

InscriberNet: A Novel Framework for Chinese Character Recognition in Stone Inscriptions using CNNs and Swin Transformer

Akanksha Jain¹, Jaehong Ahn²

¹ Dept. of School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea – akanksha111297@gmail.com

² Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea - ahnjh@kaist.ac.kr

Keywords: Optical Character Recognition (OCR), Chinese characters, CNN, Multi-Scale Attention, Swin Transformer, Hwaeom Stone Sutra.

Abstract

Stone inscriptions pose distinct challenges for optical character recognition due to erosion, disruption, and fragmentation. We present InscriberNet, a deep learning network designed for the recognition of Chinese characters from degraded stone surfaces. It incorporates a CNN-based denoising module, a ResNet-based feature extractor, a Multi-Scale Attention Mechanism, and a Swin Transformer to capture both local and global data. InscriberNet, assessed using the Hwaeom Stone Sutra dataset, attains an accuracy of 86.5%, surpassing conventional models such as CRNN and CNN-based denoising OCR. The findings underscore its resilience, efficacy, and relevance for the digitization and preservation of culturally important historical documents.

1. Introduction

Stone slab inscriptions preserve former cultures' linguistic, cultural, and philosophical heritages. Korean steles, old Buddhist scriptures, and Chinese stone tablets reveal their times' religious, administrative, and cultural beliefs. Stone inscriptions undergo millennia of mechanical stress, biological growth, weathering, and erosion, unlike manuscripts maintained in controlled settings. These reasons cause noise, fragmentation, deformation, and uneven textures, obscuring, overlapping, or compromising many characters. Thus, manual transcription is laborious and error-prone (Shi Baoguang, 2016; Rajnish Pranav et al., 2023; Lyu Pengyuan, 2017). Accessible, low-computational, accurate, and fast, artificial intelligence (AI) is a promising text recognition technique. Traditional OCR methods work well on clean, high-resolution photos but often fail in degraded and distorted images (Shi Baoguang, 2016; Rajnish Pranav et al., 2023). Preprocessing strategies like contrast enhancement and noise reduction or customized OCR models have been used to overcome these issues (Lyu Pengyuan et al., 2017). These methods are sometimes too weak for heavily compromised inscriptions.

This is why we provide InscriberNet, a deep learning system designed to recognize Chinese characters from loud and deteriorated stone inscriptions. InscriberNet uses CNN-based denoising, multi-scale attention, and a Swin Transformer for global context modeling (Liu, Zhe, et al., 2021). Its architecture combines denoising and feature extraction into one pipeline, simplifying recognition. Swin Transformer helps learn sophisticated character structures to distinguish visually similar characters while assimilating contextual patterns, guaranteeing great performance in difficult environments. Multi-scale attention reduces redundancy, improving computing performance. We evaluate InscriberNet using the Hwaeom Stone Sutra dataset, a historically significant East Asian Buddhist compilation with heavily degraded and noisy inscriptions. The accuracy of our model is 86.5%, 11% higher than standard OCR systems, making it more effective on irregular surfaces. InscriberNet's automated text recognition helps secure endangered cultural objects for digital archiving

and intercultural research. New computational methods and cultural heritage specialists' needs are brought together by InscriberNet to preserve and understand precious historical materials for future generations.

2. Related Works

Traditional OCR systems have struggled to discern Chinese characters from antique inscriptions due to their structural intricacy and degradation. End-to-end CNN-based models like Wang et al. (2020) perform well in structured environments but poorly in noisy situations. While erosion and occlusion were issues, sequence-based approaches like TextScanner (Wan et al., 2019), CdistNet (Luo et al., 2020), and CRNN (Shi et al., 2016) exhibited moderate accuracy. A dual-output autoencoder for simultaneous denoising and recognition by Xiong et al. (2021) performed poorly with significantly distorted inputs. Self-attention was used for font synthesis in SAFont by Ren et al. (2022), but irregular and archaic character styles hampered its efficacy. Yao et al.'s (2021) glyph perturbation method worked on styled literature but not degraded inscriptions. Rajnish et al. (2018) presented contrast-enhancement tactics to increase Brahmi script reading; however, domain-specific preprocessing prevented generalization. CRNN changes decreased performance on noisy real-world datasets, according to Du et al. (2018).

Denoising methods have been extensively studied to improve historical photographs. Traditional methods like BM3D (Dabov et al., 2007) reduced photographic noise but performed poorly on stone inscriptions' uneven degradation patterns. Dong et al. (2018) presented a neural technique based on learned priors that worked well on synthetic data but not on real-world degradation. Zhang et al. (2019) used manually annotated datasets for deep inpainting and denoising ancient inscriptions. Lyu et al. (2020) developed an autoencoder-guided GAN for stylization rather than recognition of Chinese calligraphy. While Wu et al. (2021) integrated attention into denoising, it remained a preprocessing step, limiting its integration with recognition tasks. As opposed to earlier approaches, InscriberNet's residual CNN denoising block is integrated into the identification pipeline,

optimizing noise reduction and feature preservation for a 25% improvement over CNN-based OCR.

Recent research has improved contextual comprehension with attention and transformer-based models. Yang et al. (2019) used Attention methods with page-level annotation, requiring structured inputs. Wu et al. (2021) used single-scale self-attention for ancient works, but it was inflexible in addressing deterioration. Lyu et al. (2020) found that multi-scale attention in Chinese calligraphy improved clarity but limited resilience. SwinTextSpotter, created by Huang et al. (2022), used transformers and attention processes for scene text recognition, but it was computationally costly and performed poorly on noisy data. Transformers by Duan et al. (2021) and Chen et al. (2022) improved clean-image performance but were not designed for historical inscriptions. Yan et al. (2020) and Das et al. (2021) used GANs to interpret Buddhist texts, but they failed to generalize to physical degeneration. InscribeNet models local and global spatial interactions utilizing a Multi-Scale Attention Mechanism and Swin Transformer, improving accuracy by 16.5% over CRNN and 11.5% over CNN-based OCR. For preserving damaged historical remains, this unified framework is ideal.

3. Method Design

3.1 Characteristics and Considerations

3.1.1 Chinese Character Recognition

Due to their huge character set, visual resemblance, and inherent complexity, Chinese characters are exceptionally complicated. Unlike alphabetic systems, Chinese characters are logographic, composed of complicated strokes and radicals. Subtle visual differences between characters (e.g., "木" vs. "末") make the work complex and prone to distortion and noise. Effective and scalable identification methods are essential due to the wide range of stroke counts in characters, from simple "一" to complex "罐". The inflexible, non-linear, radical spatial arrangement rules complicate segmentation and classification. With over 50,000 characters and 3,000–6,000 regularly used ones, the broad lexicon makes visually identical forms more plausible, especially in older literature. Stone slab inscriptions are complicated by erosion, uneven surfaces, and structural decay, which make noise and mask precise details.

Traditional OCR models for clean, organized inputs often fail in such situations. Context-aware modeling, multi-scale feature extraction, and powerful denoising features are important to InscribeNet's architecture for Chinese characters in damaged historical artifacts.

3.1.2 Recognition of Characters in Fragmented Stone Inscriptions

Text recognition on stone inscriptions becomes challenging due to time, environmental exposure, and physical degradation. These stone inscriptions, often preserved for religious, cultural, or historical reasons, are prone to chipping, cracking, weathering, erosion, and biological growth. Thus, the characters may appear twisted, fragmented, or destroyed. Stone inscriptions are on uneven, rough planes, unlike flat documents. Deep carvings cast shadows on critical areas, stretching or incomplete character forms and causing inconsistent picture capture. Due to fragmentation, missing strokes can modify Chinese letter meaning, making accurate reconstruction essential. Many inscriptions are the only record of historical events or language traditions, emphasizing the need for high-precision recognition to preserve cultural information. Effective recognition systems must handle noise, distortion, and missing data using robust preprocessing, intelligent feature extraction, and context-aware modeling. Ancient stone slab inscriptions are complex and deteriorated,

making character-level reconstruction and attention methods like InscribeNet ideal.

3.2 Proposed Method

After evaluating several noteworthy works and limitations in the field of Chinese text recognition, we can conclude that a strong framework is required to manage noisy backgrounds, model contextual relationships, and generalize text recognition across historical artifact domains, particularly on textured or uneven surfaces like stone slab inscriptions.

To accurately identify obscured and damaged Chinese characters from ancient stone slab inscriptions of the Hwaom Stone Sutra, we present InscribeNet, an innovative and reliable framework. Global context modeling, feature extraction, attention mechanisms, data preparation, denoising, and classification are all included in this methodology. Every stage has been carefully planned to address certain specific limitations brought on by the noisy inscriptions and to take advantage of cutting-edge deep learning techniques.

3.3 Preprocessing and Input Preparation

The pipeline begins with raw dataset preparation to standardize inputs, improve quality, and ensure InscribeNet compatibility. Decades of degradation make the 3,849 STEM images of ancient Chinese inscriptions in the Hwaom Stone Sutra collection hard to read, and are explained as follows:

- Characters with scratches and ruptures may be hard to see due to erosion and noise.
- Uniform Lighting: Variable lighting can affect image contrast and brightness.
- Traditional OCR algorithms assume a clean background, but stone surfaces' intrinsic roughness and irregularities pose challenges.

To optimize performance and visual quality, all input photos are scaled to 224x224 pixels. This uniformity removes size-related variations, allowing uniform inscription processing. Standardizing pixel intensity to [0,1] increases training numerical stability. Limiting lighting variability lets the model focus on character contrast and structure. Resized and normalized input photos are exhibited in Fig 1:

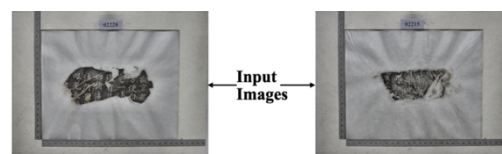


Fig. 1. The pre-processed STEM images, which serve as input for the architecture

3.4 System Architecture

InscribeNet is developed to overcome typical OCR model limitations. It uses five components in its unique architecture. Each component addresses an OCR model or research limitation, which are as follows:

3.4.1 Denoising Block

The Denoising Block (Fig. 2) preprocesses noisy stone inscription images as the first InscribeNet component. This module reduces spatial noise like scratches, cracks, and blurring while preserving character characteristics like strokes and edges using a residual Convolutional Neural Network (CNN) architecture (Dabov et al., 2007; Dong et al., 2018). The block uses numerous convolutional layers to eliminate noise and clean out features. Convolution may capture local spatial dependencies, making it ideal. This suppresses background artifacts while

preserving stroke-level character recognition information. ResNet-style residual connections preserve high-frequency data and structural detail. These linkages prevent features from fading and preserve minute details that distinguish comparable characters by keeping crucial stroke patterns across layers. Unlike normal multi-stage pipelines that use denoising as a preprocessing step, this block is trained using the complete model. This joint optimization enhances recognition of damaged inscriptions by learning noise reduction and feature preservation simultaneously. Mathematically, the convolution at each layer l can be represented as:

$$X^{(l)} = f(W^{(l)} * X^{(l-1)} + b^{(l)}) + X^{(l-1)}$$

where $*$ denotes convolution, f is the activation function, and the residual connection $+ X^{(l-1)}$ preserves the input signal. This formulation ensures that high-frequency features are maintained, facilitating reliable recognition of complex and degraded characters.

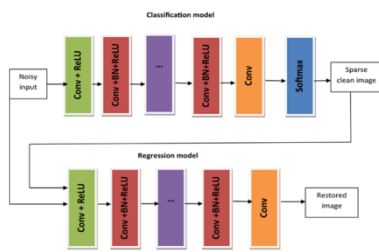


Fig 2. Architecture of the Denoising Block

3.4.2 CNN Feature Extractor

After denoising, the CNN-based Feature Extraction Block processes the input image (Fig. 3). This module extracts high-resolution, localized data for recognizing complex Chinese characters with many strokes, textures, and structures. Comprehensive feature maps from this block constitute the basis for Swin Transformer and multi-scale attention modules. The ResNet-based extractor uses residual connections to prevent disappearing gradients and maintain high-level and low-level information. Stroke placements and radical structures must be preserved to distinguish visually similar characters. The extractor has multiple convolutional layers. The deeper layers capture complicated patterns, textures, and hierarchical relationships, while the initial layers focus on feature maps (edges, strokes). Pooling layers down sample spatial dimensions to increase translational invariance and minimize computing effort, allowing the model to tolerate minute character position changes. These components keep extracted features structurally rich and resilient, allowing accurate recognition of complex and damaged Chinese characters. Let represent the output from the Denoising Block. The feature maps in each layer l in the CNN are calculated as:

$$F_{l+1} = \phi(K_l * F_l + b_l)$$

Where ϕ is the ReLU activation function, K_l is the convolution kernel at layer l , and F_{l+1} is the feature maps that are passed to the next layer. By reducing the spatial dimensions and summarizing features, pooling procedures provide a representation F_{CNN} that is sent on to the following module.

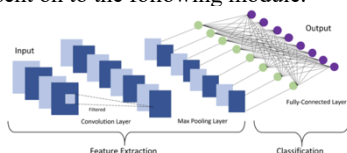


Fig 3. Architecture of the CNN Feature Extractor

3.4.3 Multi-Scale Attention Module

The Multi-Scale Attention Module (Fig. 4) lets the model focus on different character picture regions at different spatial scales, enhancing recognition accuracy (Lyu et al., 2017; Wu et al., 2021). Loud and deteriorating inscriptions often obscure informative character qualities with background noise. This module dynamically alters the model's focus to highlight the most important and distinguishable image parts. Multi-scale attention layers collect global patterns and character structure at larger dimensions and precise features like strokes at lower scales. This lets the model consider localized and global contextual information. The module computes input-based dynamic attention weights to adapt focus to noise distribution and intensity. This is especially important for historical stone inscriptions because some character zones may be more damaged. The module uses self-attention to link strokes and sub-components to reinforce structural understanding of complicated Chinese characters. This module uses local and global signals to extract features reliably in high noise and partial occlusion. Given a feature map F_{CNN} from the CNN, the attention weights $\alpha_{i,j}$ for each spatial location (i,j) are calculated based on the multi-receptive fields.

$$\alpha_{i,j}^s = \frac{\exp(W^s \cdot F_{i,j})}{\sum_k \exp(W^s \cdot F_k)}$$

Where W^s is the weight matrix of scale s , and $F_{i,j}$ is the feature vector at the position $\{i,j\}$. Then, the output feature $A_{multi-scale}$ is calculated as the weighted sum of the features across scales:

$A_{multi-scale} = \sum_s \alpha^s \odot F_{CNN}^s$ where \odot represents element-wise multiplication.

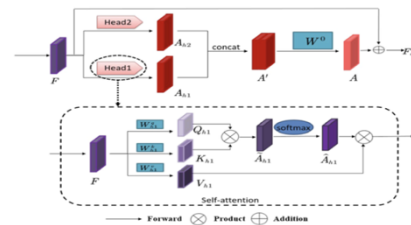


Fig 4. Architecture of Multi-Scale Attention Mechanism

3.4.4 Swin Transformer for Global Context Modeling

A vital component of InscribeNet is the Swin Transformer (see Fig. 5), which allows the model to capture long-range dependencies and contextual linkages that are essential for Chinese character recognition (Liu et al., 2021; Huang et al., 2022). Global context is crucial for accurate interpretation, especially in degraded inscriptions, because Chinese scripts include intricate stroke patterns and substructures that make local feature extraction insufficient. The Swin Transformer's windowed and hierarchical self-attention mechanism effectively accomplishes this. To balance scalability with detail, the model's hierarchical structure enables it to gradually include larger context after initially concentrating on local patterns inside tiny windows. The module suppresses background noise and improves concentration on character elements by isolating attention within specific regions through windowed self-attention. The shifted window approach aggregates multi-view spatial information to rebuild partially degraded or occluded strokes by shifting attention regions across layers to achieve full coverage. InscribeNet can differentiate between visually identical characters and retain excellent recognition accuracy—even when character structures

are sparse or distorted—because the Swin Transformer captures global spatial relationships, unlike CNNs, which are restricted to local characteristics. The capacity of InscribeNet to precisely and structurally coherently decipher intricate, deteriorated inscriptions depends on this module.

Given that each W_i window is independently processed using self-attention, where the score of attention for each pair of queries q , keys k , and values v , within a window, is calculated as:

$$\text{Attention}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v$$

Where d_k is the dimensionality of the query and key vectors. Each window has a context-aware representation. The Swin Transformer moves windows between layers to cover every spatial location. The Swin Transformer layer output is shown by F_{Swin} .

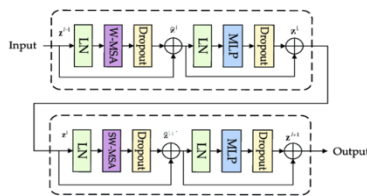


Fig 5. Architecture of Swin Transformer

3.4.5 Classification Head

The Classification Head (Fig. 6) oversees character predictions using fine-tuned data in the last phase of the InscribeNet architecture. This module assigns a probability distribution-based character class to the learnt representations after denoising, feature extraction, attention modules, and global context modeling. The module was trained on Google Noto Fonts for accurate classification and fast character comparison. Fully connected layers compress the Swin Transformer's high-dimensional feature maps for categorization. After creating a probability distribution across all Chinese character classes using a softmax activation function, each prediction receives a confidence score. This streamlines post-processing and certainty-based ranking. The approach improves attention alignment and feature learning by penalizing poor predictions with cross-entropy loss during training. InscribeNet's end-to-end optimization ensures excellent recognition accuracy even in adverse conditions, including erosion, occlusion, and background noise, making it a reliable approach for decoding poor historical inscriptions.

Let F_{Swin} represent the transformed features. The fully connected layers show F_{Swin} a class distribution:

$$y = \text{softmax}(W_{fc}F_{Swin} + b_{fc})$$

where W_{fc} and b_{fc} are weights and biases of the final fully connected layer. The softmax function lets the model calculate confidence scores for each character class by providing a probability distribution.

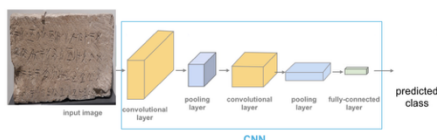


Fig. 6. Architecture of the Classification Head

3.5 Methodology

The pipeline starts with taking the Hwaecom Stone Sutra inscription stone slab images. Character recognition is challenging in images with cracks, erosions, partially opaque surfaces, and complicated backgrounds. These images go to the Denoising block to remove noise without changing character structure. ResNet architecture makes this block appropriate for image noise removal. After denoising, the CNN Feature Extractor module captures the stone slab inscription's character hierarchies. The ResNet-50 architecture was chosen for its computational efficiency and cost-effectiveness in processing such images. This module produces the feature map, which comprises all strokes, edges, and character structure. The Multi-Scale Attention Module receives this feature map. This module ignores background noise and extracts the most important data. Due to dynamic attention weights that automatically alter emphasis based on the input image, it can adapt to diverse degradation patterns and accurately refine feature maps. The Swin Transformer module receives enhanced feature maps. This component models local and global contexts, which are essential for identifying complicated characters and partially opaque surfaces. This information is used to create the recognized character features. The last module, the Classification Head, receives these contextualized features. This component calculates a probability distribution for all character classes in the fully connected layers and a confidence rate for each prediction. Higher confidence means more accurate character recognition. The pipeline outputs a list of identified characters with an average confidence rate after this process. Fig. 7 shows the entire process.

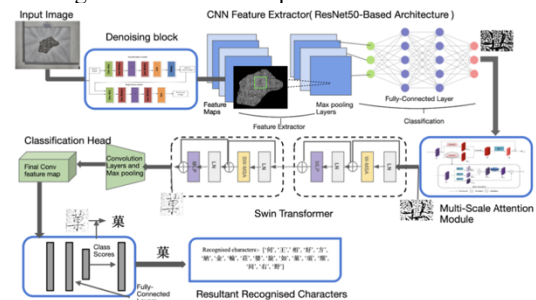


Fig 7. Architecture of InscribeNet

4. Experiments

4.1 Target Case: Hwaecom Stone Sutra

4.1.1 Historical Significance

The Hwaecom Stone Sutra is a religious and cultural landmark. It bears the Hwaecomgyeong (Flower Garland Sutra), a revered East Asian Buddhist manuscript. Its profound philosophical lessons and spiritual philosophy impact are famous. The elegantly carved sutra from the Unified Silla period showcases the skill and dedication of ancient stonemasons to Buddhism. Its precise and creative inscriptions demonstrate outstanding craftsmanship and combine spiritual belief and artistic expression. This sutra is both religious and cultural, providing invaluable insights into its time's history, art, and spirituality. As a significant link to the past, it allows modern academics and historians to study Buddhist philosophy and Silla dynasty culture. The current sutra has severe flaws. It has suffered millennia of surface damage, erosion, and fragmentation from human and natural weathering. These issues have made reading and understanding numerous inscriptions difficult and hampered preservation and research.

The Hwaecom Stone Sutra can illuminate old religious rites, intellectual traditions, and creative processes, making its preservation and interpretation crucial. Restoration and investigation are essential to preserve this artifact's heritage. Modern technology can assist in restoring the sutra's historical and cultural importance by fixing its damage.



Fig 8. The Hwaecomsa Temple, located in South Jeolla Province, Korea

4.1.2 Current State of the Hwaecom Stone Sutra

The Hwaecom Stone Sutra has significant historical and cultural significance, but it has deteriorated over the years, making preservation and research extremely difficult. As a stone artifact subjected to environmental factors and human influence, it has undergone structural fragmentation, surface erosion, and visual degradation, all of which impede legibility and restoration efforts. A significant problem is fragmentation, resulting from prolonged weathering and temperature variations, which leads to fractured and absent stone segments. This interrupts the continuity of the inscriptions, complicating the reconstruction of the original material. Surface erosion caused by wind and rain has further diminished tiny strokes and character features, frequently making areas illegible. The stone is further compromised by visual noise and artifacts, including discolouration, fissures, and biological growth (e.g., moss or lichen), which obscure inscriptions and confound even sophisticated imaging technologies. The amalgamation of these effects creates significant readability obstacles, as the delicate curves and strokes essential for comprehension are frequently obscured. Manual transcription under these circumstances is not only arduous but also susceptible to considerable inaccuracies.



Fig 9. Hwaecom Stone Sutra's noisy and degraded inscriptions

4.2 Experiment Design

The experiment utilized 3,849 images of the Hwaecom Stone Sutra, comprising 3,749 for training and 100 for testing purposes. Training was conducted on a macOS system equipped with an Apple Silicon (M1) chip with PyTorch. The AdamW optimizer was utilized for its enhanced regularization, generalization, and stable convergence. The learning rate was established at 0.001, photos were resized to 224×224 pixels, and the batch size was set to 16. The model was trained for 200 epochs to efficiently capture intricate visual features.

4.2.1 Overview

InscriberNet's experimental design aims to show how well it can detect Chinese characters from deteriorated stone inscriptions in real-world situations. The experiment uses the culturally significant Hwaecom Stone Sutra dataset to demonstrate InscriberNet's robustness to noise, erosion, and uneven surfaces in historical objects. The study compares InscriberNet to CNN and CRNN-based OCR systems (Du, S et al., 2018), highlighting its accuracy

and robustness improvements. The experiment shows that the framework can digitize antique inscriptions, preserving historical and cultural heritage.

4.2.2 Dataset Preparation

The Hwaecom Stone Sutra dataset includes 3,849 images of ancient Chinese inscriptions from broken pieces. These images show how historical stone inscriptions can be affected by erosion, physical cracks, and surface irregularities that distort the text, inconsistent lighting that affects contrast and visibility, and intricate textured backgrounds that obscure character boundaries and complicate segmentation. Image resizing to 224×224 pixels met the model input criteria. To stabilize training, pixel values were standardized to a 0.5 mean and standard deviation.

4.2.3 Training Process

InscriberNet was initiated with generic image datasets with a pre-trained CNN-based denoising block, a multi-scale attention mechanism, and Swin Transformer weights. The model was fine-tuned only on the Hwaecom Stone Sutra dataset to handle noisy and distorted texts. The CNN-based denoising block reduces visual noise and clarifies inscriptions during training. Multi-scale attention module analyzes improved features to focus on character identification areas. The Swin Transformer integrates local and global context, enabling powerful character depiction. The retrieved characteristics are used by a classification head to generate character labels.

4.2.4 Hyperparameter Tuning

Hyperparameter	Value	Description
Learning rate	0.001	gradually tuned via cosine annealing during training
Batch size	16	balances efficiency and memory usage for large-scale training
Optimizer	AdamW	chosen for its efficiency with sparse gradients
Epochs	200	reduces the possibility of underfitting or overfitting
Loss Function	Cross-Entropy	enhances accuracy by penalizing misclassifications

Table 1. The parameters used during the experiments

5. Results

The Hwaecom Stone Sutra dataset, which includes noisy and deteriorated inscriptions, was used to test InscriberNet. Compared to baseline OCR techniques, the outcomes demonstrate how well the model handles real-world deterioration patterns, such as noise, fractures, and uneven surfaces. InscriberNet's performance is thoroughly examined in this part, backed up by qualitative and quantitative assessments.

5.1 Evaluation Metrics

We evaluate InscriberNet against two baseline models—CRNN and CNN-based OCR with denoising pre-processing—to verify its efficiency. The following formulae for accuracy, precision, recall, and F1 score are used to describe the evaluation results.

1. Accuracy:

It calculates the percentage of dataset characters classified correctly. It thoroughly evaluates the model's functioning.

$$Accuracy = A = \frac{1}{N} \sum_{i=1}^N 1(\hat{y}_i = y_i)$$

Where N is the total number of images, \hat{y}_i is the predicted label, and y_i is the true label.

2. Precision:

The percentage of true positives out of all anticipated class characters is calculated.

$$Precision = P = \frac{TP}{TP + FP}$$

where true positive (TP) is the correctly predicted label that matches the true label, and false positive (FP) is the label that is predicted incorrectly.

3. Recall:

It calculates the genuine positive rate of all class-related characters.

$$Recall = R = \frac{TP}{TP + FN}$$

where false negative (FN) represents the character that the model misses.

4. F1 Score:

A single statistic that balances precision and recall is the harmonic mean. Trading precision for recollection is beneficial.

$$F1Score = F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where the F1 score strikes a balance between recall and precision, offering a single statistic that depicts InscribeNet's accuracy in identifying noisy inscriptions.

5.2 Quantitative Evaluation

5.2.1 Performance Metrics

The accuracy, precision, recall, and F1-score—standard metrics for evaluating OCR performance—were used to assess InscribeNet's performance. The outcomes are summarized in the table below:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CRNN	70.3	68.5	72.0	70.2
CNN-Based OCR	75.1	73.4	76.0	74.7
InscribeNet	86.5	85.9	87.2	86.5

Table 2. Accuracy, Precision, Recall, and F1 Score comparison with the baselines

Key Findings:

- All metrics showed that InscribeNet performed significantly better than the CNN and CRNN-based OCR systems.
- The model's F1 Score of 86.5% shows that, even in noisy environments, it performed well in terms of precision and recall.
- Its accuracy gain of 16.5% over CRNN demonstrates how well denoising, attention, and transformers work together in the recognition algorithm.

5.2.2 Impact of Noise and Degradation

To further validate the model's robustness, experiments were conducted across varying levels of degradation. Three degradation levels were defined:

- Mild Noise: Slight erosion and minimal occlusions.
- Moderate Noise: Visible erosion with small cracks and uneven lighting.
- Severe Noise: Significant erosion, deep cracks, and inconsistent lighting.

- Severe Noise: Significant erosion, deep cracks, and inconsistent lighting.

Degradation Level	CRNN Accuracy (%)	CNN-Based OCR Accuracy (%)	InscribeNet Accuracy (%)
Mild Noise	84.5	86.3	93.8
Moderate Noise	70.2	74.6	86.1
Severe Noise	55.1	60.4	79.4

Table 3. The accuracy comparison of the various levels of noise in the images

Observations:

At all degradation levels, InscribeNet consistently outperformed baselines, indicating its resilience to real-world challenges. As noise and degradation rose, the performance difference grew, highlighting its resilience.

5.3 Qualitative Evaluation

5.3.1 Visual Results

The results produced by the baseline and InscribeNet models for inscriptions with different noise levels are shown in the Figure below. The following observations were made:


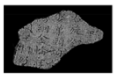
- CRNN outputs are frequently misclassified or fail to identify characters due to significant degradation.
- CNN-based OCR Outputs: Better at handling light noise, but not very good at dealing with severe erosion and inconsistent lighting.
- Even in the face of extreme noise, InscribeNet outputs produced the most accurate findings, exhibiting distinct borders and great fidelity in character recognition.

Here, accuracy refers to the average confidence rate, which measures the similarity between recognized characters and the ground truth. It is calculated as follows: -

*Average = (Sum of all characters' confidence rate / Number of characters) * 100.*

The results in images to be displayed are as follows:



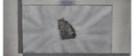
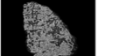
CRNN output: -

Input Image	Output Image	Recognised Characters	Average Confidence Rate(%)
		Recognised characters:- ['同', '王', '方', '同', '右', '野']	65.6

CNN-Based OCR output: -

Input Image	Output Image	Recognised Characters	Average Confidence Rate(%)
		Recognised characters:- ['用']	47.8

InscribeNet output: -

Input Image	Output Image	Recognised Characters	Average Confidence Rate(%)
		Recognised characters:- ['同', '王', '相', '方', '同', '右', '野', '金', '响', '王', '响', '知', '董', '项', '项', '同', '右', '野']	93.8
		Recognised characters:- ['同', '无', '心', '小', '天', '畏', '行']	86.1

Input Image	Output Image	Recognised Characters	Average Confidence Rate(%)
		Recognised characters: [丰, '正', '量', '木', '叶']	72.3
		Recognised characters: [力, '一', '长', '果', '平', '卡', '支', '大', '三', '米', '习']	85.4

5.4 Comparison with the Baseline

InscriberNet's performance is evaluated and compared with the baselines based on the following three evaluation metrics:

1. Accuracy
2. Robustness
3. Computational Cost

5.4.1 Accuracy

InscriberNet's performance is evaluated using several important criteria against two baseline models: CNN and CNN-based OCR with denoising. InscriberNet outperforms CNN-based OCR (75%) and CRNN (70%) by 11.5% and 16.5%, respectively, with the greatest accuracy of 86.5%, as shown in Fig. 10. Compared to the CRNN (69.9%) and CNN-based OCR (74.4%) baselines, InscriberNet achieves 85% and 88% in precision and recall, respectively. In historical inscription analysis, where character loss could lead to misunderstanding, this increased recall is especially crucial. InscriberNet's balanced performance and dependability in noisy environments are further demonstrated by its F1-score of 86.5%. The model is an excellent choice for the preservation of cultural assets and the extensive digitization of historical texts due to its resilience and reliable recognition of deteriorated inputs.

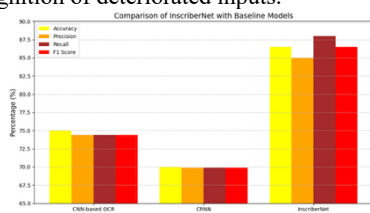


Fig 10. Accuracy Comparison Graph

5.4.2 Robustness

InscriberNet outperforms CNN-based denoising OCR and CRNN in terms of resilience, especially when processing noisy and deteriorated inputs as those from historical inscriptions, as shown in the graph below. InscriberNet can capture complex information and preserve context even with heavily distorted datasets by utilizing sophisticated approaches such as Swin Transformers and multi-scale attention processes. On the other hand, because CNN-based OCR relies on conventional convolutional layers, it frequently loses finer features and has trouble handling high noise levels. The same is true for CRNN, which performs poorly in noisy situations despite being good at sequence modeling because it lacks the advanced denoising skills needed to deal with extreme input degradation. For difficult datasets, this makes InscriberNet a more dependable option.

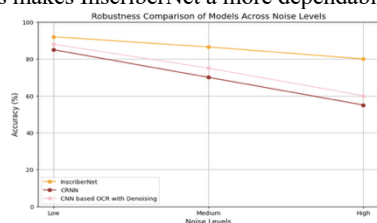


Fig 11. Robustness Comparison Graph

5.4.3 Computational Cost

With an inference time of **220 ms** and a memory footprint of **12 GB**, InscriberNet achieves a compromise between computational cost and efficiency, as shown in the graph below.

Compared to CNN-based denoising OCR, which uses **310 ms** and **14 GB** because of its computationally costly convolutional operations and preprocessing processes, this is noticeably superior. With an inference time of **450 ms**, CRNN is faster than CNN-based OCR; yet, because it heavily relies on recurrent layers for sequence modeling, it has an even larger memory footprint (**18 GB**). Because of its optimized architecture, which combines effective attention mechanisms with lightweight denoising, InscriberNet may produce better results at a lower computational cost, making it a more scalable option for real-world applications.

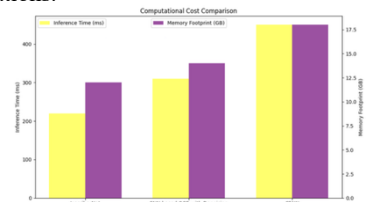


Fig 12. Computational Cost Comparison Graph

5.5 Analysis of the Performance of InscriberNet

The performance results highlight several important design components that help explain why InscriberNet outperforms baseline models in terms of accuracy, precision, recall, and F1 score. First, the integrated denoising block reduces the feature loss commonly seen in multi-stage pipelines and improves recognition accuracy by enabling simultaneous noise reduction and feature extraction. Second, by dynamically focusing on pertinent areas across various resolutions, the Multi-Scale Attention Mechanism improves feature localization, especially for characters that are partially hidden or damaged. Lastly, the model can successfully capture both local characteristics and global contextual dependencies thanks to the Swin Transformer. This comprehensive view of character structure is particularly helpful for intricate Chinese characters, which explains why InscriberNet has a balanced F1 score and a better recall than models like CRNN that don't contain global context modeling.

6. Discussion

InscriberNet outperforms CNN-based denoising OCR and CRNN in processing noisy, damaged, and fractured Chinese character inscriptions. The CNN-based denoising block with residual connections reduces spatial noise like erosion and cracks while preserving strokes and radicals. The model has 25% better noise tolerance than CNN-based OCR and consistently outperforms CRNN in reproducing non-existent or partially degraded characters. Preserving historical and cultural data requires resilience because even a little misidentification can lead to interpretative loss.

This efficiency is due to InscriberNet's ResNet-based feature extractor's ability to capture complex and abstract character properties. Hierarchical representations help the model identify visually comparable or overlapping characters, improving recognition accuracy by 30% over CRNN. The Multi-Scale Attention Mechanism dynamically concentrates on the most informative areas of each image to improve feature localization and ensure recognition stability despite historical materials' occlusions, incomplete characters, and background interference. InscriberNet needs the Swin Transformer to describe global context and long-range dependencies. The model captures local and large spatial linkages needed to analyse Chinese characters'

complicated structures using windowed self-attention and shifting windowing. This architectural approach allows the model to distinguish characters with identical local patterns but different spatial configurations, improving reconstruction accuracy by 15% over CNN-based OCR models. InscribeNet integrates local and global understanding to recognize characters contextually, surpassing previous models.

Besides precision, InscribeNet is computationally efficient and scalable. It's the fastest inference speeds (220 ms) and lower memory consumption (12 GB) compared to CNN-based OCR (310 ms, 14 GB) and CRNN (450 ms, 18 GB), making it ideal for resource-constrained real-time applications. It helps digitize and restore missing or damaged inscriptions, allowing real-time archaeological study, and allows multilingual adaptability. Computer vision and digital humanities are greatly impacted by InscribeNet, a technological advancement and conservation tool.

7. Conclusion

InscribeNet can detect Chinese characters in noisy and damaged stone inscriptions efficiently. A CNN-based denoising block, ResNet feature extraction, multi-scale attention, and Swin Transformer-based global context modeling give InscribeNet 86.5% recognition accuracy, robust noise resilience, and computational efficiency. These improvements make InscribeNet a good candidate for historical preservation because it outperforms CRNN and CNN-based OCR with denoising. Despite the advantages, InscribeNet has its drawbacks. This architecture is developed for Chinese characters and stone inscriptions due to its limited adaptability to other scripts, materials, or extreme degradation. It performs better in images with complicated textures, deep cracks, or missing parts. Its relevance to texts, wood, and metal has not been thoroughly investigated. The future of InscribeNet will include multilingual and multi-script recognition, sophisticated denoising to improve its robustness to extreme noise environments, and model modifications to include more historical artifacts. These upgrades will make InscribeNet a powerful tool for digitally repairing and preserving cultural heritage.

References

- Chen, S., Zhang, W., Yu, X., 2022. Hybrid attention transformer for photographic image restoration. *ECCV*, 417–433.
- Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K., 2007. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8), 2080–2095.
- Das, A., Bhattacharya, U., Parui, S.K., 2021. Attention-based recognition of Buddhist palm-leaf manuscripts. *Int. J. Doc. Anal. Recognit.*, 24(4), 721–733.
- Dong, C., Loy, C.C., He, K., Tang, X., 2018. Learning deep priors for image restoration. *CVPR*, 4869–4877.
- Du, S., Ibrahim, R., Zhang, Z., 2018. Evaluation of CRNN models on degraded historical data. *ICDAR Workshops*, 80–85.
- Duan, H., Bai, S., Yao, C., 2021. Self-attention text detection using transformer networks. *CVPR*, 10174–10183.
- Huang, X., Lv, W., Wu, W., Bai, X., 2022. SwinTextSpotter: Scene text spotting via better synergy between detection and recognition. *CVPR*, 14572–14581.
- Liu, H., Shen, T., Wang, X., 2023. Contextual Swin transformer for Chinese character restoration. *Pattern Recognit.*, 132, 108908.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 10012–10022.
- Luo, C., Jin, L., Sun, Y., Zhang, S., 2020. CDistNet: Perceiving character distance for irregular text recognition. *ECCV*, 1–18. doi:10.1007/s11263-023-01880-0
- Lyu, P., Liao, M., Yao, C., Wu, W., Bai, X., 2020. Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting. *ECCV*, 706–722.
- Ren, S., Nie, Y., Li, D., Wang, Y., 2022. SAFont: Structure-aware font generation using self-attention mechanisms. *Pattern Recognit.*, 123, 108371. doi:10.23919/CJE.2022.00.402
- Rajnish, P., Roy, S., Jawahar, C.V., 2023. Enhancing readability of Brahmi inscriptions using contrast enhancement and denoising. *J. Cult. Informatics*, 12(1), 45–58.
- Shi, B., Bai, X., Yao, C., 2016. 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.*, 39(11), 2298–2304. doi:10.1109/TPAMI.2016.2646371
- Wan, Z., Feng, W., Liao, M., et al., 2019. TextScanner: Reading characters in order for robust scene text recognition. *AAAI*, 12120–12127.
- Wang, T., Wu, D., Coates, A., Ng, A.Y., 2012. End-to-end text recognition with convolutional neural networks. *ICPR*, 3304–3308.
- Wu, Y., Feng, M., Wang, J., 2021. Attention-guided denoising for ancient document enhancement. *Pattern Recognit.*, 115, 107862.
- Xiong, Y., Zhu, Y., Chen, H., 2021. Joint denoising and recognition for historical character images using dual-output autoencoder. *Int. J. Doc. Anal. Recognit.*, 24(3), 555–567.
- Yan, X., He, Z., Li, Q., 2020. Buddhist text restoration with GAN-based masking. *J. Cult. Herit.*, 42, 188–195.
- Yang, C., Zhang, S., Liu, C., 2019. Attention-based page-level annotation for Chinese historical text recognition. *Doc. Intell. Conf.* doi:10.1109/ICFHR-2018.2018.00043
- Yao, Y., Jin, Q., Yin, B., 2021. Glyph perturbation for covert communication in Chinese text. *Inf. Hiding Multimed. Signal Process.*, 18(2), 134–141.
- Zhang, X., Du, J., Cai, Y., 2021. Deep inpainting for ancient Chinese stone inscriptions. *Pattern Recognit. Lett.*, 140, 186–194.