Diffusion-Based Restoration of Japanese Classical Documents with Multi-Level Synthetic Degradation

HANZHE GAO[1,a)] HONGHUI YUAN[1,b)] JUNWEN CHEN[1,c)] PEILIN XIONG[1,d)] KEIJI YANAI[1,e)]

Abstract. We explore diffusion-based restoration for Japanese classical document images under multi-level synthetic degradation. Pages from ROIS-DS Center for Open Data in the Humanities (CODH) [6] are degraded with ink stains, water stains, mold spots, and smudges at three severity levels. A Palette-style diffusion model with RePaint-style masked sampling restores damaged regions while preserving intact regions, without OCR annotations. Under the proposed synthetic benchmark, our method improves medium and severe degradation compared with classical inpainting baselines, LaMa, and a cross-domain DiffHDR reference. Mask-aware metrics further show that masked sampling mainly benefits known-region preservation. Remaining limitations include synthetic-only degradation, oracle mask dependency, and the absence of character-level readability evaluation.

1. Introduction

Japanese classical documents are valuable cultural resources, but many surviving pages contain ink stains, water damage, mold, smudges, and paper aging. These artifacts obscure characters and make both human reading and downstream document analysis difficult. A large body of Japanese document analysis has focused on recognition or generation of kuzushiji characters, including benchmark datasets for classical Japanese literature [2], character-level restoration [3], and diffusion-based font generation [11]. In contrast, full-page restoration of Japanese classical documents remains relatively underexplored.

Recent diffusion-based historical document repair methods provide useful references for this problem. DiffHDR [9] formulates damaged historical document repair using paired damaged-clean data and character-level annotations, while AutoHDR [12] extends restoration to full-page settings through a modular pipeline with OCR-assisted localization and context prediction. These methods show the benefit of task-specific supervision, but their datasets and annotation assumptions do not directly match Japanese classical pages where dense character-level restoration labels are expensive.

This paper studies a complementary setting: restoration from synthetically degraded CODH pages [6] with oracle masks available from the synthesis process. We do not claim to solve real-world manuscript restoration. Instead, we construct a controlled benchmark that lets us analyze whether a conditional diffusion model can remove localized page degradation while preserving intact regions. The main contributions are a multi-level synthetic degradation pipeline, a Palette-style restoration model with RePaint-style masked sampling, and an evaluation against OpenCV inpainting, LaMa, and a cross-domain DiffHDR reference using both global and mask-aware metrics.

2. Related Work

2.1 Historical Document Restoration

Diffusion models have recently been applied to historical document repair and document enhancement. DiffHDR [9] introduces HDR28K and uses a conditional diffusion model with semantic and spatial guidance for damaged Chinese historical documents. AutoHDR [12] addresses a more comprehensive full-page setting by combining damage localization, text context prediction, and patch-level appearance restoration. DocDiff [10] uses residual diffusion for document enhancement tasks such as deblurring, denoising, and removal of watermarks or seals. These studies show the strength of diffusion models for text-like structures, but usually assume task-specific training data or degradation patterns different from Japanese classical pages.

2.2 Japanese Document Analysis

The CODH resources provide open data for digital humanities and have supported machine learning studies on classical Japanese literature [6]. Clanuwat et al. [2] introduced Kuzushiji-MNIST and related benchmarks to connect machine learning with classical Japanese text analysis. Kaneko et al. [3] studied deteriorated character restoration using an enhanced CycleGAN, while KuzushijiDiffuser [11] explores diffusion-based kuzushiji font generation. These

[1] The University of Electro-Communications, Tokyo, Japan
[a)] gao-h@mm.inf.uec.ac.jp
[b)] yuan-h@mm.inf.uec.ac.jp
[c)] chen-j@mm.inf.uec.ac.jp
[d)] xiong-p@mm.inf.uec.ac.jp
[e)] yanai@cs.uec.ac.jp

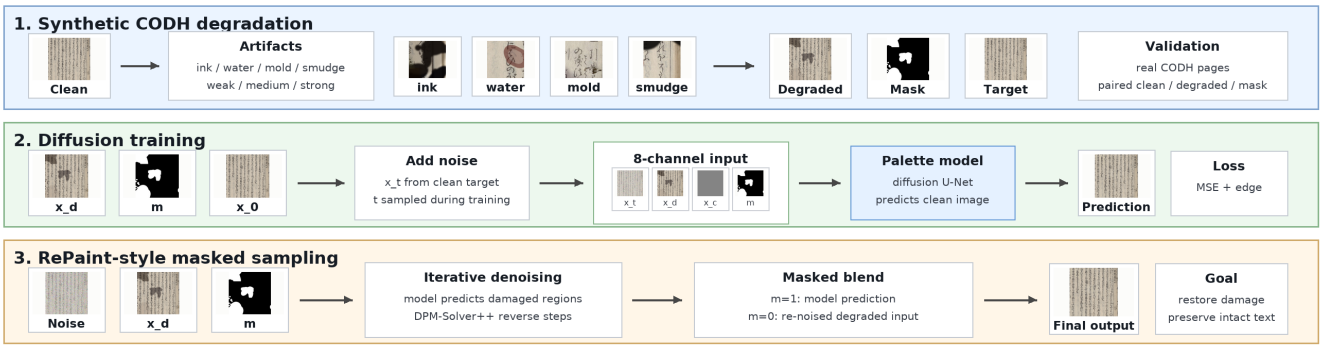


Fig. 1 Overview of the proposed framework. Clean CODH pages are synthetically degraded with four artifact types and three severity levels to generate degraded-clean pairs and oracle masks; a Palette-style diffusion model restores damaged regions with RePaint-style masked sampling.

works are domain-related but focus mainly on characters or fonts rather than damaged full-page images.

2.3 Image-to-Image Diffusion and Masked Sampling

Palette [7] shows that conditional diffusion can be used as a general image-to-image framework for tasks such as inpainting and image restoration. RePaint [5] demonstrates that diffusion sampling can be guided by a known-region mask so that generated content remains consistent with unmasked pixels. DPM-Solver++ [4] provides an efficient sampler for diffusion models. We adapt these ideas to Japanese classical document restoration by conditioning on a degraded page, an oracle mask, and a constant map, then enforcing known-region consistency during sampling.

3. Method

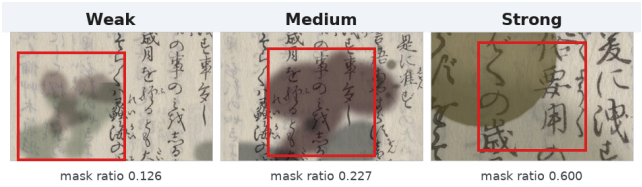
3.1 Synthetic Degradation

Clean CODH pages are split into 214 training pages and 24 validation pages before degradation synthesis, preventing page-level leakage. For each clean page, we generate three variants at each of three severity levels, resulting in 1,926 training triplets and 216 validation triplets. The degradation process combines four artifact types: ink stains, water stains, mold spots, and smudges. Each artifact is sampled with a text-density heatmap so that damage is likely to interact with written regions while still covering the page background.

The purpose of the synthetic design is twofold. First, it provides paired degraded-clean supervision for a domain where naturally paired damaged and undamaged pages are rarely available. Second, the severity levels create a controlled way to test whether a method only handles near-clean images or can also recover from large localized damage. The binary mask m is generated during synthesis and is treated as an oracle mask in this paper. Pixels with $m = 1$ are damaged and pixels with $m = 0$ are intact. This setting is useful for studying restoration and preservation separately, but it also means that automatic damage detection remains outside the present scope.

Table 1 Degradation parameters per level.

Parameter	Weak	Medium	Strong
Ink opacity	0.28–0.45	0.40–0.65	0.78–0.95
Ink count/variant	2–3	2–4	3–6
Stain size (% dim)	12–28%	15–40%	20–50%
Target damage ratio	6–10%	16–25%	35–50%
Max damage ratio	20%	35%	55%



Examples of the synthesized degradation levels. Representative CODH validation crops show increasing degradation severity from weak to strong; mask ratios indicate the damaged-pixel proportion used during synthesis.

3.2 Preliminaries: PaletteStyle

Palette [7] is an image-to-image diffusion framework that learns a conditional restoration distribution by denoising a noisy target image. At timestep t, the model receives a noisy image x_t derived from the clean target x_0 and one or more conditioning images. The network is trained to predict the clean target from this noisy state and the conditions, commonly called x_0-prediction. This formulation is suitable for restoration because the conditioning image can provide page layout, stroke positions, and visible background context while the diffusion process models uncertainty in the missing or corrupted regions.

In our setting, the condition consists of the degraded image x_d, the binary damage mask m, and a constant map x_c. The constant map is an all-ones channel following the Palette-style conditioning convention; it does not contain clean-target information. The target is the clean CODH image before synthetic degradation.

3.3 Proposed Diffusion Model

The denoising network takes an 8-channel input formed as $[x_t, x_d, x_c, m]$. Here, x_t and x_d are RGB images, x_c is a one-channel all-ones map, and m is a one-channel binary mask. The model predicts $\hat{x}_0$, the restored clean image. Providing

both x_d and m is important: the degraded image preserves the visible page content, while the mask tells the network which pixels should be regenerated and which pixels should mainly be copied or preserved.

This design differs from annotation-heavy document repair pipelines. We do not use OCR labels, character bounding boxes, or text recognition losses. The model therefore learns restoration from image pairs and masks alone. This makes the setting simpler and applicable to pages without transcriptions, but it also limits explicit character-level control over ambiguous strokes.

3.4 RePaint-style Masked Sampling

Without an explicit known-region constraint, diffusion sampling may slightly modify intact areas even when the damage mask is provided as input. Such changes can improve global visual smoothness but are undesirable for historical document restoration, where unchanged strokes and paper texture should remain stable. We therefore use RePaint-style masked sampling. At each reverse step, damaged regions use the model prediction, whereas intact regions are replaced by a re-noised version of the degraded input:

$$x_{t-1} = (1 - m) \cdot \mathrm{noise}(x_d, t - 1) + m \cdot x_{t-1}^{\mathrm{pred}}.$$

The sampler is implemented with DPM-Solver++ for efficient inference. This procedure explains the behavior observed in the mask-aware ablation: RePaint greatly improves $\mathrm{PSNR}_{\mathrm{known}}$ because known pixels are constrained by the degraded input, while $\mathrm{PSNR}_{\mathrm{mask}}$ changes little because damaged pixels are still generated by the same restoration model.

3.5 Training Objective

The training loss combines image reconstruction with a weak edge-preservation term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + 0.05\mathcal{L}_{\mathrm{edge}}.$$

The MSE term reconstructs the clean target and provides stable pixel-level supervision. The Sobel edge loss encourages the restored image to preserve stroke boundaries, which are important for classical Japanese characters and can become blurred under purely pixel-wise optimization. The weight 0.05 is an empirical choice used to keep the edge term auxiliary rather than dominant.

Training uses random 512×512 resized crops from the synthetic pages. The diffusion process uses 1000 DDPM steps with a linear beta schedule. We train with Adam, learning rate 10^{-4}, cosine warmup for 10 epochs, batch size 4 per GPU with 4-step gradient accumulation, fp16, EMA 0.9999, and 500 epochs. The model has 113.7M parameters.

4. Experiments

4.1 Evaluation Setting

The validation set contains 24 CODH pages, three variants, and three degradation levels, giving 216 validation images. Although 72 validation samples are available per level, all methods are evaluated on the same 15 center crops per level due to inference cost. The clean-page split is performed before synthesis, so no clean page appears in both training and validation. This protocol makes the comparison controlled: all methods receive the same degraded inputs and the same masks when a mask is required.

We report PSNR, SSIM, and LPIPS with AlexNet features as global image metrics. Global metrics alone can be misleading in masked restoration because most pixels may be intact, especially under weak degradation. We therefore also compute $\mathrm{PSNR}_{\mathrm{mask}}$ over damaged pixels and $\mathrm{PSNR}_{\mathrm{known}}$ over intact pixels. These mask-aware metrics separate two goals that are both important for document restoration: recovering corrupted regions and avoiding unnecessary changes to readable regions.

Compared methods include the degraded input as a lower bound, Telea and Navier-Stokes inpainting from OpenCV [1], LaMa [8], a cross-domain DiffHDR reference [9], and our model with RePaint disabled or enabled. DiffHDR is included as a reference rather than as a same-domain trained model, since its original training data targets Chinese historical document repair.

4.2 Quantitative Results

Table 2 reports the main comparison. At the weak level, the degraded input is already relatively close to the clean image because only a small portion of pixels is damaged. Even in this setting, the proposed method improves global PSNR and LPIPS, indicating that localized restoration can be performed without heavily disturbing the rest of the page. However, weak-level results should be interpreted with care because global metrics are strongly influenced by intact background and strokes.

The medium and strong levels show clearer differences. At medium degradation, LaMa remains a strong mask-based inpainting baseline with 27.5 dB PSNR, while Ours(+RP) reaches 39.9 dB. At strong degradation, Ours(+RP) reaches 29.1 dB and leads the best classical or LaMa baseline by 4.8 dB. The cross-domain DiffHDR reference performs worse under strong synthetic damage, which is expected because its training distribution and annotation assumptions differ from the Japanese CODH benchmark used here. These results suggest that same-domain synthetic training is useful when the target artifacts and page layout are controlled.

4.3 Mask-Aware Analysis

Table 3 Mask-aware RePaint ablation in dB, evaluated on the same 15 samples per level.

	Weak	**Medium**	**Strong**
Ours(-RP) $\mathrm{PSNR}_{\mathrm{mask}}$	34.1	32.7	23.8
Ours(+RP) $\mathrm{PSNR}_{\mathrm{mask}}$	34.0	32.6	23.8
$\Delta\mathrm{PSNR}_{\mathrm{mask}}$	−0.1	−0.0	−0.0
Ours(-RP) $\mathrm{PSNR}_{\mathrm{known}}$	41.7	42.0	41.0
Ours(+RP) $\mathrm{PSNR}_{\mathrm{known}}$	65.1	63.3	62.3
$\Delta\mathrm{PSNR}_{\mathrm{known}}$	+23.4	+21.3	+21.4

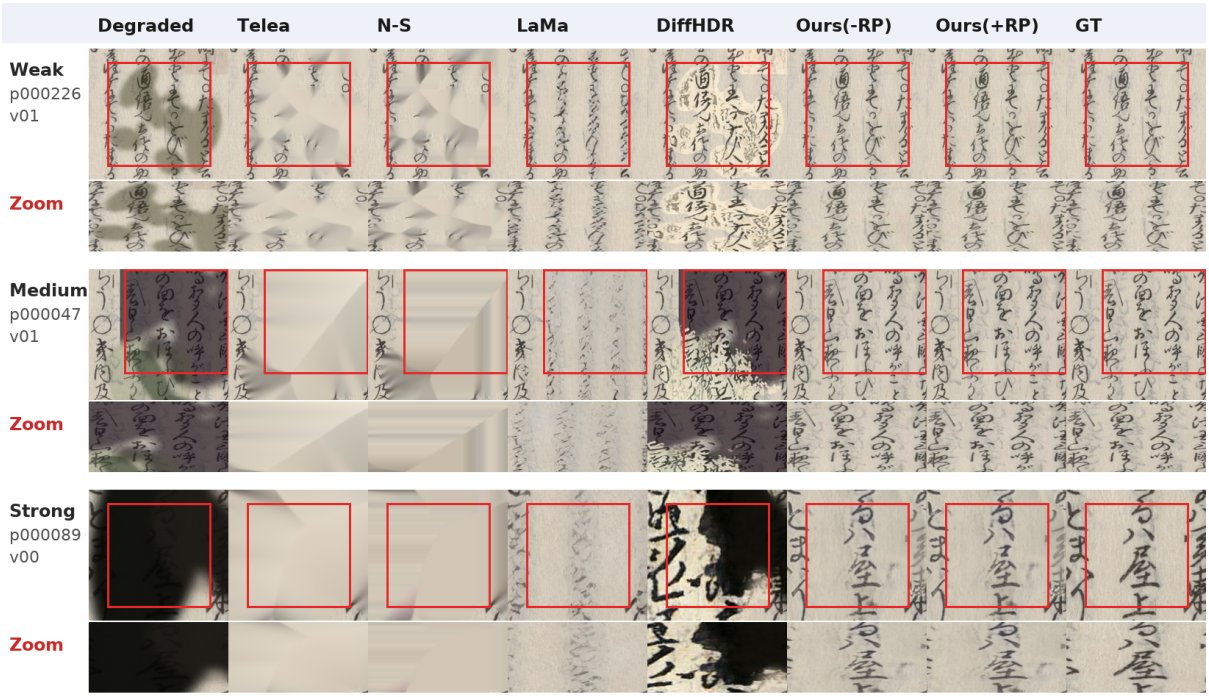


Fig. 2 Qualitative comparison on one weak, one medium, and one strong sample. Main panels show context crops around damaged regions; zoom rows further enlarge the red-boxed regions. Columns show Degraded, Telea, Navier-Stokes (N-S), LaMa, cross-domain DiffHDR, Ours without RePaint, Ours with RePaint, and GT.

Table 2 Quantitative comparison under the proposed synthetic benchmark. All methods are evaluated on the same 15 center crops per degradation level.

	Weak			Medium			Strong		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LowerBound	28.3	.9831	.0322	20.2	.9459	.0909	13.3	.7449	.2947
Telea	31.1	.9535	.0788	26.9	.8906	.1754	23.2	.7796	.3726
N-S	31.1	.9534	.0793	27.0	.8896	.1760	24.2	.7761	.3846
LaMa	31.5	.9512	.0270	27.5	.8846	.0599	23.9	.7447	.2560
DiffHDR	30.5	.9593	.0449	22.4	.8960	.1050	13.8	.6950	.3192
Ours(-RP)	40.2	.9882	.0130	38.0	.9844	.0167	28.0	.9024	.1266
Ours(+RP)	44.8	.9960	.0055	39.9	.9914	.0097	29.1	.9056	.1206

Ours(+RP) medians: Weak 44.9/.9968/.0015; Medium 39.9/.9919/.0028; Strong 28.6/.9506/.0627

Table 3 shows that RePaint-style sampling has a highly specific effect. It improves known-region PSNR by 21–23 dB across all degradation levels, while damaged-region PSNR remains almost unchanged. This is desirable for the restoration setting considered here. The model is still responsible for generating damaged pixels, but the sampler prevents unnecessary drift in regions that were not synthetically corrupted. The analysis also explains why mask-aware metrics are necessary: a method can improve global PSNR by preserving intact pixels even if damaged-region quality is similar, or conversely can restore damaged regions while making small but undesirable changes elsewhere.

4.4 Qualitative Results

Figure 2 illustrates the visual tendencies behind the quantitative results. Classical inpainting removes stains but often leaves smooth artifacts, and LaMa can oversmooth characters. Ours(-RP) restores masked regions but may alter intact pixels, whereas Ours(+RP) better preserves unmasked content.

5. Discussion

The proposed setting separates damaged-pixel restoration from intact-pixel preservation, which matters because readable strokes should not be beautified. RePaint-style sampling therefore helps by reducing known-region drift, although damaged-region generation is still handled by the same Palette-style model. The benchmark is synthetic, real pages may contain warping, bleed-through, scanning artifacts, or chemically complex stains, and practical use requires automatic or user-provided masks with robustness to mask noise. Image metrics should also be complemented by readability, OCR, full-page, and expert evaluation before making stronger claims for real archives.

6. Conclusion

We presented a synthetic benchmark and a conditional diffusion baseline for Japanese classical document restoration. The Palette-style model uses degraded images, damage masks, and a constant map, while RePaint-style masked sampling constrains intact regions. Under this benchmark, the method compares favorably with OpenCV inpainting, LaMa, and a cross-domain DiffHDR reference, especially at medium and strong degradation levels.

References

[1] Bradski, G.: The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000).

[2] Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K. and Ha, D.: Deep Learning for Classical Japanese Literature, arXiv:1812.01718 (2018).

[3] Kaneko, H., Ishibashi, R. and Meng, L.: Deteriorated Characters Restoration for Early Japanese Books Using Enhanced CycleGAN, *Heritage*, Vol. 6, No. 5, pp. 4345–4361 (2023).

[4] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. and Zhu, J.: DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models, arXiv:2211.01095 (2022).

[5] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R. and Van Gool, L.: RePaint: Inpainting Using Denoising Diffusion Probabilistic Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471 (2022).

[6] ROIS-DS Center for Open Data in the Humanities: CODH Japanese Dataset, `https://codh.rois.ac.jp/` (2026). National Institute of Informatics, Research Organization of Information and Systems.

[7] Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J. and Norouzi, M.: Palette: Image-to-Image Diffusion Models, *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10 (2022).

[8] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K. and Lempitsky, V.: Resolution-Robust Large Mask Inpainting with Fourier Convolutions, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2149–2159 (2022).

[9] Yang, Z., Peng, D., Shi, Y., Zhang, Y., Liu, C. and Jin, L.: Predicting the Original Appearance of Damaged Historical Documents, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9382–9390 (2025).

[10] Yang, Z., Liu, B., Xiong, Y., Yi, L., Wu, G., Tang, X., Liu, Z., Zhou, J. and Zhang, X.: DocDiff: Document Enhancement via Residual Diffusion Models, *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2795–2806 (2023).

[11] Yuan, H. and Yanai, K.: Japanese Kuzushiji Font Generation with FontDiffuser, *Proceedings of International Conference on MultiMedia Modeling*, pp. 142–149 (2025).

[12] Zhang, Y., Zhang, P., Yang, Z., Yan, P., Shi, Y., Liu, P., Guo, F. and Jin, L.: Reviving Cultural Heritage: A Novel Approach for Comprehensive Historical Document Restoration, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 28876–28892 (2025).