Computer Vision*, pp. 74–91. Springer, 2024. doi: 10.1007/978-3-031-73247-8_5.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, Hongen Liu, Suyin Liang, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PP-OCRv6: From 1.5M to 34.5M parameters, surpassing billion-scale VLMs on OCR tasks, 2026.

Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25827–25836, 2024.

Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: Scene text image super-resolution with triple clues. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1707–1713, 2022. doi: 10.24963/ijcai.2022/238.