Predicting the Original Appearance of Damaged Historical Documents

Zhenhua Yang[1*], Dezhi Peng[1*], Yongxin Shi[1], Yuyi Zhang[1], Chongyu Liu[1], Lianwen Jin[12†]

[1]South China University of Technology
[2]INTSIG-SCUT Joint Lab on Document Analysis and Recognition
{eezhyang, pengdezhi000, scut.yxshi, yuyizhang.scut, liuchongyu1996}@gmail.com, eelwjin@scut.edu.cn

Abstract

Historical documents encompass a wealth of cultural treasures but suffer from severe damages including character missing, paper damage, and ink erosion over time. However, existing document processing methods primarily focus on binarization, enhancement, etc., neglecting the repair of these damages. To this end, we present a new task, termed Historical Document Repair (**HDR**), which aims to predict the original appearance of damaged historical documents. To fill the gap in this field, we propose a large-scale dataset **HDR28K** and a diffusion-based network **DiffHDR** for historical document repair. Specifically, HDR28K contains 28,552 damaged-repaired image pairs with character-level annotations and multi-style degradations. Moreover, DiffHDR augments the vanilla diffusion framework with semantic and spatial information and a meticulously designed character perceptual loss for contextual and visual coherence. Experimental results demonstrate that the proposed DiffHDR trained on HDR28K significantly surpasses existing approaches and exhibits remarkable performance in handling real scenarios. Notably, DiffHDR can also be extended to document editing and text block generation, showcasing its high flexibility and generalization capacity. We believe this study could pioneer a new direction of document processing and contribute to the inheritance of invaluable cultures and civilizations.

Dataset, Code — https://github.com/yeungchenwa/HDR

Introduction

Historical documents play a pivotal role in the transmission of cultural heritage. However, during prolonged preservation, they are susceptible to oxidization, insect damage, water erosion, *etc*., leading to character missing, paper damage, and ink erosion, as shown in Figure 1. Nevertheless, the manual repair of damaged characters and corrupted backgrounds is a complex and time-consuming endeavor.

Recently, generic document image processing primarily concentrates on the low-level vision tasks, such as rectification (Li et al. 2023a; Jiang et al. 2022), binarization (Yang et al. 2023a; Yang and Xu 2023), enhancement (Hertlein and Naumann 2023; Wang et al. 2022a; Xue et al. 2022), and deshadowing (Li et al. 2023c; Lin, Chen, and Chuang 2020). However, they fall short of understanding the semantics and stylistic elements within document images, thereby hindering their capability to repair the damaged documents. Moreover, existing historical document processing methods also address tasks such as text restoration (Assael et al. 2022) and individual character restoration (Nguyen et al. 2019; Amin, Siddiqi, and Moetesum 2023); however, these unimodal methods are unsuitable for document repair because predicting the original appearance of damaged documents is a highly challenging multimodal task, requiring the understanding of the context and the pixel-level restoration. Though the recent work (Zhu et al. 2024) conducts inscription restoration task, it specifically targets inscriptions with simple backgrounds, exclusively characterized by white text on a black background. Additionally, the font generation task (Kong et al. 2022; Wang et al. 2023; Yang et al. 2023b) exhibits a higher resemblance, involving character generation under the conditions of content and style. Nevertheless, this task is only employed for individual characters and it is not feasible to reconstruct the documents.

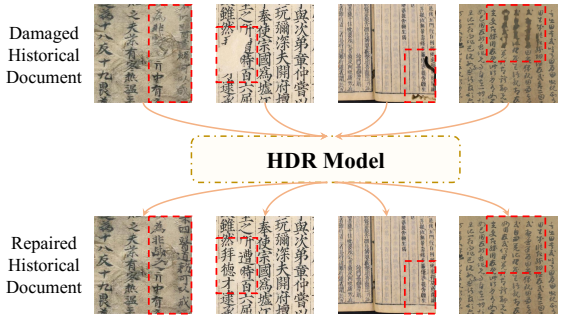


Figure 1: Definition of Historical Document Repair (HDR) task. The green boxes represent the damaged regions and the blue boxes denote the repaired regions.

Therefore, to fill the gap in this field, we introduce a new task, termed *Historical Document Repair* (**HDR**), which involves predicting the original appearance of damaged historical document images. As shown in Figure 1, the damaged historical document images are fed into the HDR model to repair the damaged regions. The output images of HDR

*These authors contributed equally.
†Corresponding author

model, termed repaired images, should not only capture precise character content and style but also harmonize with the surrounding background within the repaired region.

As there is no dataset available for historical document repair, we contribute a large-scale dataset, named **HDR28K**, which comprises a total of 28,552 damaged-repaired image pairs with character-level annotations and multi-style degradation. As shown in Figure 2, the undamaged images are corrupted by three meticulously designed degradations to simulate character missing, paper damage, and ink erosion, which is intended to faithfully replicate the visual effects of damages observed in historical documents.

Additionally, to facilitate the development of HDR task, we propose **DiffHDR**, a Diffusion-based Historical Document Repair network, which frames the HDR task as a series of diffusion steps that progressively transform the damaged regions to match the target character content and character style with an accurate background. In our method, we first crop the damaged region from the historical document to obtain a fixed-size damaged patch image, then DiffHDR leverages the damaged images, along with semantic and spatial information, as conditions for appearance reconstruction. To further improve the content preservation of repaired characters, we introduce a character perceptual loss to penalize the misalignment of character features.

Extensive experiments demonstrate that the models trained using HDR28K can reconstruct the original appearance of damaged historical document images and achieve state-of-the-art performance. Moreover, we have gathered a collection of real damaged samples from the Internet and applied our method for repair, which shows that DiffHDR, trained on synthetic data, is proficient in real scenarios, highlighting its significant potential for the preservation of cultural heritage. Furthermore, DiffHDR can be extended to document editing and text block font generation, exhibiting the flexibility and generalization of our proposed method.

We summarize our main contributions as follows:

- We introduce a Historical Document Repair (HDR) task, which endeavors to predict the original appearance of damaged historical document images.
- We build a large-scale historical document repair dataset, termed HDR28K, which includes 28,552 damaged-repaired image pairs with character-level annotations and multi-style degradation.
- We propose a Diffusion-based Historical Document Repair method (DiffHDR), which augments the DDPM framework with semantic and spatial information and incorporates a meticulously designed character perceptual loss to enhance the contextual and visual coherence.
- DiffHDR trained on HDR28K outperforms other methods and is capable of repairing real damaged historical documents. Moreover, our method can be extended to document editing and text block font generation.

Related Work

Image Restoration

Generic image restoration methods (Li et al. 2023b; Cui et al. 2023; Chen et al. 2022; Wang et al. 2022b; Zamir et al. 2022) are primarily focused on image deraining, defogging, deblurring, or denoising. For example, Restormer (Zamir et al. 2022) proposes an efficient transformer model to capture long-range pixel interactions and is applicable to large images. In document image restoration, they mainly address the tasks of rectification (Li et al. 2023a; Jiang et al. 2022; Das et al. 2019), binarization (Yang et al. 2023a; Yang and Xu 2023), enhancement (Hertlein and Naumann 2023; Wang et al. 2022a), and deshadowing (Li et al. 2023c; Lin, Chen, and Chuang 2020). Although the above methods have achieved remarkable performance, they cannot comprehend the semantics and stylistic elements present in document images, thereby hindering the repair of damaged documents.

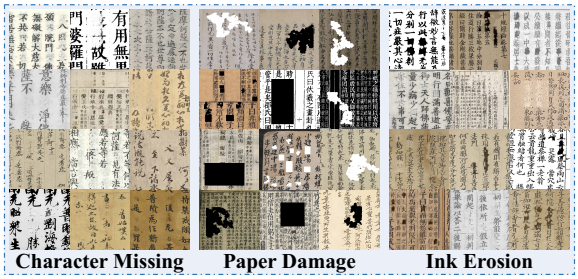


Figure 2: Damaged Samples in HDR28K.

Historcial Document Image Processing

Some approaches have been proposed for historical document image processing. Ithaca (Assael et al. 2022) utilizes the transformer block to conduct the sequence modeling for textual restoration, geographical attribution, and chronological attribution. Some methods (Nguyen et al. 2019; Amin, Siddiqi, and Moetesum 2023) focus on individual character restoration. (Amin, Siddiqi, and Moetesum 2023) focuses on the isolated Greek characters and applies an auto-encoder to reconstruct the missing parts of characters. Moreover, to alleviate the unreadability of damaged historical documents, (Ech-Cherif and Cheriet 2022) propose a multi-task learning module to conduct the binarization task. Furthermore, some methods (Hedjam and Cheriet 2013; Raha and Chanda 2019; Wadhwani et al. 2021) are proposed to conduct a historical document image enhancement. Nevertheless, these unimodal methods prove inadequate for document repair, as historical document repair is a highly challenging multimodal task, which demands an understanding of both the context and pixel-level restoration. The recent work (Zhu et al. 2024) introduces an inscription restoration dataset; however, it lacks diversity in styles. Thus, to fill the gap in this field, we contribute a large-scale historical document repair dataset with diverse complex backgrounds, termed HDR28K, and propose a diffusion-based model.

Historical Document Repair

The objective of historical document repair (HDR) is to accurately predict the original appearance of damaged historical document images. Specifically, as illustrated in Figure 1, when presented with a damaged historical document image

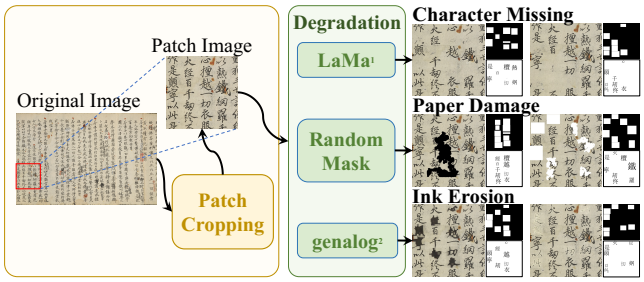


Figure 3: Construction pipeline of the HDR28K dataset.

$\boldsymbol{x}_d$, the HDR model focuses on reconstructing the original state, obtaining the repaired image $\boldsymbol{x}_r$. This reconstruction process requires the HDR model to repair both the damaged characters and the corrupted background. The repaired output should not only precisely capture the content and style of characters but also seamlessly integrate with the surrounding background within the repaired region. Thus, the repaired result $\boldsymbol{x}_r$ can be formulated as follows:

$$\boldsymbol{x}_r = \mathcal{F}_{HDR}(\boldsymbol{x}_d). \tag{1}$$

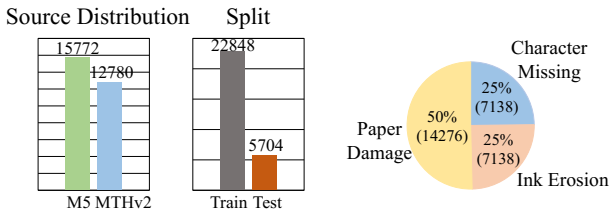
$\mathcal{F}_{HDR}$ denotes the historcial document repair model. In our work, we leverage the priors of semantic and spatial cues (provided by our HDR28K dataset) to support the repair of damaged historical documents. Thus, the repair of our method is as follows:

$$\boldsymbol{x}_r = \mathcal{F}(\boldsymbol{x}_d, \boldsymbol{x}_c, \boldsymbol{x}_m), \tag{2}$$

where $\boldsymbol{x}_c$ represents the content image (senmantic prior) and $\boldsymbol{x}_m$ denotes the mask image (spatial prior).

HDR28K Dataset

As there is no dataset specifically designed for historical document repair, we construct HDR28K, which consists of 28,552 damaged-repaired image pairs with OCR annotations. In this section, we provide the construction details and analysis of the proposed dataset.



(a) Sample Distribution (b) Degradation Distribution

Figure 4: Statistics of the HDR28K dataset.

Data Construction

To accurately simulate real damaged scenarios in historical documents, it is necessary to degrade the characters, which requires character-level bounding boxes and content annotations. Therefore, we constructed the HDR28K dataset by building upon MTHv2 (Ma et al. 2020) and M5HisDoc (Shi et al. 2023) and implementing meticulously designed degradations. Specifically, as illustrated in Figure 3, for efficiency in memory and computation, we first crop 512×512 patch images from high-resolution original images using sliding windows. During cropping, our automated schemes focus exclusively on text regions, and we manually filter out the images that are low resolution or lack text intensity. Subsequently, we apply three degradations on patch images, which replicate the real scenarios of character missing, paper damage, and ink erosion in damaged situations.

The details of the three degradations are: **(1) Character Missing**: We randomly generate masks and employ LAMA (Suvorov et al. 2022) to erase the content in mask regions. The generated masks consist of character-level and block-level types. Because MTHv2 and M5HisDoc provide the location annotations for corresponding individual character regions, we randomly select some of these character regions as the character-level masks. In the generation of block-level masks, we randomly sample a rectangular region from the patch image as the erasing mask. **(2) Paper Damage**: Due to insect infestation, oxidization, contamination, etc., the papers in historical documents suffer from severe damage. To replicate this scenario, we randomly mask some regions in the patch image using black or white pixels. Similar to the character missing, the masked regions include character-level and block-level types, but they take the form of either a rectangular or an irregular shape. **(3) Ink Erosion**: We utilize genalog[1] to simulate scenarios involving water erosion and character fading. We first randomly sample rectangular regions from patch images similar to the mask generation in character missing. Then we apply diverse degradation modes and convolution kernels in genalog to induce degradation to the sampled regions. The examples of the above three degradations are listed on the right of Figure 3.

Data Analysis

We randomly select 536 original images from the testing set of MTHv2 and 891 original images from the testing set of M5HisDoc to construct our HDR28K testing set. The HDR28K training set is sourced from the remaining samples in the two datasets. Note that the patch images from the same historical documents are not assigned to both training and testing sets. As shown in Figure 4(a), after the cropping in the construction pipeline, the training set in HDR28K comprises 22,848 patch images, while the testing set consists of 5,704 patch images. Moreover, 12,780 patch images originate from MTHv2 (Ma et al. 2020) while 15,772 patch images are sourced from M5HisDoc (Shi et al. 2023). As depicted in Figure 4(b), the degradation of paper damage accounts for 50% of the HDR dataset, while the other two degradations account for 25%, respectively. Finally, we present some samples in Figure 2, which demonstrates that our dataset can realistically replicate the damage observed in historical document images.

[1]https://github.com/microsoft/genalog

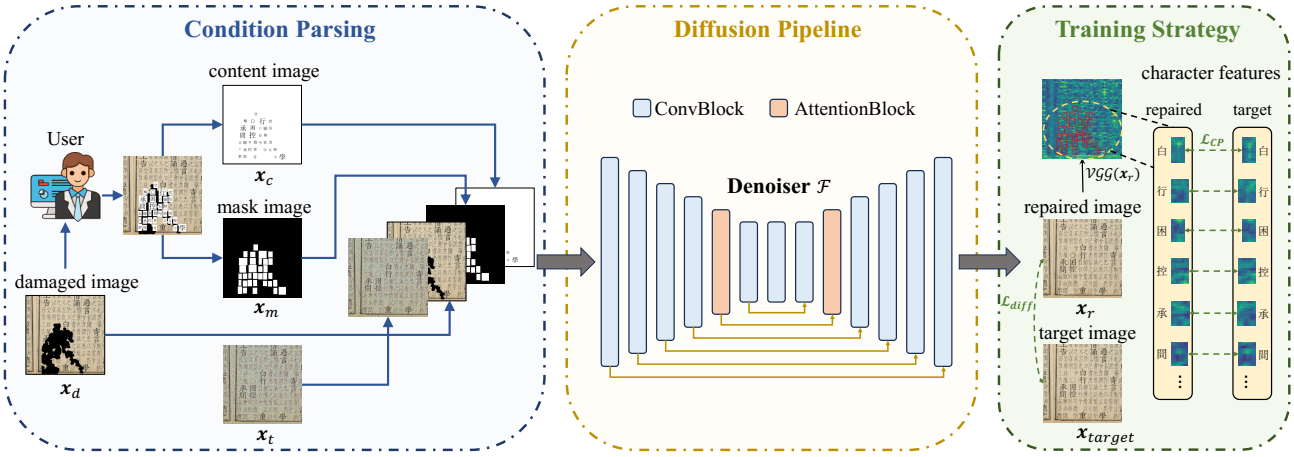


Figure 5: Overview of our proposed method. DiffHDR comprises a condition parsing and a diffusion pipeline. In the condition parsing, the user provides the content and location of damaged characters, obtaining the content image $\boldsymbol{x}_c$ and mask image $\boldsymbol{x}_m$. In the diffusion pipeline, our denoiser $\mathcal{F}$, a UNet-based network, outputs the repaired image $\boldsymbol{x}_r$ conditioned on noised image $\boldsymbol{x}_t$, damaged image $\boldsymbol{x}_d$, mask image $\boldsymbol{x}_m$ and content image $\boldsymbol{x}_c$. During training, in addition to using diffusion loss $\mathcal{L}_{diff}$, we introduce a character perceptual loss $\mathcal{L}_{CP}$ to enhance the content preservation of repaired characters.

DiffHDR: Diffusion-based HDR Network

Framework

As depicted in Figure 5, the framework of DiffHDR consists of a condition parsing and a diffusion pipeline. During condition parsing, the user provides the content and location of the damaged characters and we parse them out to obtain the content image $\boldsymbol{x}_c$ and the mask image $\boldsymbol{x}_m$. Subsequently, our diffusion pipeline gathers the damaged image $\boldsymbol{x}_d$ with the user's guidance to predict the original appearance $\boldsymbol{x}_r$ of the damaged image $\boldsymbol{x}_d$.

Specifically, we randomly sample a time step $t \sim Uniform(0, T_{max})$ and a Gaussian noise $\boldsymbol{\epsilon}_t$ to corrupt the damaged image x_0, yielding the noised image $\boldsymbol{x}_t$ following (Ho, Jain, and Abbeel 2020):

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t}(1-\beta_i)$, $\beta_i \sim (0,1)$. Then we concatenate $\boldsymbol{x}_t \sim \mathbb{R}^{3\times H\times W}$, $\boldsymbol{x}_d \sim \mathbb{R}^{3\times H\times W}$, $\boldsymbol{x}_c \sim \mathbb{R}^{1\times H\times W}$ and $\boldsymbol{x}_m \sim \mathbb{R}^{1\times H\times W}$ in channel dimension as a 8-channel input to the following denoiser $\mathcal{F}$. Our denoiser $\mathcal{F}$ is a UNet-based network, which directly predicts the repaired result $\boldsymbol{x}_r$ rather than the added noise $\boldsymbol{\epsilon}_t$. In this manner, our method is capable of performing pixel-level repairs, significantly reducing both labor and time costs while advancing the field of digital humanities.

Training Objective

Diffusion Loss Because our proposed method directly predicts the original appearance $\boldsymbol{x}_r$ of the damaged images rather than the added noise $\boldsymbol{\epsilon}_t$, we optimize DiffHDR with the diffusion loss as follows:

$$\mathcal{L}_{diff} = \left\|\boldsymbol{x}_{target} - \mathcal{F}(\boldsymbol{x}_t; \boldsymbol{x}_d, \boldsymbol{x}_c, \boldsymbol{x}_m)\right\|^2, \tag{4}$$

where $\boldsymbol{x}_{target}$ denotes the target image.

Character Perceptual Loss To further improve the content preservation of repaired characters, we introduce a Character Perceptual Loss (CPLoss) to provide guidance to our model. Specifically, as shown in right of Figure 5, we first utilize the pretrained VGG (Simonyan and Zisserman 2014) to extract the feature $\mathcal{VGG}(\boldsymbol{x}_r)$ from the repaired image $\boldsymbol{x}_r$. Then we penalize the misalignment of feature between $\mathcal{VGG}(\boldsymbol{x}_r)$ and the target feature $\mathcal{VGG}(\boldsymbol{x}_{target})$ within the repaired regions. The CPLoss is formulated as follows:

$$\mathcal{L}_{CP} = \sum_{i=1}^{L} \omega_i(\|\mathcal{VGG}_i(\boldsymbol{x}_r) - \mathcal{VGG}_i(\boldsymbol{x}_{target})\|)\boldsymbol{x}_m, \tag{5}$$

where $\mathcal{VGG}_i$ represents the i-th VGG layer feature. ω_i denotes the layer weight. To effectively capture both global and local representations of repaired characters, we utilize multi-scale features to penalize the misalignment. $\boldsymbol{x}_m$ enables the concentration of DiffHDR solely on damaged characters. CPLoss not only ensures the preservation of content and style for characters within the repaired regions but maintains the compatibility of the repaired background as well.

Attribute-Sensitive Repair Strategy

In HDR task, our denoiser $\mathcal{F}$ has three attributes: the damaged image $\boldsymbol{x}_d$, content image $\boldsymbol{x}_c$ and mask image $\boldsymbol{x}_m$. Inspired by InstructPix2Pix (Brooks, Holynski, and Efros 2023), it is beneficial to utilize the classifier-free guidance (Ho and Salimans 2022) in relation to the conditional inputs. Therefore, during training, we randomly set only $\boldsymbol{x}_d = \emptyset$, both $\boldsymbol{x}_c = \emptyset$ and $\boldsymbol{x}_m = \emptyset$, and all $\boldsymbol{x}_d = \emptyset$, $\boldsymbol{x}_c = \emptyset$ and $\boldsymbol{x}_m = \emptyset$ with an 8% probability, respectively (where $\emptyset$ indicates setting the unconditional inputs to pixel values of 255 or 0). This strategy enables our method more sensitive to the three attributes.

During sampling, we introduce the guidance scales s_d and $s_{c,m}$, which can be viewed as the sensitivity of the repaired

Model	FID↓	LPIPS↓	Rec-ACC(%)↑
UNet (Ronneberger, Fischer, and Brox 2015)	1.5504	0.0638	55.5547
Pix2Pix-ResNet (Isola et al. 2017)	7.9586	0.0821	38.7583
Pix2Pix-UNet (Isola et al. 2017)	12.4989	0.0816	39.7935
CycleGAN-ResNet (Zhu et al. 2017)	4.6521	0.0935	33.2306
CycleGAN-UNet (Zhu et al. 2017)	12.4847	0.1192	27.5988
Uformer-Tiny (Wang et al. 2022b)	1.4743	0.0626	57.7111
Uformer-Small (Wang et al. 2022b)	1.1858	0.0547	67.5760
Uformer-Big (Wang et al. 2022b)	1.026	0.0510	<u>69.9545</u>
Restormer (Zamir et al. 2022)	1.1632	0.0584	60.4338
NAFNet (Chen et al. 2022)	<u>0.8588</u>	<u>0.0435</u>	69.5821
GRL (Li et al. 2023b)	3.5562	0.0819	41.0406
FocalNet (Cui et al. 2023)	14.7901	0.1034	44.0407
UPOCR (Peng et al. 2023)	6.5690	0.0903	40.1415
Ours	**0.7499**	**0.0384**	**81.9180**

Table 1: Quantitative comparison. The bold indicates the state-of-the-art and the underline indicates the second best.

results with the damaged image $\boldsymbol{x}_d$ and the content and location cues $\boldsymbol{x}_c$, $\boldsymbol{x}_m$, respectively. Thus, the repair strategy can be formulated as:

$$\begin{aligned}\tilde{\mathcal{F}}(\boldsymbol{x}_t;\boldsymbol{x}_d,\boldsymbol{x}_c,\boldsymbol{x}_m) &= \mathcal{F}(\boldsymbol{x}_t;\emptyset,\emptyset,\emptyset) \\ &+ s_d(\mathcal{F}(\boldsymbol{x}_t;\boldsymbol{x}_d,\emptyset,\emptyset) - \mathcal{F}(\boldsymbol{x}_t;\emptyset,\emptyset,\emptyset)) \\ &+ s_{c.m}(\mathcal{F}(\boldsymbol{x}_t;\boldsymbol{x}_d,\boldsymbol{x}_c,\boldsymbol{x}_m) - \mathcal{F}(\boldsymbol{x}_t;\boldsymbol{x}_d,\emptyset,\emptyset)). \quad (6)\end{aligned}$$

Experiment

Evaluation Metrics

We utilize FID (Heusel et al. 2017), LPIPS (Zhang et al. 2018), and the accuracy of character recognizer (Rec-ACC) for quantitative comparison. FID measures the distribution distance between the output and the target domain. LPIPS is closer to human visual perception. We utilize a trained character recognizer to evaluate the accuracy of the individual characters within the repaired regions in $\boldsymbol{x}_r$ and the result is called Rec-ACC. We employ VGG19 (Simonyan and Zisserman 2014) as the character recognizer and it is trained using all individual character data in MTHv2 and M5HisDoc. Because HDR task focuses on the repaired region, we replace the non-damaged region of $\boldsymbol{x}_r$ by the target $\boldsymbol{x}_{target}$ before the evaluation. In short, FID and LPIPS are the image-level metrics, while Rec-ACC is the instance-level metric. Additionally, we do not evaluate two commonly used metrics, PSNR and SSIM, as they are unsuitable for HDR task (Zhang et al. 2018). The evidence is provided in Section 6.3.

Implementation Details

We adopt an AdamW optimizer to train DiffHDR with $\beta_1 = 0.95$ and $\beta_2 = 0.999$. The image size is 512×512. During classifier-free training, we set the conditional dropout probability as 8% and we train the model with a batch size of 32 and a total epoch of 165. The learning rate is set as 1×10^{-4} with the linear schedule. The training is conducted on 8 NVIDIA A6000 GPUs. We set the guidance scales $s_d = 1.2$ and $s_{c,m} = 1.5$ and adopt the DPM-Solver++ (Lu et al. 2022) as our sampler with the inference step of 20.

Comparison with Existing Methods

We compare our method with 9 methods, including GAN-based methods (Pix2Pix and CycleGAN), CNN-based methods (UNet, NAFNet, and FocalNet), and Transformer-based methods (Uformer, Restormer, GRL, and UPOCR). Since these existing methods are not originally designed for the HDR task, they are adapted to use the concatenation of damaged image $\boldsymbol{x}_d$, content image $\boldsymbol{x}_c$, and mask image $\boldsymbol{x}_m$ as a 5-channel input and generate a 3-channel repaired image $\boldsymbol{x}_r$ as output. For a fair comparison, all these methods are trained on 8 NVIDIA A6000 GPUs with the same batch size and epochs as DiffHDR. Moreover, we adopt ResNet or UNet as the generator of Pix2Pix and CycleGAN, as shown in the *2nd* to *5th* rows of Table 1. Note that we do not conduct the experiments on the CIRI dataset (Zhu et al. 2024), nor do we use its method for comparison, as they have not yet been made open-source.

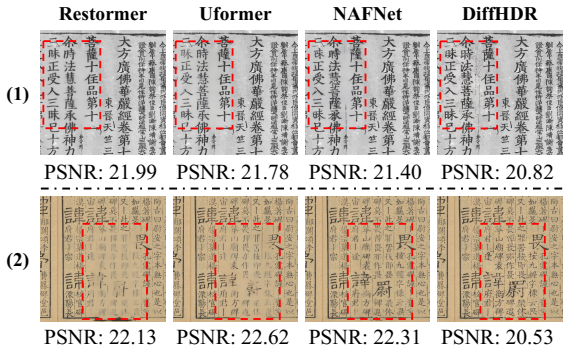
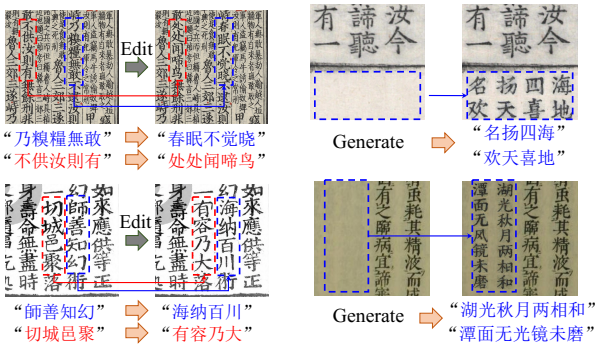


Figure 6: Unsuitableness of PSNR and SSIM.



(a) Document editing. (b) Text block font generation.

Figure 7: Document editing and text block font generation.

Quantitative Comparison The quantitative results are shown in Table 1. DiffHDR achieves state-of-the-art performance, surpassing other methods by a substantial margin in FID, LPIPS and Rec-ACC. Our method achieves a 12.7% lower FID and an 11.7% lower LPIPS compared to the second-best NAFNet. Notably, DiffHDR outperforms the second-best Uformer-Big by 11.9635% in Rec-ACC, highlighting its advantage in character correctness. We find that PSNR and SSIM are unsuitable for HDR task. Supporting evidence is presented in Figure 6, where blurring causes higher values of PSNR and SSIM (Zhang et al. 2018).

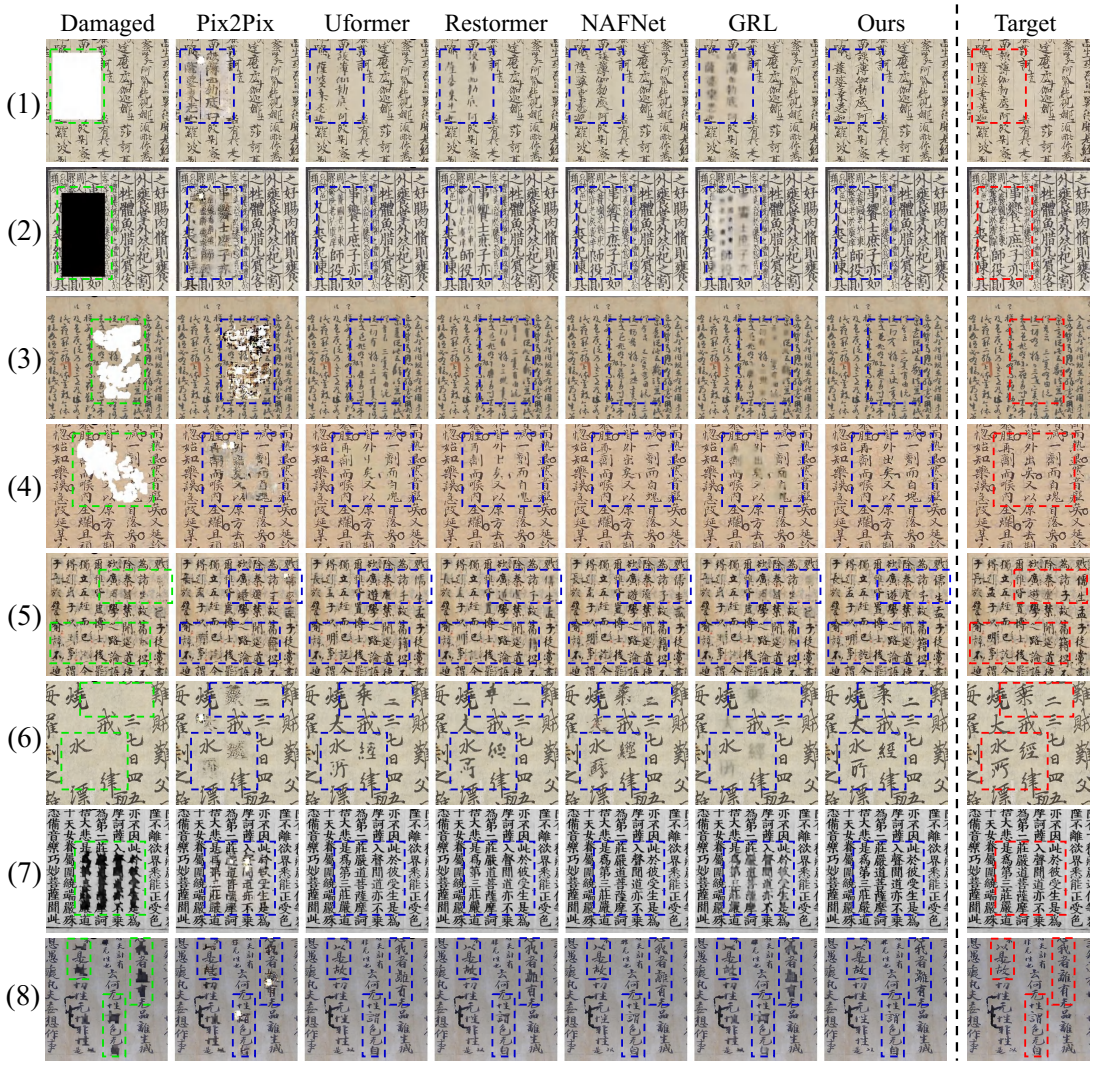


Figure 8: Qualitative comparison. We visualize the results of some evaluated methods. The green, blue, and red boxes represent the damaged regions, the repaired regions, and the target, respectively.

Qualitative Comparison As illustrated in Figure 8, we present the visualizations of DiffHDR and the existing methods on HDR28K testing set. DiffHDR can reconstruct the original appearance of damaged images with both realism and high quality. The second-best NAFNet and the third-best Uformer-Big encounter the problems such as blurring (see Figure 8(1)(3)(5)), missing character strokes (see Figure 8(1)(2)(4)(6)) and style inconsistency of background (see Figure 8(3)(7)) within the repaired regions. In contrast, DiffHDR excels in these aspects and shows the superiority on the generation of scribble characters (see Figure 8(1)(3)(5)(8)), complex characters (see Figure 8(2)(4)(7)), intensive text (see Figure 8(1)(2)(3)), and complex background (see Figure 8(4)(5)) within repaired regions.

Real Damaged Document Image Repair

In this section, we utilize the trained DiffHDR to repair real damaged historical document images that are obtained from the Internet. As shown in Figure 9, DiffHDR is capable of generating realistic characters coherent with the background during the repair, demonstrating the adaptability of our method in real-world scenarios. Additionally, it validates the appropriateness of our HDR28K dataset though it is constructed through synthetic degradations. Note that real damaged-repair image pairs in historical documents are exceptionally rare, making it challenging to collect sufficient data for evaluation purposes. In the future, we intend to work with relevant restoration institutions to acquire real damaged-repair pairs to address this issue.

Effectiveness of CPLoss $\mathcal{L}_{CP}$

We investigate the advantage of the proposed Character Content Perceptual Loss $\mathcal{L}_{CP}$, in which we trained the DiffHDR with and without $\mathcal{L}_{CP}$. As shown in Figure 2, incorporating the CPLoss improves the repair performance in terms of FID, LPIPS, and Rec-ACC.

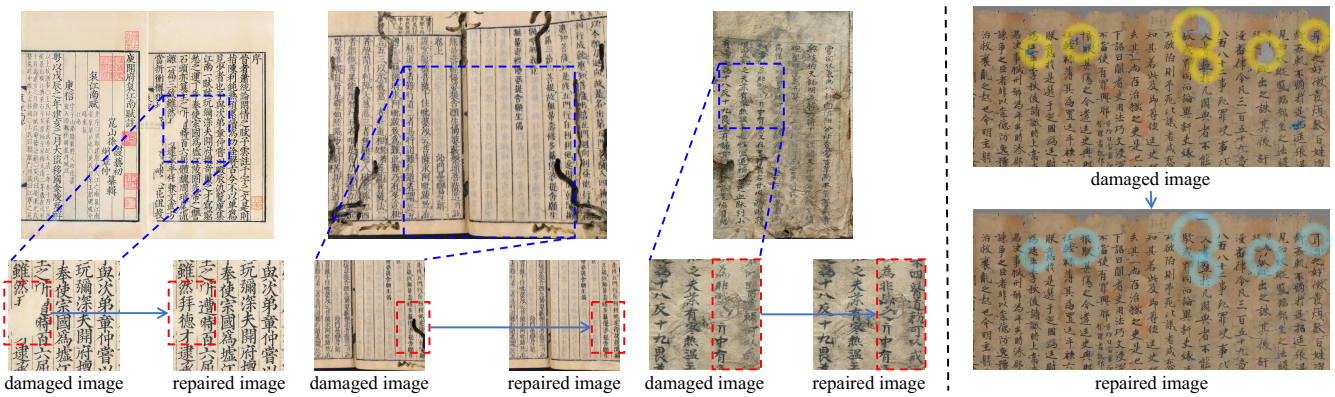


Figure 9: Real damaged historical documents repair by DiffHDR.

Method	FID↓	LPIPS↓	Rec-ACC(%)↑
wo $\mathcal{L}_{CP}$	0.8416	0.0450	64.4897
w $\mathcal{L}_{CP}$	**0.7499**	**0.0384**	**81.9180**

Table 2: Effectiveness of character perceptual loss $\mathcal{L}_{CP}$.

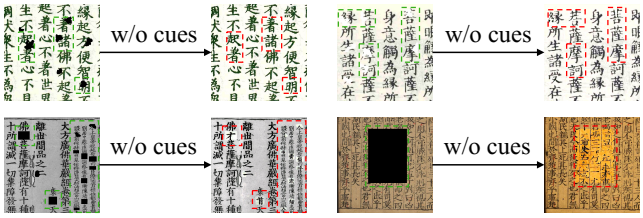


Figure 10: Damaged historical documents repair by DiffHDR when not provided with semantic and spatial cues.

Editing and Text Block Font Generation

In this section, we explore the capabilities of DiffHDR in historical document editing and text block font generation. **(1)** Document editing is to modify the text content to our target while maintaining consistency of style in the edited characters and the surrounding background. In our method, given the edited location $\boldsymbol{x}_m$ and content $\boldsymbol{x}_c$, we mask the input image by $\boldsymbol{x}_m$ and feed it to DiffHDR. As shown in Figure 7(a), our method generates readable characters, which are also harmonized with the surrounding background. **(2)** Text block font generation is to generate a group of characters within the specified region, while the text block adopts the style of the remaining areas. As shown in Figure 7(b), DiffHDR generates characters are coherent with the background, though the background area is filled with noise.

Limitation

As depicted in the 1*st* row of Figure 10, when DiffHDR is not provided with the semantic and spatial information of damaged characters (setting $\boldsymbol{x}_c$ and $\boldsymbol{x}_m$ to pixel 255), our method can comprehend the character semantics and repair the characters correctly. However, when the damage is severe, our method is unable to repair the image in the absence of semantic and spatial information, as shown in the 2*nd* row of Figure 10. For future work, we will address the prediction of damaged character content and location. Specifically, we plan to leverage large vision-language models (such as Qwen2-VL(Wang et al. 2024) and InternVL2(Chen et al. 2024)) for automated content prediction of damaged characters and to train a detection model (such as the DINO(Zhang et al. 2022)) for identifying damaged character locations. Moreover, we will collect a certain amount of real damaged-repair image pairs to better evaluate the repair performance of real damaged documents using different methods.

Conclusion

In this paper, we introduce a new task, *Historical Document Repair* (HDR), which aims to predict the original appearance of damaged historical documents. To fill the blank in this field, we contribute a large-scale HDR dataset, named HDR28K, which contains 28,552 damaged-repaired document image pairs and employs three meticulously designed synthetic degradations to simulate real damages typically observed in historical documents. Furthermore, a novel DiffHDR model is proposed to solve the HDR problem. Specifically, DiffHDR follows a diffusion-based paradigm conditioned on semantic and spatial priors for context correctness and visual truthfulness. During training, a new character perceptual loss is incorporated to enhance the content preservation of repaired characters. Extensive experiments demonstrate that DiffHDR achieves state-of-the-art performance and is capable of repairing real damaged documents though trained with synthetic damages of HDR28K. Thanks to its highly flexible framework, DiffHDR also exhibits impressive performance in document editing and text block generation. We believe this study could be the cornerstone of the new HDR field and significantly contribute to the preservation of invaluable cultural heritage.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (Grant No.: 62441604, 62476093) and IntSig-SCUT Joint Lab Foundation.

References

Amin, J.; Siddiqi, I.; and Moetesum, M. 2023. Reconstruction of Broken Writing Strokes in Greek Papyri. In *International Conference on Document Analysis and Recognition*, 253–266. Springer.

Assael, Y.; Sommerschield, T.; Shillingford, B.; Bordbar, M.; Pavlopoulos, J.; Chatzipanagiotou, M.; Androutsopoulos, I.; Prag, J.; and de Freitas, N. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900): 280–283.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European Conference on Computer Vision*, 17–33. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023. Focal Network for Image Restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13001–13011.

Das, S.; Ma, K.; Shu, Z.; Samaras, D.; and Shilkrot, R. 2019. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 131–140.

Ech-Cherif, M. E.-A.; and Cheriet, M. 2022. Frank-Wolfe-based multi-task learning for historical document restoration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 3900–3907. IEEE.

Hedjam, R.; and Cheriet, M. 2013. Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8): 2297–2312.

Hertlein, F.; and Naumann, A. 2023. Template-Guided Illumination Correction for Document Images with Imperfect Geometric Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 904–913.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, X.; Long, R.; Xue, N.; Yang, Z.; Yao, C.; and Xia, G.-S. 2022. Revisiting document image dewarping by grid regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4543–4552.

Kong, Y.; Luo, C.; Ma, W.; Zhu, Q.; Zhu, S.; Yuan, N.; and Jin, L. 2022. Look closer to supervise better: one-shot font generation via component-based discriminator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13482–13491.

Li, H.; Wu, X.; Chen, Q.; and Xiang, Q. 2023a. Foreground and Text-lines Aware Document Image Rectification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19574–19583.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023b. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18278–18289.

Li, Z.; Chen, X.; Pun, C.-M.; and Cun, X. 2023c. High-Resolution Document Shadow Removal via A Large-Scale Real-World Dataset and A Frequency-Aware Shadow Erasing Net. *arXiv preprint arXiv:2308.14221*.

Lin, Y.-H.; Chen, W.-C.; and Chuang, Y.-Y. 2020. Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12905–12914.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Ma, W.; Zhang, H.; Jin, L.; Wu, S.; Wang, J.; and Wang, Y. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 31–36. IEEE.

Nguyen, K. C.; Nguyen, C. T.; Hotta, S.; and Nakagawa, M. 2019. A character attention generative adversarial network for degraded historical document restoration. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 420–425. IEEE.

Peng, D.; Yang, Z.; Zhang, J.; Liu, C.; Shi, Y.; Ding, K.; Guo, F.; and Jin, L. 2023. UPOCR: Towards unified pixel-level ocr interface. In *Forty-first International Conference on Machine Learning*.

Raha, P.; and Chanda, B. 2019. Restoration of historical document images using convolutional neural networks. In *2019 IEEE region 10 symposium (TENSYMP)*, 56–61. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shi, Y.; Liu, C.; Peng, D.; Jian, C.; Huang, J.; and Jin, L. 2023. M5HisDoc: A Large-scale Multi-style Chinese Historical Document Analysis Benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Wadhwani, M.; Kundu, D.; Chakraborty, D.; and Chanda, B. 2021. Text extraction and restoration of old handwritten documents. *Digital Techniques for Heritage Presentation and Preservation*, 109–132.

Wang, C.; Zhou, M.; Ge, T.; Jiang, Y.; Bao, H.; and Xu, W. 2023. CF-Font: Content Fusion for Few-shot Font Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1858–1867.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Zhou, W.; Lu, Z.; and Li, H. 2022a. Udocgan: Unpaired document illumination correction with background light prior. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5074–5082.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022b. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Xue, C.; Tian, Z.; Zhan, F.; Lu, S.; and Bai, S. 2022. Fourier document restoration for robust document dewarping and recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4573–4582.

Yang, M.; and Xu, S. 2023. A novel Degraded Document Binarization model through vision transformer network. *Information Fusion*, 93: 159–173.

Yang, Z.; Liu, B.; Xxiong, Y.; Yi, L.; Wu, G.; Tang, X.; Liu, Z.; Zhou, J.; and Zhang, X. 2023a. DocDiff: Document enhancement via residual diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2795–2806.

Yang, Z.; Peng, D.; Kong, Y.; Zhang, Y.; Yao, C.; and Jin, L. 2023b. FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning. *arXiv preprint arXiv:2312.12142*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, S.; Xue, H.; Nie, N.; Zhu, C.; Liu, H.; and Fang, P. 2024. Reproducing the Past: A Dataset for Benchmarking Inscription Restoration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7714–7723.