applied sciences

MDPI

Article

UniText: A Unified Framework for Chinese Text Detection, Recognition, and Restoration in Ancient Document and Inscription Images

Lu Shen [1,2,*], Zewei Wu [1,2], Xiaoyuan Huang [1], Boliang Zhang [1], Su-Kit Tang [1], Jorge Henriques [2] and Silvia Mirri [3]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China; zewei.wu@mpu.edu.mo (Z.W.); huangxy@dgut.edu.cn (X.H.); p1807471@mpu.edu.mo (B.Z.); sktang@mpu.edu.mo (S.-K.T.)
[2] Department of Informatics Engineering, University of Coimbra, 3004-531 Coimbra, Portugal; jh@dei.uc.pt
[3] Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy; silvia.mirri@unibo.it
* Correspondence: lu.shen@mpu.edu.mo

Abstract

Processing ancient text images presents significant challenges due to severe visual degradation, missing glyph structures, and various types of noise caused by aging. These issues are particularly prominent in Chinese historical documents and stone inscriptions, where diverse writing styles, multi-angle capturing, uneven lighting, and low contrast further hinder the performance of traditional OCR techniques. In this paper, we propose a unified neural framework, UniText, for the detection, recognition, and glyph restoration of Chinese characters in images of historical documents and inscriptions. UniText operates at the character level and processes full-page inputs, making it robust to multi-scale, multi-oriented, and noise-corrupted text. The model adopts a multi-task architecture that integrates spatial localization, semantic recognition, and visual restoration through stroke-aware supervision and multi-scale feature aggregation. Experimental results on our curated dataset of ancient Chinese texts demonstrate that UniText achieves a competitive performance in detection and recognition while producing visually faithful restorations under challenging conditions. This work provides a technically scalable and generalizable framework for image-based document analysis, with potential applications in historical document processing, digital archiving, and broader tasks in text image understanding.

Keywords: ancient Chinese characters; text detection and recognition; glyph restoration

Academic Editors: Danila Germanese and Andrea Loddo

Received: 28 May 2025
Revised: 25 June 2025
Accepted: 27 June 2025
Published: 8 July 2025

Citation: Shen, L.; Wu, Z.; Huang, X.; Zhang, B.; Tang, S.-K.; Henriques, J.; Mirri, S. UniText: A Unified Framework for Chinese Text Detection, Recognition, and Restoration in Ancient Document and Inscription Images. *Appl. Sci.* **2025**, *15*, 7662. https://doi.org/10.3390/app15147662

1. Introduction

Processing ancient text images involves a number of visual and structural challenges that complicate automated analysis. Over time, physical artifacts such as paper manuscripts and stone inscriptions are subject to deterioration due to aging, environmental exposure, and human activity [1]. This results in faded ink, missing or distorted glyphs, surface erosion, and variable background textures. As illustrated in Figure 1, the texts in ancient books and stone inscriptions exhibit different writing styles, and some characters become blurred or even entirely missing due to paper deterioration or stone weathering. Image acquisition under natural conditions, commonly adopted for its non-intrusive and flexible nature, adds additional complexity through variations in lighting, viewpoint, and focus, as well as

occlusions and shadows [2,3]. These factors collectively impair the clarity and continuity of textual content, posing significant obstacles to traditional optical character recognition methods, which typically assume clean, well-aligned, and high-contrast inputs [4].

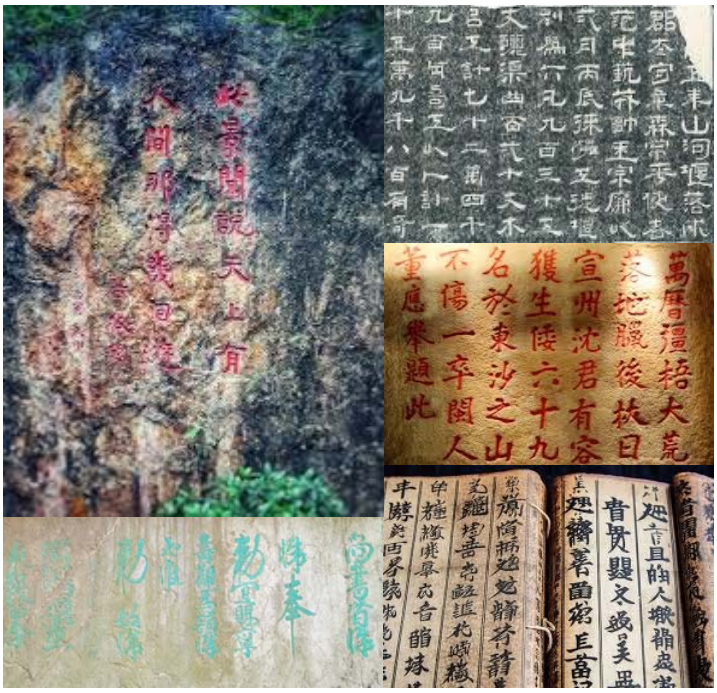


Figure 1. Sample images of Chinese ancient books and stone inscriptions collected from publicly available photos via Google Image Search. The Chinese characters shown in the images are original historical texts, which serve as the target content for handwritten text understanding.

In recent years, with the rapid development of deep learning and computer vision technologies, modern neural networks have achieved remarkable progress in image processing tasks and are increasingly applied to the automatic recognition and analysis of ancient books and stone inscriptions. Some existing studies [1,4–8] have explored the use of detection, recognition, or reconstruction techniques in the processing of historical documents. However, these methods typically focus on specific tasks and are often trained on cropped and well-aligned line-level text or character images. As a result, they struggle to handle the combined challenges found in full-page images, which often include various types of noise and interference. A few approaches attempt to process entire pages directly, yet they often assume that ancient texts exhibit regular structures and standardized layouts. These methods usually rely on rectangular bounding boxes for character detection, which limits their ability to handle more complex real-world scenarios such as changes in camera angle, character rotation, or skew, and flexible text arrangement.

To address the limitations of existing methods, such as task specificity, reliance on cropped line or character inputs, and restricted detection box shapes, we propose a unified multi-task learning framework for character detection, recognition, and glyph restoration in page-level images of historical books and stone inscriptions. This framework enables the joint modeling and processing of entire pages under complex natural conditions, and is particularly suited for challenging scenarios involving varying camera angles, uneven lighting, and irregular text arrangements.

The proposed architecture is built upon a shared backbone network, with an interactive multi-scale decoding module and task-specific output branches. The glyph restoration branch outputs a segmentation map of character structures to recover the visual appearance of individual glyphs. Meanwhile, the detection and recognition branches jointly perform

geometric localization and category prediction of characters through convolutional and fully connected layers. By introducing a mechanism for feature sharing across tasks and designing specialized decoders for each task, the framework effectively integrates both visual and semantic information, enabling the efficient joint modeling of character structure and meaning. In particular, to improve the learning efficiency of the recognition branch, we propose a training strategy that combines fine-grained instance-level mask supervision with target-aware classification loss. During training, the model produces pixel-level category predictions across the entire image, but classification loss is computed only within valid labeled regions. This reduces the influence of background noise on recognition and enhances the model's robustness in complex environments.

Overall, the main contributions of our work are as follows:

- We propose an end-to-end multi-task learning framework for character detection, recognition, and glyph reconstruction in page-level images of Chinese historical documents and inscriptions.
- We introduce a fine-grained instance mask-based supervised signal for text regions, and propose multi-task loss guided by text mask, char category, and context position. This strategy is designed to suppress misleading gradients from non-text areas and improve the model's robustness to background noise.
- The proposed approach is validated on the test set of historical documents and stone inscriptions, showing its effectiveness in improving character readability and supporting digital preservation.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we detail the proposed method. Section 4 describes the dataset and evaluation metrics for the three tasks. Section 5 analyzes the experimental results. Finally, Section 6 concludes the paper.

2. Related Work

In the field of historical document digitization, several studies on non-Chinese handwritten texts have provided valuable insights. For example, in the study of ancient Hebrew ostraca, Sober et al. [9] propose a variational approach for stroke restoration, modeling each character as a trajectory of a pen with a variable radius. They optimize a cubic spline representation via gradient descent, guided by manually sampled anchor points, to structurally reconstruct over one thousand characters dating back to the 8th–7th century BCE. TexRGAN [10], on the other hand, introduces an end-to-end restoration framework based on Generative Adversarial Networks (GANs), which is the first to simultaneously handle struck-out and underlined handwritten text within a single deep model. By incorporating both spatial and structural loss functions and adopting a weakly supervised training scheme, the model significantly improves robustness and OCR accuracy under various types of textual distortions.

To meet the digitization needs of medieval Latin dictionaries, Koch et al. [11] design a complete pipeline for handwritten text recognition, covering lemma localization, extraction, and transcription. Their approach integrates image segmentation for preprocessing with a Transformer-based architecture composed of vision encoders and a GPT-2 decoder and applies extensive data augmentation techniques to enhance generalization. In the context of damaged Latin inscriptions, Locaputo et al. [12] propose several neural network-based methods for restoring missing text, including Transformer models based on the Ithaca architecture, masked language models, diffusion-based models, and blank language models, forming a cross-disciplinary digital restoration framework. Additionally, facing the challenges of diverse scripts and complex layouts in 12th–15th century Latin and French manuscripts, Aguilar et al. [13] employ CRNN-CTC architectures for

handwritten text recognition and explore cross-model fusion strategies to further improve recognition performance.

Although these methods provide effective solutions for Western scripts, their direct applicability to Chinese characters remains limited due to the significantly higher visual complexity, dense stroke patterns, and structural variability of Chinese writing. In addition, inscriptions in stone steles and ancient books often present challenges related to non-linear degradation. For ancient Chinese books and stone inscriptions, existing research primarily focuses on single-task or dual-task settings, such as text detection, character recognition, or glyph restoration. One representative work, proposed by Zhang et al. [5], addresses glyph restoration from stone inscription images using pre-cropped single-character inputs. To tackle the high self-similarity between foreground characters and background textures, they introduce a character image segmentation framework that combines stacked U-Nets [14] with generative adversarial networks (GANs), aiming to enhance character extraction and restoration accuracy.

The research in [4] tackles visual corruption in inscription rubbings caused by weathering and erosion. This work presents a multimodal benchmark framework for ancient text restoration, including the construction of the Chinese Inscription Rubbing Image (CIRI) dataset and a diffusion-based restoration pipeline. The proposed pipeline integrates a global restoration module for structural recovery, a local refinement module for ambiguous characters, and a text rectification module for recognition, clustering, rendering, and semantic correction.

Shimoyama et al. [8] focus on character detection in Han dynasty bamboo slip images, which present challenges such as dense character layouts, scale variations, and minimal spacing. Their model, based on a U-Net [14] architecture, learns both character regions and boundary information to improve detection and separation accuracy. For more complex visual environments, AncientGlyphNet [1] introduces a deep learning framework for ancient character detection across diverse media including stone inscriptions, calligraphy, and couplets. The framework incorporates the ACHaar downsampling module to retain key glyph features, the Glyph Focus Module (GFM) to enhance structural modeling, and the Character Contour Refinement Layer (CCRL) to sharpen character boundaries. The accompanying HUSAM-SinoCDCS dataset supports robust evaluation under real-world conditions.

To address the variety and scarcity of ancient character data, Xu et al. [6] propose a large-scale incremental learning framework based on Convolutional Prototype Networks (CPNs). This framework enables the continual learning of new character categories without storing data from previously seen classes, making it suitable for expanding character sets in low-resource scenarios. Unlike studies that target a single task, the work in [7] presents an end-to-end joint model capable of performing character detection, recognition, and document layout analysis simultaneously. This multi-task approach offers a unified solution for a comprehensive understanding of historical document images.

However, many existing methods assume horizontally aligned text and rely on standard rectangular bounding boxes, which limits their performance when applied to text images with multi-angle perspectives, complex background noise, and varying lighting conditions. Furthermore, considering the rich diversity of glyph forms and their cultural significance in ancient books and stone inscriptions, it is essential to develop an end-to-end framework for page-level character detection, recognition, and glyph restoration that can robustly handle text under such challenging conditions.

3. Methodology

3.1. The Proposed Model

We propose an end-to-end multi-task learning model integrating glyph restoration, text detection, and recognition into a unified framework. The model is designed to improve the robustness and overall accuracy of text extraction from historical book images and stone inscription images, particularly in challenging scenarios with low image quality, complex background noise, and unclear character structures. By leveraging multi-task learning on shared features, the model effectively captures both the structural characteristics of characters and contextual semantics, which enhances the stability and generalization of text extraction. As shown in Figure 2, unlike traditional multi-stage approaches, our method employs a single shared backbone network for hierarchical feature extraction from the input image. Task-specific branches are then used to perform image-level glyph denoising and region-level text detection and recognition. This architecture simplifies the overall pipeline, reduces deployment and maintenance complexity, and improves cross-task learning efficiency through feature sharing. The network structure and parameters of each module in Figure 2 correspond to those summarized in Table 1.

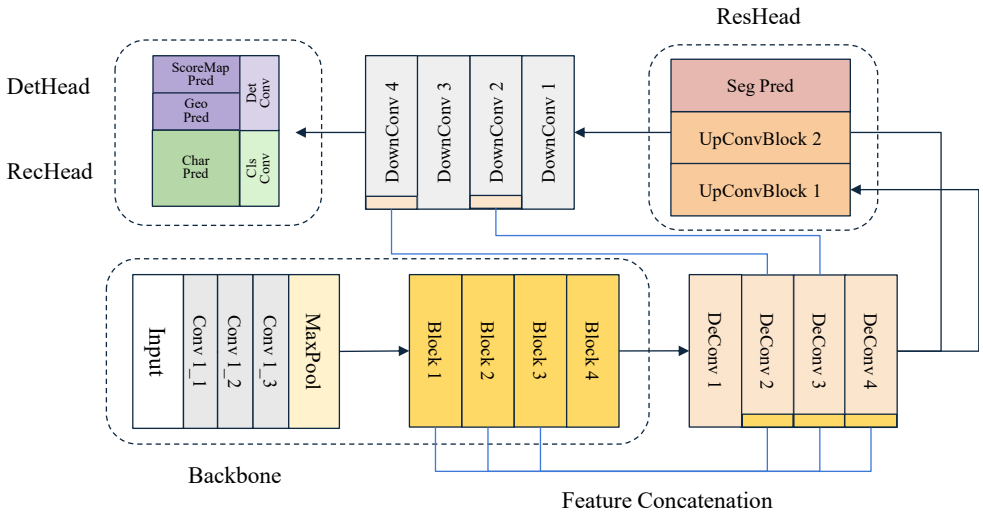


Figure 2. Architecture of the proposed model.

Specifically, the model consists of three main components: a shared feature extraction network, a glyph restoration branch, and a text detection and recognition branch. The shared backbone network is based on a ResNet-like multi-layer residual structure, which extracts bottom-up multi-scale hierarchical features. These features serve as common inputs to the task-specific branches, providing rich and stable representations. The glyph restoration branch combines shallow visual details with deep semantic features. It performs image-level reconstruction and denoising of character regions through upsampling and feature fusion, generating a full-image character segmentation map. This output enhances the readability of degraded text and supports text extraction. The text detection and recognition branch focuses on extracting high-level semantic information related to textual content from the shared features. It integrates both detection and recognition sub-modules to predict the spatial locations and class labels of characters. By sharing backbone features and jointly optimizing the three tasks during training, the model improves text extraction performance in complex scenes while maintaining overall computational efficiency.

Table 1. Detailed parameters of the proposed network.

Level	Layer	Params	Output
Input	-	-	512×512
Conv_1_1	ConvBNLayer	3×3, in = 3, stride = 2	256×256
Conv_1_2	ConvBNLayer	3×3, in = 32, stride = 1	256×256
Conv_1_3	ConvBNLayer	3×3, in = 32, stride = 1	256×256
MaxPool1	MaxPool2d	3×3, stride = 2, padding = 1	128×128
Block1	ResBlock	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	128×128
Block2	ResBlock	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	64×64
Block3	ResBlock	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	32×32
Block4	ResBlock	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	16×16
DeConv1	DeConvBNLayer	4×4, in = 512, stride = 2	32×32
DeConv2	ConvBNLayer	3×3, in = 384, stride = 1	32×32
	DeConvBNLayer	4×4, in = 128, stride = 2	64×64
DeConv3	ConvBNLayer	3×3, in = 256, stride = 1	64×64
	DeConvBNLayer	4×4, in = 128, stride = 2	128×128
DeConv4	ConvBNLayer	3×3, in = 192, stride = 1	128×128
	ConvBNLayer	3×3, in = 128, stride = 1	128×128
UpConvBlock1	Conv2d	1×1, in = 128, stride = 1	128×128
	Upsample	size $= (256, 256)$	256×256
	Conv2d	3×3, in = 32, stride = 1	256×256
UpConvBlock2	Upsample	size $= (512, 512)$	512×512
	Conv2d	3×3, in = 32, stride = 1	512×512
DownConv1	Conv2d	3×3, in = 128, stride = 2	64×64
DownConv2	Conv2d	1×1, in = 384, stride = 1	64×64
DownConv3	Conv2d	3×3, in = 256, stride = 2	32×32
DownConv4	Conv2d	1×1, in = 640, stride = 1	32×32
DetConv	ConvBNLayer	3×3, in = 512, stride = 1	32×32
	ConvBNLayer	3×3, in = 128, stride = 1	32×32
ClsConv	ConvBNLayer	3×3, in = 512, stride = 1	32×32
	ConvBNLayer	3×3, in = 512, stride = 1	32×32
ResHead	Conv2d	1×1, in = 128, stride = 1	512×512
DetHead	ConvBNLayer	1×1, in = 64, stride = 1	32×32
	ConvBNLayer	1×1, in = 64, stride = 1	32×32
RecHead	ConvBNLayer	1×1, in = 512, stride = 1	32×32
	fc	in = 1024	out = 745
	softmax	in = 745	32×32

Among them, the glyph restoration branch adopts a typical encoder–decoder architecture. The encoder is implemented by the shared backbone network, which extracts hierarchical visual features from the input image. The decoder progressively upsamples the feature maps to restore the original spatial resolution. During the encoding stage,

the input image $x \in {}^{3\times H\times W}$ is passed through multiple residual blocks in the backbone to obtain intermediate representations $\mathcal{F}_{\text{backbone}}$, which capture both the local textures and global structures of the characters. These features contain essential visual cues such as shape, contour, and texture, which support the subsequent restoration task. In the decoding stage, the branch performs a series of upsampling and convolution operations to gradually reconstruct the spatial resolution of $\mathcal{F}_{\text{backbone}}$. The decoding process consists of four stages of transposed convolutions (stride = 2), each followed by feature fusion via concatenation with corresponding encoder layers. This hierarchical design enables progressive spatial resolution recovery from 16×16 to 512×512.

To preserve the structural integrity of characters during reconstruction, we introduce a multi-scale feature fusion mechanism. At each upsampling stage, the decoder combines the current feature maps with intermediate features from the corresponding resolution levels of the backbone, integrating high-level semantics with low-level details. The fused features are then refined through a standard sequence of convolution, normalization, and activation operations, specifically using a 3×3 convolution layer, batch normalization, and ReLU activation. This combination, commonly adopted in dense prediction architectures such as U-Net [14] and FPN [15], has been shown to enhance the spatial coherence of feature representations and suppress artifacts introduced during feature fusion. The final output is a prediction map that matches the size of the original image. The overall restoration process can be summarized as the following function:

$$\mathcal{F}_{\text{restored}} = f_{\text{restore}}(\mathcal{F}_{\text{backbone}}; \theta_d) \tag{1}$$

where $f_{\text{restore}}(\cdot)$ denotes the sequence of upsampling, fusion, and convolutional operations in the restoration module, and θ_d represents the learnable parameters. The final output $\mathcal{F}_{\text{restored}}$ is a character-level segmentation map that indicates the spatial distribution of character foreground regions in the input image.

Meanwhile, the upsampled features are further processed through a downsampling path, where features at different scales are fused and passed to the text detection and recognition branch for subsequent processing. In the text detection module, we adopt a polygon-based regression approach to achieve the flexible and accurate localization of text regions. The core idea of the detector is to generate multiple feature maps from polygon annotations, which are used for both text region confidence prediction and geometric parameter regression. These outputs jointly provide information about the position and shape of the text.

The objective of text region segmentation is to determine which pixels in the image belong to text. We represent the probability of text presence using a binarized region mask, where a foreground probability map indicates the confidence score for each pixel that is part of a text region. The segmentation module applies convolutional layers to the fused features and outputs a probability map of size $H \times W$, representing the spatial distribution of text likelihood across the image. For the geometric regression part, to better model text regions with inclined angles or irregular shapes, we introduce a quadrilateral-based regression strategy. This strategy predicts the relative offsets between each pixel and the four corner points of the corresponding text region, enabling the flexible modeling of text boundaries. Specifically, for each pixel (x, y), we compute the offset (d_{xi}, d_{yi}) to each of the four vertices $P_i = (x_i, y_i)$, and encode this information into the geometric feature map $\mathcal{G}(x, y)$.

$$P_i = \left(x + d_{xi},\ y + d_{yi}\right), \quad i = 1, 2, 3, 4 \tag{2}$$

In practice, the detection branch first applies two convolutional layers to extract intermediate features from the fused representation. These features are then passed through

two parallel 1×1 convolutions to generate a score map and a geometry map. The score map indicates the probability of each pixel belonging to a text region and is normalized to the range $[0, 1]$ using a sigmoid function. The geometry map encodes the distances from each pixel to the four corner points of the predicted quadrilateral, with 8 output channels corresponding to the (dx, dy) offsets of the four vertices. This allows the coordinates of the quadrilateral vertices to be reconstructed from pixel-wise offsets and enables the model to effectively capture the geometric structure of text regions with various orientations and shapes. It is worth noting that, to prevent excessively large offset values from causing training instability, the geometry outputs should be normalized, typically by applying a sigmoid activation followed by scaling. Additionally, to ensure consistent vertex ordering, a fixed-point sequence (e.g., starting from the top-left corner in a clockwise direction) should be used during both label generation and model prediction.

In the text recognition task, the fused visual features are further processed by a classification branch. Unlike the traditional Convolutional Recurrent Neural Network (CRNN) [16], we adopt a lightweight fully convolutional structure as the classification head to ensure overall efficiency and enable end-to-end training. In this branch, the input features are first passed through two convolutional layers to extract semantic information, followed by a 1×1 convolution that expands the channel dimension, resulting in the recognition feature map $\mathbf{f}_{\text{cls}}$ with a shape of $B \times C \times H \times W$, where B denotes the batch size, C is the number of channels, and H and W are the spatial dimensions. To perform character classification, $\mathbf{f}_{\text{cls}}$ is flattened along the spatial dimensions and fed into a fully connected network with dropout, producing a set of class probabilities for each spatial location. The flattened feature has a shape of $B \cdot H \cdot W \times N$, where N is the number of character categories. The softmax function is applied along the class dimension to normalize the output into a valid probability distribution for each pixel. Finally, the output is reshaped back to $B \times N \times H \times W$, restoring the spatial layout of the predictions. As a result, each pixel position carries a complete probability distribution over all character classes, enabling dense, pixel-level character recognition.

$$\mathbf{f}_{\text{cls}} = \text{Softmax}\Big(\text{FC}\Big(\text{Conv}_{1\times1}\Big(\text{Conv}^{(2)}_{3\times3}(\mathbf{f}_{\text{det-rec}})\Big)\Big)\Big) \tag{3}$$

3.2. Fine-Grained Text Region Annotation Strategy

In object detection and classification tasks, a target region annotation map is used to indicate which areas of an image contain the objects of interest (e.g., text). Traditional methods typically construct such maps based on bounding boxes. Specifically, for each pixel in the image, a value of 1 is assigned if the pixel lies within the annotated region, indicating that it belongs to the target area. Otherwise, a value of 0 is assigned, indicating background. The target region annotation map $M(x, y)$ is defined as follows:

$$M(x, y) = \begin{cases} 1, & \text{if } (x, y) \in \mathcal{R}_{\text{bbox}} \\ 0, & \text{otherwise} \end{cases}$$

where (x, y) denotes the pixel position and $\mathcal{R}_{\text{bbox}}$ represents the annotated region provided by bounding boxes.

Although this strategy is commonly used in general object detection and recognition tasks, it presents certain limitations in text-related scenarios. First, text region annotations are typically based on rectangular bounding boxes, which not only cover the actual character strokes but also inevitably include surrounding background areas, such as inter-character spacing, noise around the text, image textures, and even printing or carving artifacts. These non-text pixels, although labeled as part of the target region, do not carry

meaningful semantic information and may interfere with the model's ability to accurately learn the structure of the text during training.

Second, in page-level images, text usually occupies only a small portion of the image. This is particularly true in historical documents or stone inscription images, where text regions coexist with large and complex backgrounds, and the characters are dense and structurally detailed. In such cases, directly using the entire annotated region as the training supervision signal may reduce the model's focus on critical stroke features and lead to the learning of background patterns, thereby affecting the recognition accuracy.

Furthermore, characters exhibit highly structured spatial patterns, such as the relative positions among horizontal, vertical, and diagonal strokes. However, traditional coarse-grained region annotations are insufficient for capturing such fine-grained details and only provide a broad, weak supervisory signal. This overly relaxed supervision may cause the model to learn redundant or irrelevant information, making it difficult to focus on the true boundaries and structural features of characters, which in turn limits its generalization ability and robustness in complex scenarios.

Figure 3 illustrates this issue. The green rectangle represents the bounding box used to localize a textual region. A common approach to converting this box into a supervision map is to fill the entire area with a uniform foreground label, resulting in what we refer to as a box-based mask annotation (indicated by the gray-green shaded area). This method provides weak supervision, as it assumes all pixels within the bounding box contribute equally to the semantic content of the region. In reality, only a subset of these pixels corresponds to actual character strokes.

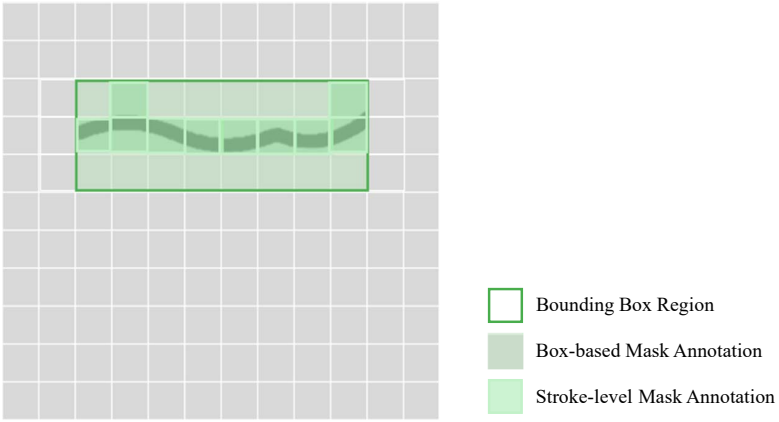


Figure 3. Comparison of annotation strategies. The green rectangle denotes the bounding box. The box-based mask annotation (gray-green area) labels all pixels within the bounding box as foreground. In contrast, the stroke-level mask annotation (light green area) includes only the pixels corresponding to actual character strokes, providing more precise supervision.

To address the mentioned limitations of traditional target region annotation methods in text-based scenarios, we propose a fine-grained instance mask-based annotation strategy for the classification task. The goal is to improve the model's learning efficiency and recognition accuracy in the presence of complex backgrounds. This method introduces structure-aware supervision at the glyph level by leveraging character-level segmentation labels, also referred to as glyph masks. In the supervision signal, only the pixels that correspond to character strokes are preserved. Specifically, we construct a binary mask where pixels belonging to the strokes are assigned a value of 1, and all other pixels, such as those in inter-character spaces and background regions, are assigned a value of 0. This produces a more discriminative and semantically precise target region annotation map. The target region annotation map is defined as follows:

$$M(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \mathcal{R}_{\text{stroke}} \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{S}_{\text{stroke}}$ denotes the set of pixels corresponding to character strokes as defined by the segmentation labels, and (x,y) represents a pixel location in the image.

The contrast between the two approaches is visually summarized in Figure 3. The box-based annotation treats the entire bounding box as foreground, while the stroke-level annotation selectively includes only the actual stroke pixels. This difference significantly impacts the supervision signal used during training. The stroke-level mask offers a more semantically aligned and spatially precise indication of target regions, allowing the model to focus directly on the most informative parts of the image. From a technical perspective, this shift from coarse to fine-grained annotation also affects the distribution of training samples. In box-based masks, the foreground occupies a larger area, potentially leading to class imbalance when the foreground is overrepresented by semantically irrelevant pixels. In contrast, stroke-level masks reduce this imbalance by restricting positive labels to meaningful regions only. This enhances the signal-to-noise ratio in the supervision and prevents the model from being biased toward background characteristics.

In addition, stroke-level supervision is inherently more suitable for character-level classification and recognition tasks. When used in conjunction with fully convolutional classification heads, these masks enable the model to associate local visual features with specific character categories on a per-pixel basis. This facilitates the learning of highly localized representations, which are critical for distinguishing between nuanced glyph structures, especially in low-quality or degraded inputs.

3.3. Joint Optimization Framework

3.3.1. Loss Components and Formulation

To simultaneously optimize text detection, character recognition, and glyph restoration, we design a multi-task joint loss function that integrates region localization, category prediction, and structural reconstruction into a unified optimization framework. The joint loss consists of three components, corresponding to text detection loss, character classification loss, and glyph restoration loss. By jointly optimizing these objectives during training, the model aims to accurately localize text regions, recognize character categories, and recover detailed glyph structures under complex backgrounds. The overall loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{res}} \tag{4}$$

The text detection loss, denoted as $\mathcal{L}_{\text{det}}$, includes two components. The first is a region segmentation loss, where Dice Loss [17] measures the overlap between the predicted text areas and the corresponding ground truth regions. The second is a geometric regression loss, where Smooth L1 Loss [18] is used to optimize the geometric parameters of text boxes, such as the distances to the top, bottom, left, and right boundaries, as well as the rotation angles. This combination is designed to improve both the segmentation quality and the precision of geometric regression.

$\mathcal{L}_{\text{cls}}$ is the character classification loss, which enhances the model's ability to distinguish character categories at the pixel level. We adopt a cross-entropy-based strategy that selects character stroke regions as the supervision area, focusing on the foreground to suppress background interference and improve classification performance. The detailed formulation of this loss is provided in Section 3.3.2.

The glyph restoration loss, denoted as $\mathcal{L}_{\text{res}}$, aims to recover character details from degraded or noisy input images, thereby improving the quality of character reconstruction. This loss consists of three components. The first uses mean squared error (MSE) [19] to

perform pixel-wise regression between the restored output and the target image, enhancing the accuracy of grayscale reconstruction. The second introduces binary cross-entropy with logits (BCEWithLogitsLoss) [19] to guide the segmentation of foreground regions and ensure the structural completeness of the characters. The third employs the soft-clDice [20] loss to preserve the topological structure of character skeletons, which helps maintain the connectivity of elongated strokes and improves the recovery of fine-grained glyph details.

3.3.2. The Stroke-Aware Classification Loss Computation Strategy

In the text recognition task, the model outputs a category distribution for each pixel. However, in full-page images, text regions typically occupy only a small proportion of the total pixels, while the majority of pixels belong to the background. Directly computing the cross-entropy loss over all pixels introduces a severe class imbalance, as the background dominates the loss. This imbalance leads to misleading gradients during training and weakens the model's ability to discriminate fine-grained features within text regions. To mitigate this issue, we propose a classification loss computation strategy based on a stroke region mask. During training, the loss is computed only within character stroke regions, thereby reducing the influence of background pixels and improving the model's capability to distinguish character categories.

Specifically, let the model's pixel-wise classification prediction be denoted as $P_{\text{cls}} \in \mathbb{R}^{B\times C\times H\times W}$. The corresponding ground-truth label map is denoted as $G_{\text{cls}} \in \mathbb{Z}^{B\times H\times W}$, and the stroke region mask is denoted as $M_{\text{stroke}} \in \{0,1\}^{B\times H\times W}$, where a value of 1 indicates that a pixel belongs to a character stroke. We define a binary mask $M'(x,y)$ as follows:

$$M'(x,y) = \begin{cases} 1, & \text{if } M_{\text{stroke}}(x,y) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Using this mask, we select only the pixels within stroke regions and compute the classification loss as follows:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|M'|} \sum_{(x,y)\in M'} \text{CrossEntropy}(P_{\text{cls}}(x,y), G_{\text{cls}}(x,y)) \tag{6}$$

where $|M'|$ denotes the total number of foreground pixels with $M'(x,y) = 1$. This formulation ensures that the loss computation is constrained to semantically relevant regions, thereby reducing the risk of overfitting to background noise and enabling the model to better capture the visual patterns associated with individual character classes. By applying this targeted supervision mechanism, the model is encouraged to focus its learning capacity on the spatial locations that carry meaningful semantic information.

4. Dataset and Evaluation Protocol

4.1. Dataset Constructed with Poisson Blending and Relief Simulation

Existing datasets [1,4,7,21–23] related to traditional Chinese characters often lack comprehensive annotations at the character level for full-page texts, including detection, recognition, and glyph-level labeling. In addition, they rarely address practical challenges such as noise interference, multi-angle imaging, and complex backgrounds, which limit their effectiveness for training high-quality models.

To address these issues, we consider the complexity of ancient books and stone inscriptions in real-world scenarios and synthesize page-level images of ancient texts with diverse fonts and backgrounds that simulate aged paper textures and stone surfaces. This approach reduces manual annotation costs and facilitates the efficient creation of a large and diverse set of character-level annotations. Each image includes quadrilateral bounding

boxes for character detection, category labels, and glyph segmentation masks, supporting multi-task model training and evaluation. Specifically, we construct a synthetic dataset consisting of 14,000 images and 744 commonly used ancient character categories. Detailed statistics of the dataset are shown in Table 2.

Table 2. Statistics of the constructed dataset based on ancient manuscripts and stone inscriptions. The dataset is designed for three tasks: text detection (Det), recognition (Rec), and glyph restoration (Gly).

Dataset	Chars	Classes	Images	Noise	Angles	Det	Rec	Gly
Train	430,552	744	12,000	Yes	Yes	✓	✓	✓
Test	71,757	744	2000	Yes	Yes	✓	✓	✓

Figure 4 presents several examples of text images from ancient books and stone inscriptions in our dataset. The dataset includes 14 commonly used font styles, such as cursive, clerical, and running scripts. The text appears with varying degrees of slant and is combined with diverse backgrounds to simulate the visual characteristics found in real-world manuscripts and inscriptions. The background images are collected from a wide range of sources. For ancient books, the backgrounds are derived from high-resolution scans of historical documents, featuring yellowed paper textures, natural creases, and stain marks. In contrast, the stone inscription backgrounds are taken from photographs of actual steles, cliff carvings, and epitaphs, which exhibit irregular stone textures, weathering cracks, and lighting variations. To enhance the realism of the synthetic images and achieve natural integration between text and background, we apply a series of image processing techniques.

Figure 4. Sample text images of ancient books and stone inscriptions from the constructed dataset. The Chinese characters are retained as they represent the primary visual content processed by our models.

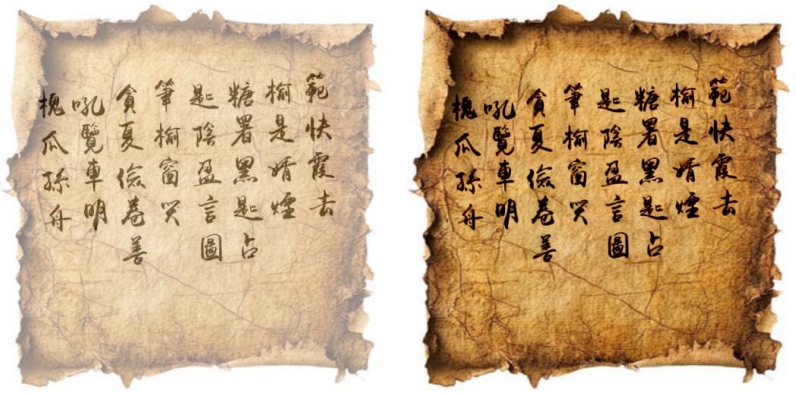
To improve the visual quality of synthetic text–background images, we adopt a gradient domain-based method, known as Poisson Image Blending [24], as an alternative to the commonly used alpha blending approach. This method formulates image fusion as an optimization problem in the gradient field, allowing foreground content to be integrated seamlessly into the background in terms of lighting, texture, and tone. Alpha blending combines the foreground image $S(x, y)$ and the background image $B(x, y)$ by linear interpolation. The resulting image $I(x, y)$ is computed as follows:

$$I(x, y) = \alpha \cdot S(x, y) + (1 - \alpha) \cdot B(x, y), \quad \alpha \in [0, 1] \tag{7}$$

Although simple and efficient, this method often leads to visible seams when the foreground and background differ in texture, brightness, or material. As shown in Figure 5, the left image produced by alpha blending exhibits semi-transparent text and a washed-out background. The result lacks realism, and the text appears visually detached from the page. To address these limitations, we apply Poisson image blending [24], which is formulated as a gradient-preserving optimization problem. The goal is to retain the gradient field of the foreground image within a selected region Ω, while ensuring a smooth transition at the boundary $\partial\Omega$ with the background. The problem is defined as follows:

$$\min_I \int_\Omega \|\nabla I - \nabla S\|^2 \quad \text{subject to} \quad I|_{\partial\Omega} = B|_{\partial\Omega} \tag{8}$$

Here, ∇ denotes the gradient operator, S is the source (foreground) image, B is the background image, and I is the output image. This formulation ensures that the gradients inside the pasted region match those of the source, while boundary values are constrained by the background. As a result, the output exhibits consistent illumination and texture continuity across the fusion boundary. As illustrated in Figure 5, the right image produced by Poisson blending demonstrates a more natural integration. The text appears to be part of the background surface, with preserved texture and no visible seams. Compared to alpha blending, Poisson blending yields improved edge smoothness, texture consistency, and overall visual coherence.



(a) (b)

Figure 5. Comparison between (**a**) alpha blending with transparency $\alpha = 0.5$ and (**b**) Poisson image blending. The Chinese text is used as a representative example for evaluating the visual effectiveness of image blending techniques. No translation is provided, as the semantic content is not essential to the analysis.

To generate realistic and visually diverse stone inscription text images, we propose a synthesis method based on emboss filtering and alpha blending. Given a grayscale text image as input, the method simulates the visual characteristics of carved stone text, such as depth variation, edge shadows, material reflectivity, and surface texture. We first apply a directional emboss filter to the input text image using a convolution operation:

$$I_{\text{emboss}} = S * K_{\text{emboss}} \tag{9}$$

where S is the input grayscale text image, and $*$ denotes the 2D convolution operator. The emboss kernel K_{emboss} is defined as follows:

$$K_{\text{emboss}} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix} \tag{10}$$

This operator enhances gradient edges in specific directions. It generates a pseudo-3D effect that mimics the appearance of chiseled text on stone surfaces. The embossed text image is then embedded into a resized stone background image. The placement is centered to ensure spatial alignment. We apply linear alpha blending to combine the text and background images:

$$I(x,y) = (1-\alpha) \cdot B(x,y) + \alpha \cdot I_{\text{emboss}}(x,y) \tag{11}$$

where $B(x,y)$ is the background image and $\alpha \in [0,1]$ controls the visibility of the text. The embedding position (x_0, y_0) is computed as follows:

$$x_0 = \left\lfloor \frac{W-w}{2} \right\rfloor, \quad y_0 = \left\lfloor \frac{H-h}{2} \right\rfloor \tag{12}$$

where (W,H) and (w,h) denote the dimensions of the background and text images, respectively. This blending preserves the details of the stone texture while maintaining the visual coherence between text and background.

To further enhance realism and variability, we introduce random perturbations in the HSV color space. This simulates natural variations in stone surfaces under different lighting and aging conditions. Specifically, the hue, saturation, and value channels are perturbed as follows:

$$\begin{aligned} H' &= H + \delta_h, \quad \delta_h \sim \mathcal{U}(-\Delta_h, +\Delta_h) \\ S' &= S + \delta_s, \quad \delta_s \sim \mathcal{U}(-\Delta_s, +\Delta_s) \\ V' &= V + \delta_v, \quad \delta_v \sim \mathcal{U}(-\Delta_v, +\Delta_v) \end{aligned} \tag{13}$$

where the variables δ_h, δ_s, and δ_v are drawn from uniform distributions, and the parameters Δ_h, Δ_s, and Δ_v denote the predefined perturbation ranges. This operation enhances the diversity of visual styles, mimicking stone materials such as bluish-gray slate, reddish sandstone, and weathered limestone.

As shown in Figure 6, the synthesized result in (c) demonstrates improved visual integration in terms of lighting consistency, surface realism, and perceptual depth. In contrast, the baseline results in (a) and (b) provide only basic compositing without capturing the physical characteristics of stone inscriptions. The proposed method effectively models the visual properties of carved text by combining directional filtering, transparency-based blending, and color perturbation. It produces high-quality samples that are useful for training robust models.

(a) (b) (c)

Figure 6. Comparison of stone-style text synthesis results using (**a**) alpha blending with transparency $\alpha = 0.5$, (**b**) Poisson image blending, and (**c**) the proposed method based on relief simulation. The Chinese text is used as visual content to evaluate synthesis quality and does not require translation.

4.2. Harmonic Evaluation Metric

To comprehensively evaluate the performance of the proposed model on the three subtasks, we design a set of task-specific metrics as well as joint evaluation indicators. For the text detection task, we adopt Hmean [23] as the primary evaluation metric to assess both the accuracy and completeness of text region localization. Hmean is defined as the harmonic mean of detection precision and recall, as given by the following formula:

$$\text{Hmean} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Precision indicates the proportion of predicted bounding boxes that are successfully matched to ground-truth boxes, while recall measures the proportion of ground-truth boxes that are correctly detected by the model. During evaluation, we determine matches by computing the Intersection over Union (IoU) between predicted and ground-truth boxes, using a fixed threshold of 0.5. A detection is considered correct only if the IoU exceeds this threshold and the match is unique. Precision reflects the model's ability to avoid false positives, whereas recall captures its capacity to cover all relevant text regions without omissions. Since these two metrics may conflict, Hmean serves as a balanced measure that considers both, providing a more comprehensive assessment of the overall performance of the text detection module.

In the character recognition task, we use Character Accuracy (CA) as the evaluation metric to measure the model's performance at the character classification level. This metric is defined as the ratio of correctly recognized characters to the total number of characters. N_{correct} denotes the number of characters whose predicted labels exactly match the ground truth, and N_{total} represents the total number of characters to be recognized. The evaluation is conducted on a per-character basis, which directly reflects the classification accuracy of the character recognition module.

$$\text{CA} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{15}$$

In the glyph restoration task, we adopt mean Intersection over Union (mIoU) and Root Mean Square Error (RMSE) as joint evaluation metrics to assess the differences between the predicted and ground-truth images from the perspectives of structural similarity and pixel-level accuracy. mIoU measures the overlap between the predicted and reference regions and reflects the accuracy of structural reconstruction. RMSE quantifies the overall pixel-wise error between the two images, where a lower value indicates a closer match to the ground truth. To provide a more comprehensive evaluation of restoration quality, we introduce a combined score termed GlyphScore. This metric averages the normalized values of mIoU and RMSE, balancing both structural fidelity and fine-grained detail preservation. A higher GlyphScore indicates better performance in both restoring the underlying glyph structure and reconstructing pixel-level details.

We introduce two joint evaluation metrics to assess the synergy between tasks to comprehensively evaluate the overall performance of the model under a multi-task learning framework. Hmean-CA is designed to measure the combined performance of the model on character detection and recognition. This metric is computed as the harmonic mean of Hmean and CA. Hmean-CA considers both the localization accuracy of the detection module and the classification accuracy of the recognition module, thus reflecting the model's ability to coordinate text extraction with semantic understanding. As the harmonic mean is sensitive to lower values, this metric penalizes poor performance in either subtask, encouraging the model to maintain a balanced and stable performance across the two tasks. The definition is as follows:

$$\text{Hmean-CA} = \frac{2 \cdot \text{Hmean} \cdot \text{CA}}{\text{Hmean} + \text{CA}} \tag{16}$$

HCG serves as a joint metric for all three tasks, evaluating the overall performance of character detection, recognition, and glyph restoration simultaneously. It is computed as the harmonic mean of three individual task metrics: Hmean, CA, and GlyphScore. This metric reflects the model's ability to localize text regions, recognize character semantics, and reconstruct glyph-level visual details. By integrating these aspects, HCG provides a comprehensive measure of the model's effectiveness and stability across the full text understanding pipeline. It enables the effective assessment of how multi-task learning contributes to performance improvements across tasks and offers a unified criterion for model comparison and optimization. The definition is as follows:

$$\text{HCG} = \frac{3 \cdot \text{Hmean} \cdot \text{CA} \cdot \text{GlyphScore}}{\text{Hmean} \cdot \text{CA} + \text{Hmean} \cdot \text{GlyphScore} + \text{CA} \cdot \text{GlyphScore}} \tag{17}$$

5. Experiments

5.1. Implementation Details

We implement our model with PyTorch 2.1.0 and evaluate it alongside methods from the PaddleOCR library [25]. All experiments were run on Ubuntu 20.04.6 LTS with an Intel E5-2680 v3 CPU and an RTX 3090 GPU.

5.2. Ablation Study

To validate the effectiveness of our fine-grained annotation and stroke-aware loss strategies, we compare the proposed model UniText with two alternative approaches: one that uses bbox masks for classification annotation and loss computation, and another that computes classification loss over all output pixels. The comparison is conducted on the dataset described in Section 4.1 in terms of text detection, recognition, and glyph restoration performance (corresponding to Exp. 1, Exp. 2, and Exp. 3, respectively).

Table 3 shows that UniText outperforms the other two approaches in detection metrics, including precision, recall, and Hmean. In terms of recognition, it achieves a character accuracy of 88.05%, which is significantly higher than that of Exp. 2 (80.37%) and Exp. 3 (54.05%). These results indicate that restricting classification supervision to actual stroke pixels not only enhances character recognition performance but also unexpectedly improves detection results. This improvement can be attributed to the fact that stroke regions naturally align with the true contours and boundaries of text, offering higher semantic density and stronger task relevance. By applying classification supervision only to these high-quality pixels, the model learns to better distinguish text from background, resulting in sharper feature boundaries and improved discriminative capability. This further contributes to improved localization accuracy and recall in the detection module.

It is worth noting that the fine-grained annotation and selective classification strategy demonstrates clear advantages in the classification task. As shown in Figure 7, the classification loss curves of the three annotation strategies exhibit noticeable differences during training. The conventional all-pixel supervision approach (corresponding to Exp. 3) maintains the highest loss throughout the training process, suggesting that the model is affected by a large number of irrelevant regions and struggles to converge effectively. The bbox-based selection strategy (bbox-sel, corresponding to Exp. 2) shows a faster initial decline compared to all-pixel supervision, but still involves many non-discriminative pixels, which limits the convergence speed. In contrast, our stroke-aware region selection method (strock-sel, corresponding to Exp. 1) significantly reduces the loss at the early stage of training and

consistently maintains the lowest classification error throughout the process, indicating higher learning efficiency and better convergence behavior.

Table 3. Ablation experiments on stroke-aware loss and joint tasks on the test set. Best results are highlighted in bold. "Prec." and "Rec." denote precision and recall, respectively. "GS", "Hm.", and "Hm.-CA" represent GlyphScore, Hmean, and Hmean-CA.

Exp	Models	mIoU ↑	RMSE ↓	Prec. ↑	Rec. ↑	CA ↑	GS ↑	Hm. ↑	Hm.-CA ↑	HCG ↑
1	UniText	**0.6691**	0.1426	**0.9822**	**0.9588**	**0.8805**	0.7632	**0.9704**	**0.9232**	**0.8629**
2	bbox-sel	0.6195	0.1538	0.9746	0.9050	0.8037	0.7328	0.9385	0.8659	0.8165
3	all-pixel	0.6683	**0.1418**	0.9655	0.9236	0.5405	**0.7633**	0.9441	0.6875	0.7110
4	w/o $\mathcal{L}_{res}$	-	-	0.8462	0.9077	0.6102	-	0.8759	0.7193	-
5	w/o T_{res}	-	-	0.9588	0.9479	0.8764	-	0.9533	0.9132	-
6	w/o T_{rec}	0.6237	0.1498	0.9810	0.9502	-	0.7370	0.9654	-	-
7	w/o T_{det}	0.6657	0.1451	-	-	0.6075	0.7603	-	-	-

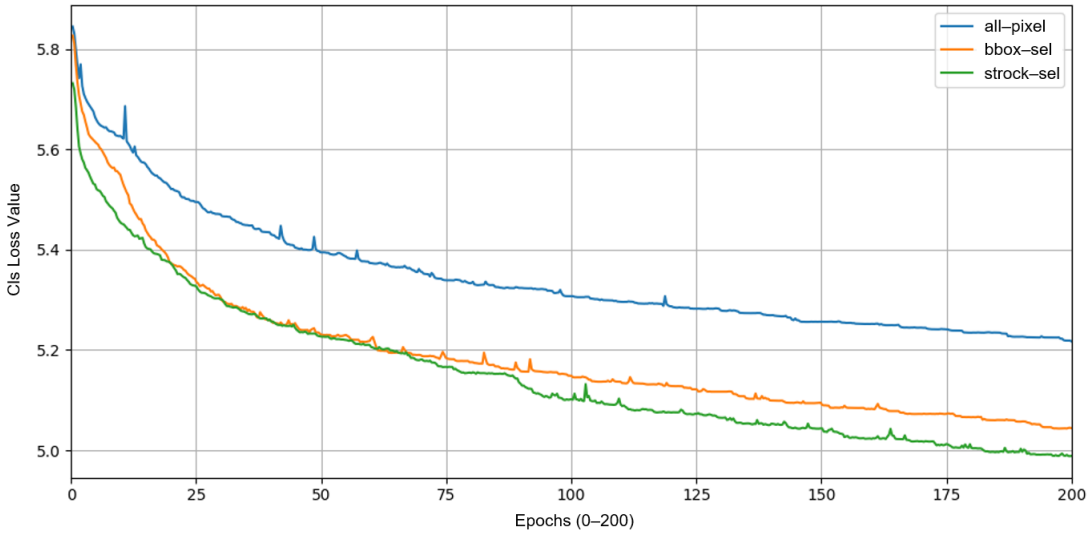


Figure 7. Comparison of classification loss across three pixel-level supervision strategies.

In particular, Exp. 3 in Table 3 shows a slight advantage in the GS, which is mainly attributed to the application of classification loss to all output pixels. This approach introduces more boundary and background information during training, helping the model capture the completeness of character contours and the continuity of structural patterns. However, its CA score drops significantly to 54.05%, indicating that applying classification supervision uniformly to all pixels introduces a large number of irrelevant or weakly relevant regions. This may lead to overfitting or semantic drift in the classification decision process. Such all-pixel supervision lacks selectivity toward target regions and ultimately compromises the overall performance of the model.

A comparison between removing the glyph restoration loss (Exp. 4) and completely removing the glyph restoration branch (Exp. 5) shows that the joint training of three tasks not only enables effective glyph reconstruction but also significantly improves character detection and recognition performance. Specifically, UniText achieves the best results in key metrics such as precision, CA, Hmean-CA, and HCG, indicating that the glyph restoration task plays a positive role in supporting the main tasks within the multi-task learning framework. This improvement is primarily attributed to the complementary relationship among tasks; the glyph restoration task provides pixel-level structural supervision, enhancing the model's ability to perceive local information such as character contours and stroke details.

As a result, it offers richer and more accurate structural features to support both detection and recognition tasks.

It is noteworthy that, compared to completely removing the glyph restoration head (Exp. 5), removing only its loss function while retaining the branch structure (Exp. 4) leads to a more pronounced drop in detection and recognition performance. This phenomenon reveals the issue of gradient interference in multi-task architectures. When an auxiliary task lacks explicit supervision during training, its branch may still generate uncontrolled or even disruptive gradients during backpropagation, which can negatively affect the optimization of the shared backbone and undermine the learning of the primary tasks. These results not only validate the positive role of the glyph restoration task in three-task joint learning, but also underscore the necessity of caution when designing unsupervised auxiliary branches in multi-task models, in order to avoid potential negative effects on the shared backbone network.

In the extended task ablation experiments, we remove the recognition task (Exp. 6) and observe a decline in segmentation performance. This indicates that the recognition task provides auxiliary semantic supervision, which helps improve segmentation accuracy. However, the detection metrics remain relatively stable, suggesting that the recognition task has a limited impact on detection. This is primarily because the recognition and detection tasks share a large portion of low-level and mid-level features, and are ultimately processed by separate task-specific heads. Under this architecture, the detection task does not directly depend on the recognition task.

When the detection task is removed (Exp. 7), the recognition performance drops significantly, with the CA score decreasing from 0.8805 to 0.6075. Although the recognition and detection tasks appear structurally independent, with shared features and separate heads, the detection task plays a critical guiding role during training. It provides region-level supervision that encourages the shared feature maps to focus on relevant text regions. This implicit guidance is essential for the recognition task to accurately extract and classify textual content. Without detection supervision, the recognition task becomes more vulnerable to background noise, leading to a substantial performance drop. In contrast, the segmentation task operates as a dense, pixel-wise prediction over the entire image and relies mainly on shared spatial features. Therefore, it exhibits only a minor decrease in performance when the detection task is removed.

5.3. Text Detection and Recognition Performance Analysis

To comprehensively evaluate the performance of the proposed UniText model on historical and stone inscription text detection tasks, we conduct comparative experiments against several mainstream text detection methods on a test set of 2000 images. The results are presented in Table 4. In terms of the key detection metric Hmean, UniText achieves a score of 0.9704, which is notably higher than DB [26] (0.7955) and EAST [27] (0.9603), and is comparable to the more accurate FCENet [28] (0.9772) and PSENet [29] (0.9847), demonstrating competitive detection capabilities.

Table 4. Detection performance comparison between the proposed UniText model and mainstream detection methods on the test set.

Methods	Hmean ↑	Precision ↑	Recall ↑	FPS ↑
Ours	0.9704	0.9822	0.9588	**74.6142**
DB [26]	0.7955	0.9068	0.7085	27.1825
EAST [27]	0.9603	0.9892	0.9331	44.8303
PSENet [29]	**0.9847**	**0.9997**	**0.9702**	35.0545
FCENet [28]	0.9772	0.9932	0.9616	39.5791

It is worth noting that, unlike other methods that focus solely on text detection, UniText additionally performs character recognition and glyph restoration, enabling a more integrated form of end-to-end text understanding. Although it incorporates multiple tasks, UniText achieves an inference speed of 74.6 FPS under a batch size of 1, which is faster than FCENet [28] (39.6 FPS) and PSENet [29] (35.1 FPS), with relative improvements of approximately 1.9× and 2.1×, respectively. These results suggest that UniText achieves a reasonable trade-off between detection accuracy, inference efficiency, and multi-task capability, making it a potentially suitable choice for practical applications.

Figure 8 presents the visualization results of our model on the historical document image (Figure 8a), the stone inscription image (Figure 8b), and a real-world scene image (Figure 8c). It can be observed that the model handles multi-oriented character-level detection and recognition in complex backgrounds effectively. In Figure 8a, although the handwritten style of the historical document exhibits distinct historical characteristics and the page contains visual noise due to aging, the model is able to accurately localize character boundaries and produce clear recognition results. In Figure 8b, we present a stone inscription image that poses additional difficulties such as low foreground-background contrast, erosion of character edges, and uneven illumination. UniText successfully detects and recognizes each character instance with high spatial precision. Furthermore, Figure 8c illustrates the model's performance on a real-world image captured under natural scene conditions. This image contains text with varying font sizes and is taken from a non-frontal angle. Despite these challenges, the model accurately detects individual characters and maintains robust recognition performance. This demonstrates the model's good generalization ability and robustness to real-world distortions.

(a) (b) (c)

Figure 8. Sample results of character-level detection and recognition by the proposed model. (**a**,**b**) Examples from the test set. (**c**) Real-world image. The Chinese characters are the primary visual content used to evaluate model performance and are not intended for linguistic interpretation. Blue frames indicate the detected bounding boxes for individual Chinese characters.

5.4. Glyph Restoration Effectiveness Analysis

Figure 9 presents the visualization results of glyph restoration by the proposed method on a historical document image and a stone inscription image. For the historical document example, despite the presence of paper aging, stains, and complex background textures, the model effectively preserves the stroke details of characters while suppressing irrelevant background noise. The restored output closely aligns with the ground-truth label in terms of character layout, spacing, and overall typesetting, demonstrating strong structural consistency.

For the stone inscription example, the model is able to extract clear character contours and reconstruct structurally complete text regions, even under challenging conditions such as intricate carved textures, low foreground–background contrast, and partial surface erosion. This performance benefits from the integration of a multi-scale feature fusion

strategy, which enables the model to capture both fine-grained stroke details and global character structures, thereby improving restoration quality in degraded scenarios.

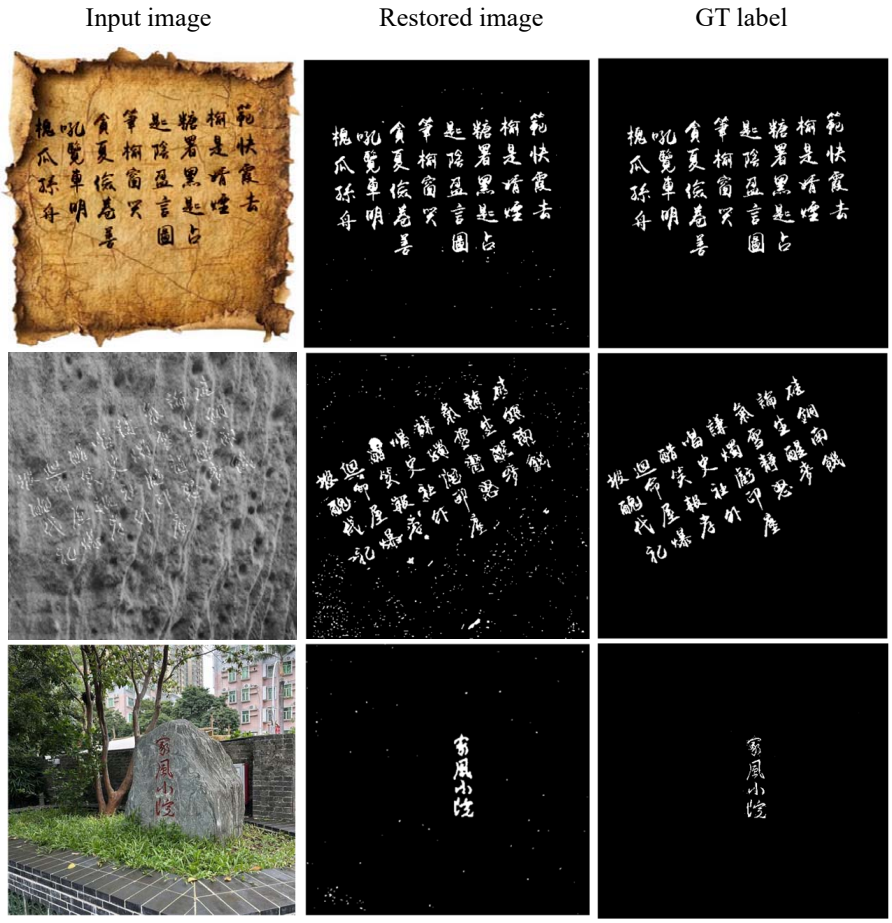


Figure 9. Examples of glyph restoration results. From left to right: input image, restored output, and ground-truth label. The first two rows are from the test set (historical documents and stone inscriptions), while the third row shows a real-world image captured in the wild. The Chinese text is used to evaluate restoration quality and does not require translation.

It is worth noting that the glyph restoration task is not performed as a separate image preprocessing step, but is integrated into an end-to-end multi-task framework jointly with text detection and recognition. This design is expected to support mutual enhancement among the three tasks. Glyph restoration has the potential to improve the structural modeling of characters for detection and recognition, while detection and recognition are able to provide more precise spatial guidance and semantic constraints for restoration.

While the restoration results are generally satisfactory, some limitations remain. In regions with severe erosion or extremely low signal-to-noise ratios, particularly evident in the stone inscription example, residual background noise or incomplete stroke recovery may still occur. Future work may explore transformer-based global attention to further enhance restoration quality under extreme deterioration conditions.

6. Conclusions

In this paper, we propose UniText, a unified end-to-end framework for page-level text image analysis, which jointly performs character detection, recognition, and glyph restoration. The goal is to enable the high-fidelity visual reconstruction of Chinese historical documents and inscriptions, which often suffer from complex corruption such as erosion, missing strokes, and uneven backgrounds. By leveraging multi-task learning and instance-

aware supervision, UniText effectively captures both the visual structures and semantic representations of ancient characters, enhancing overall robustness in image understanding and restoration. Experimental results on a curated dataset of historical books and stone inscriptions demonstrate that UniText achieves competitive performance in character detection, while also delivering strong results in glyph restoration and character recognition. These findings validate the effectiveness of our approach in handling challenging text image degradation scenarios. This work offers a novel perspective on ancient text image processing, contributing a unified solution for character-level analysis and restoration in historical contexts. With its modular and extensible design, UniText establishes a solid foundation for future applications in computational heritage, archival image analysis, and digital humanities, and shows promising potential for broader adoption in visual document intelligence.

Author Contributions: Conceptualization, L.S. and Z.W.; methodology, L.S. and Z.W.; software, L.S. and Z.W.; formal analysis, L.S. and Z.W.; investigation, L.S., X.H. and B.Z.; resources, S.-K.T., J.H. and S.M.; data curation, L.S.; writing—original draft preparation, L.S.; writing—review and editing, L.S., Z.W., S.-K.T., J.H. and S.M.; visualization, L.S.; supervision, S.-K.T., J.H. and S.M.; project administration, S.-K.T., J.H. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qi, H.; Yang, H.; Wang, Z.; Ye, J.; Xin, Q.; Zhang, C.; Lang, Q. AncientGlyphNet: An advanced deep learning framework for detecting ancient Chinese characters in complex scene. *Artif. Intell. Rev.* **2025**, *58*, 88. [CrossRef]
2. Shen, L.; Chen, B.; Wei, J.; Xu, H.; Tang, S.K.; Mirri, S. The challenges of recognizing offline handwritten Chinese: A technical review. *Appl. Sci.* **2023**, *13*, 3500. [CrossRef]
3. Fang, K.; Chen, J.; Zhu, H.; Gadekallu, T.R.; Wu, X.; Wang, W. Explainable-AI-based two-stage solution for WSN object localization using zero-touch mobile transceivers. *Sci. China Inf. Sci.* **2024**, *67*, 170302. [CrossRef]
4. Zhu, S.; Xue, H.; Nie, N.; Zhu, C.; Liu, H.; Fang, P. Reproducing the Past: A Dataset for Benchmarking Inscription Restoration. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 7714–7723.
5. Zhang, P.; Li, C.; Sun, Y. Stone inscription image segmentation based on Stacked-UNets and GANs. *Discov. Appl. Sci.* **2024**, *6*, 550. [CrossRef]
6. Xu, Y.; Zhang, X.Y.; Zhang, Z.; Liu, C.L. Large-scale continual learning for ancient Chinese character recognition. *Pattern Recognit.* **2024**, *150*, 110283. [CrossRef]
7. Ma, W.; Zhang, H.; Jin, L.; Wu, S.; Wang, J.; Wang, Y. Joint layout analysis, character detection and recognition for historical document digitization. In Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 8–10 September 2020; IEEE: New York, NY, USA, 2020; pp. 31–36.
8. Shimoyama, H.; Yoshida, S.; Fujita, T.; Muneyasu, M. U-Net Architecture for Ancient Handwritten Chinese Character Detection in Han Dynasty Wooden Slips. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2023**, *106*, 1406–1415. [CrossRef]
9. Sober, B.; Levin, D. Computer aided restoration of handwritten character strokes. *Comput.-Aided Des.* **2017**, *89*, 12–24. [CrossRef]
10. Poddar, A.; Chakraborty, A.; Mukhopadhyay, J.; Biswas, P.K. Texrgan: A deep adversarial framework for text restoration from deformed handwritten documents. In Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, Jodhpur, India, 19–22 December 2021; pp. 1–9.
11. Koch, P.; Nuñez, G.V.; Arias, E.G.; Heumann, C.; Schöffel, M.; Häberlin, A.; Aßenmacher, M. A tailored Handwritten-Text-Recognition System for Medieval Latin. *arXiv* **2023**, arXiv:2308.09368.

12. Locaputo, A.; Portelli, B.; Colombi, E.; Serra, G. Filling the Lacunae in ancient Latin inscriptions. In Proceedings of the 19th IRCDL (The Conference on Information and Research Science Connecting to Digital and Library Science), Bari, Italy, 23–24 February 2023; pp. 68–76.
13. Aguilar, S.T.; Jolivet, V. Handwritten text recognition for documentary medieval manuscripts. *J. Data Min. Digit. Humanit.* 2023. [CrossRef]
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2298–2304. [CrossRef] [PubMed]
17. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: New York, NY, USA, 2016; pp. 565–571.
18. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
20. Shit, S.; Paetzold, J.C.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J.P.; Bauer, U.; Menze, B.H. clDice-a novel topology-preserving loss function for tubular structure segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16560–16569.
21. Yang, H.; Jin, L.; Huang, W.; Yang, Z.; Lai, S.; Sun, J. Dense and tight detection of Chinese characters in historical documents: Datasets and a recognition guided detector. *IEEE Access* **2018**, *6*, 30174–30183. [CrossRef]
22. Xu, Y.; Yin, F.; Wang, D.H.; Zhang, X.Y.; Zhang, Z.; Liu, C.L. CASIA-AHCDB: A large-scale Chinese ancient handwritten characters database. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; IEEE: New York, NY, USA, 2019; pp. 793–798.
23. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on robust reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; IEEE: New York, NY, USA, 2015; pp. 1156–1160.
24. Pérez, P.; Gangnet, M.; Blake, A. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 577–582.
25. PaddleOCR. PaddleOCR Documentation. 2024. Available online: https://paddlepaddle.github.io/PaddleOCR (accessed on 15 November 2024).
26. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time scene text detection with differentiable binarization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *34*, 11474–11481. [CrossRef]
27. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. East: An efficient and accurate scene text detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5551–5560.
28. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier contour embedding for arbitrary-shaped text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3123–3131.
29. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9336–9345.