# High-Precision Character Extraction from Historical Japanese Manuscripts Using Tiled Inference with YOLOX

Yuya YOSHIZU[1], Lin MENG[2]

[1]Graduate School of Science and Engineering, Ritsumeikan University

[2]College of Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan

[1]ri0120si@ed.ritsumei.ac.jp, [2]menglin@fc.ritsumei.ac.jp

**Abstract**

Historical Japanese manuscripts are invaluable cultural assets, yet their characters are often obscured due to degradation such as stains, fading, and insect damage. To ensure reliable digital preservation and enable downstream restoration, highly accurate extraction of text regions is essential. This paper proposes a high-precision character detection framework based on YOLOX. Each manuscript page is padded and divided into overlapping $640 \times 640$ tiles, and detection is performed independently on each tile. The results are then merged using page-level non-maximum suppression (NMS). To further mitigate duplicate detections and boundary errors inherent to tiled inference, a central-region selection strategy is employed. Experiments on 11 manuscripts demonstrate that the conventional page-level YOLOX approach—processing entire pages resized to a fixed resolution—suffers from degraded detection performance, achieving only 79.8% recall due to loss of detail. In contrast, the proposed method combining tiled inference with central-region filtering achieves 0.976 precision, 0.989 recall, and 0.982 average precision (AP). It successfully separates main body characters from annotation characters and degradation-induced artifacts across RGB, grayscale, and monochrome images.

**Keywords:** Computer Vision , Deep Learning , Object Detection , YOLOX , Historical Documents

## 1 Introduction

Japan possesses a diverse range of cultural heritage, including architecture, art, and literature. Among these, ancient Japanese manuscripts serve as important records that reflect the culture, language, and philosophical thought of past eras. Preserving such materials digitally is crucial for

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365292

ensuring their accessibility to future generations and preventing further degradation due to environmental or temporal factors [1, 2, 3].

However, many historical manuscripts have deteriorated over time, suffering from stains, fading, and insect damage. These degradations obscure the original characters and complicate the process of digitization and analysis. In recent years, deep learning has been applied to cultural heritage preservation, including text detection, character recognition, and restoration. Nevertheless, accurate character extraction from ancient Japanese manuscripts remains a major challenge due to the presence of noise such as annotations, rubricated characters, and background artifacts. These non-text elements are often confused with main text regions, reducing the accuracy of subsequent restoration and recognition processes.

Traditional object detection methods, including YOLO-based approaches, have shown strong performance in document analysis. However, when applied to full manuscript pages, they often fail to detect small and fine characters accurately because resizing the entire page to a fixed input size reduces the effective pixel resolution. Furthermore, page edges and dense text areas tend to produce overlapping or missed detections.

To address these issues, this study proposes a high-precision character extraction framework using YOLOX [4] with tiled inference. In the proposed method, each manuscript page is divided into overlapping $640 \times 640$ tiles with zero-padding around the page. Only the detections whose bounding boxes fall within the central region of each tile are retained, and all results are merged through a page-level Non-Maximum Suppression (NMS). This approach improves the effective resolution for small strokes, prevents boundary duplication, and ensures consistent page-level aggregation.

Experimental results on 11 types of manuscripts demonstrate that the proposed method achieves an average text detection accuracy of 98.7% across RGB, grayscale, and monochrome documents, significantly outperforming conventional page-level inference. Moreover, the method successfully distinguishes main text characters from annotation characters and degradation-induced artifacts, enabling cleaner text regions for subsequent restoration and classification.

The major contributions of this paper are summarized as follows:

- A tiled inference pipeline for YOLOX that enables high-precision detection of small text regions in historical Japanese manuscripts.
- A page-level aggregation strategy that integrates overlapping tiles and suppresses duplicate detections using central-region filtering and non-maximum suppression.
- Comprehensive experiments showing robustness across manuscript domains (RGB, grayscale, monochrome) and superiority over conventional page-level inference.

Furthermore, this YOLOX-based character extraction is envisioned as a core component of a larger automatic document restoration system currently under development. In the future, the extracted characters obtained through this module will serve as clean inputs for diffusion-based restoration and layout reconstruction, forming an integrated framework for digital preservation of historical manuscripts.

The remainder of this paper is organized as follows. Section 2 reviews related work on text detection and document analysis. Section 3 describes the dataset and annotation process. Section 4

presents the proposed tiled inference method using YOLOX. Section 5 discusses the experimental results, and Section 6 concludes this paper with future perspectives.

## 2 Related Work

### 2.1 AI Applications in Cultural Heritage

Recent advances in artificial intelligence (AI) have profoundly impacted the field of cultural heritage conservation and documentation. In particular, AI-based image recognition, pattern analysis, and predictive modelling provide new avenues for digitising, analyzing, and preserving heritage assets such as manuscripts, artworks, monuments, and archaeological sites. For example, automated scanning and 3D modelling enable the creation of accurate digital replicas of fragile artefacts; machine learning algorithms identify damage or degradation, and environmental sensors combined with AI analytics forecast risks such as changes in temperature or humidity [5, 6]. These developments facilitate efficient and scalable heritage-management workflows, but they also prompt salient concerns about data quality, ethics, and long-term access to digital heritage. As one recent review highlights, the integration of AI into cultural heritage "raises intricate ethical questions" including authenticity, bias, and authorship issues [7, 8]. In the context of manuscript analysis, AI techniques are playing a critical role in revealing and interpreting textual, pictorial, and material features of historical documents [9, 10, 11]. By enhancing both scholarly research and preservation practice, these methods contribute to a deeper understanding and more resilient archiving of cultural heritage.

### 2.2 Text Detection and Layout Analysis for Document Images

Text detection and layout analysis remain core tasks in document image processing, especially for historical manuscripts and archival records. Traditional methods employed connected-component analysis, projection profiling, or geometric heuristics to segment text lines and blocks; however, modern deep learning approaches have demonstrated superior robustness against noise, deformation, degradation, and variant layouts [12, 13]. Object detection frameworks are increasingly applied to locate text regions, while semantic segmentation and layout parsers interpret structural elements such as columns, marginalia, annotations, and decorative features [14, 15]. Special challenges arise in historical documents: non-standard scripts, faded ink, physical damage (e.g., insect holes, stains), and complex page structures. For instance, a recent survey on historical-document datasets emphasises that the variety of scripts, degradation types, and layouts makes layout analysis particularly demanding in this domain [16]. Deep learning–based layout analysis systems support not only segmentation but also logical structure understanding, enabling improved OCR and paleographic interpretation [17]. In addition, large-scale annotated datasets such as M6Doc and BaDLAD have been released to advance multi-format, multi-layout layout-analysis research [18, 19]. Advances in layout analysis thus facilitate downstream tasks such as transcription, table and figure extraction, and faithful digitisation of historical cultural heritage texts.

### 2.3 Tiled Inference and High-Resolution Object Detection

High-resolution images present unique challenges for object detection, particularly when target instances are small relative to the image size or when fine detail is required (e.g., manuscripts or heritage images). One effective approach is tiled inference (or image tiling), where the large image is subdivided into smaller overlapping tiles, each processed by a detector and then merged via non-maximum suppression or more advanced post-processing [20]. Tiling preserves high spatial resolution and ensures complete coverage while mitigating GPU/memory constraints, making it especially suitable for large document or cultural-heritage images. Recent studies show that tiling paired with careful overlapping strategies, central-region filtering and Soft-NMS improves detection recall for small or densely packed objects [21, 22]. Meanwhile, emerging research explores more advanced merging strategies (e.g., group-evidence clustering) to handle detection fragmentation across tile boundaries [23]. Comprehensive surveys on small-object detection highlight that boosting input resolution, scale-aware training, and context fusion are critical for fine-grained detection tasks [24, 25, 26]. Moreover, tiled-training strategies (where tiles are used during training as well as inference) have been shown to improve feature resolution for tiny objects [27]. These developments make tiled inference particularly relevant for high-density manuscript pages, large maps, or historical documents where small character or degradation regions require fine-grained localisation and high recall.

## 3 Dataset and Annotation

### 3.1 Dataset Overview

In this study, we use 44 historical manuscript documents provided by the Humanities Data Repository.[1] The documents include three types of scanned images: RGB (color), B/W (binary monochrome), and grayscale. To ensure a fair evaluation across diverse manuscripts, the documents are split into training, validation, and test sets with a ratio of 7:1:2 without overlapping document IDs. Table 1 summarizes the distribution.

**Table 1:** Distribution of manuscripts and images in the dataset.

| Subset | RGB | B/W | Gray | Total |
|---|---|---|---|---|
| Train | 20 | 3 | 1 | 24 (5,113 images) |
| Validation | 8 | 1 | 0 | 9 (790 images) |
| Test | 8 | 2 | 1 | 11 (1,404 images) |

[1]Formerly provided through the Center for Open Data in the Humanities (CODH).

### 3.2 Annotation Policy

The original annotations contain bounding boxes only for the *main body text*, excluding red-ink characters and marginal notes. Fig. 1 illustrates the annotation categories: (i) main text (used in this study), (ii) red-ink characters (*shubun*), and (iii) marginal notes. Red-ink characters and annotations were excluded because they are not targets of downstream restoration tasks and may introduce noise in character-detection training.

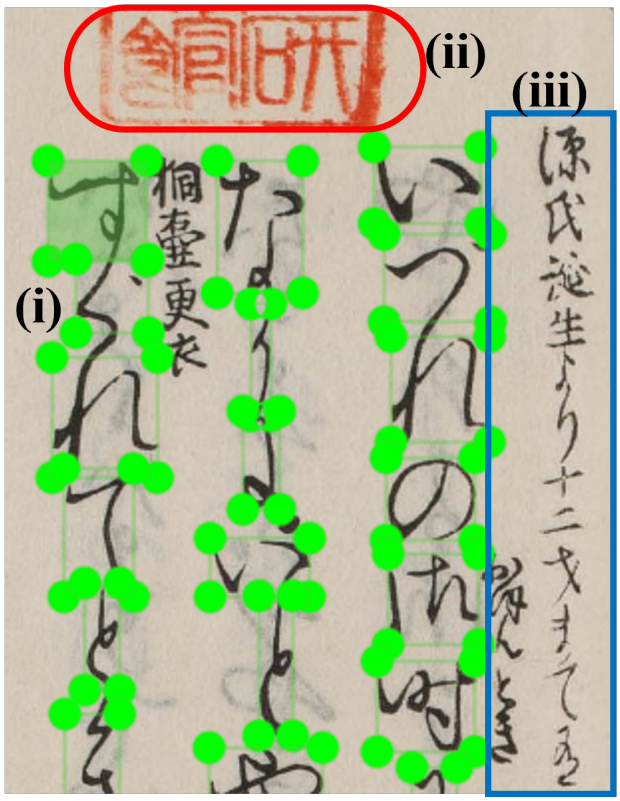



**Fig. 1.** Examples of annotation categories. Only main body text (i) is used for training and evaluation. Red-ink characters (ii) and marginal notes (iii) are excluded.

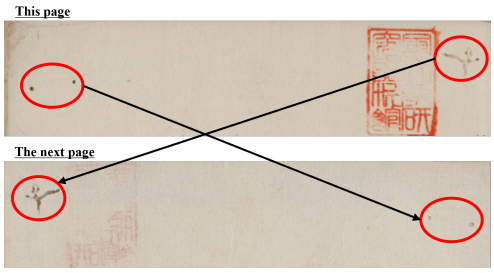
### 3.3 Insect-Damage Class Definition

In addition to the "character" class, we introduce an *insect damage* class to represent physical deterioration commonly observed in ancient manuscripts, such as surface scraping and boring holes. This class helps YOLOX distinguish deterioration artifacts from text, reducing false positives in dense regions and supporting downstream document restoration workflows.

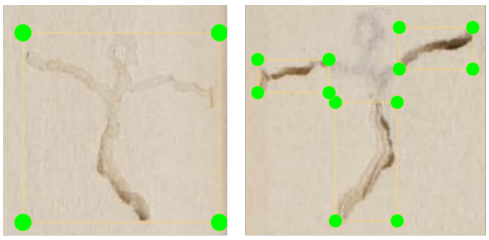
Two major types of insect damage were annotated (Fig. 2a):

1. **Y-shaped scraping:** Caused by silverfish, producing branched erosion traces.
2. **Circular perforation:** Caused by wood-boring beetles, creating round holes or tunnels.

Annotation was performed using the *LabelImg* tool, drawing rectangular bounding boxes that fully cover each missing region. If the same damage shape appears in the same position across consecutive pages, it is considered a continuous physical penetration and annotated as a single instance. This rule prevents the model from learning artificially fragmented shapes that do not correspond to the underlying physical structure.




**(a)** Examples of insect damage observed across consecutive pages.




**(b)** Annotation rule for split insect-damage regions.

If the continuous damage region splits into multiple segments on subsequent pages, each segment is annotated independently (Fig. 2b). The same annotation rules apply to RGB, B/W, and grayscale manuscripts.

### 3.4 Summary

Through this policy, the dataset contains two primary detection targets: (1) main text characters (2) insect-damage regions representing physical deterioration. This multi-class design enables YOLOX to distinguish characters from degradation artifacts, improves detection robustness in degraded areas, and provides reliable inputs for downstream restoration tasks.

## 4 Method

This section describes the proposed high-precision character detection pipeline based on YOLOX-x and tiled high-resolution inference. The method is designed to address challenges specific to historical manuscripts, including dense small characters, edge degradation, and fine stroke preservation.

### 4.1 YOLOX-Based Detection Framework

#### 4.1.1 Model Architecture and Rationale

We adopt **YOLOX-x**, the highest-capacity model in the YOLOX family, as its deeper backbone and larger feature representation provide superior recall for dense and low-contrast handwritten characters. Such high-capacity models are particularly effective for historical manuscripts where fine-grained details are essential for accurate recognition.

#### 4.1.2 Training Data Processing and Augmentation

To preserve the geometric structure of characters, we disable strong spatial-mixing augmentations and adopt only light geometric and color augmentations suited for historical documents:

- Mosaic / MixUp: disabled ($p = 0$) to avoid spatial distortion of character strokes.
- HSV color jitter: enabled ($p = 1.0$).
- Horizontal flip: probability 0.5.
- Small geometric jitters: rotation $\pm 10^{\circ}$, translation $\pm 10\%$, shear $\pm 2^{\circ}$.
- Images are resized and cropped to a fixed input size of $640 \times 640$.

These augmentations preserve the readability of fine strokes while improving generalization.

#### 4.1.3 Optimization and Training Schedule

The model is trained for 300 epochs with a 5-epoch warmup (warmup LR $= 0$; minimum LR ratio $= 0.05$). The base learning rate per image is $0.01/64$ and is scaled by the effective batch size. Optimization uses SGD with momentum 0.9, weight decay $5 \times 10^{-4}$, EMA tracking, and the `yoloxwarmcos` cosine schedule. Following the YOLOX convention, heavy augmentations are disabled for the final 15 epochs (`no_aug_epochs = 15`). Validation is performed every 10 epochs.

#### 4.1.4 Loss Function

We employ the unmodified YOLOX loss formulation, consisting of localization, objectness, classification, and coordinate refinement components. Each loss term is normalized by the number of positive foreground anchors, following the original implementation.

#### 4.1.5 Inference Settings

The evaluation uses a fixed test size of $640 \times 640$, a confidence threshold of 0.01, and a hard NMS threshold of 0.65. No Soft-NMS is applied at evaluation or merging time.

### 4.2 Tiled High-Resolution Inference

Historical manuscripts often contain thousands of small characters per page, and the resolution reduction induced by resizing can significantly degrade detection performance. Therefore, we propose a tiled inference strategy that preserves local resolution while providing full-page coverage.

#### 4.2.1 Pre-padding

Each manuscript page is padded by $p = 320$ pixels on all sides to avoid losing characters located near page borders during tiling. Padding is performed with zero values.

#### 4.2.2 Tiling and Per-Tile Detection

Each padded page is divided into overlapping square tiles of size $T = 640$, with stride $S = 320$ (50% overlap). For each tile, YOLOX predicts bounding boxes for the two target classes (character, insect). Class-specific confidence thresholds are used (e.g., $\tau_{\text{char}} = 0.25$, $\tau_{\text{insect}} = 0.10$).

#### 4.2.3 Central-Region Acceptance

To avoid duplicate detections caused by overlapping tiles, a bounding box is accepted only if its top-left coordinate lies within the **central 50% region** of the tile. This choice is motivated by the following rationale:

- With stride $S = T/2$, every pixel on the page is covered by at least one tile's central region (complete coverage).
- Objects near boundaries appear in multiple tiles, but only one tile accepts them.
- This prevents redundant predictions while preserving recall.

Fig. 3 illustrates the acceptance region.

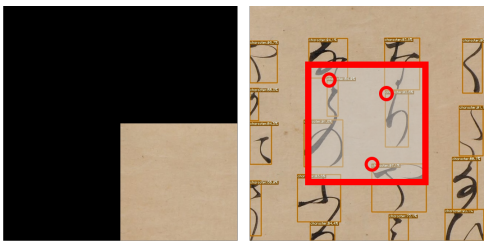



**Fig. 3.** Central-region filtering. Only boxes whose top-left coordinates fall inside the central 50% region (red square) are accepted for merging. This suppresses duplicate detections across overlapping tiles.

#### 4.2.4 Merging and Post-processing

Tile-level coordinates are mapped back to the original page coordinate system and merged using standard **hard NMS** with IoU threshold 0.5. Soft-NMS was evaluated but did not improve performance for densely packed characters, so it is not used in the final system.

After merging, we apply lightweight post-processing to remove tiny boxes, extreme aspect ratios, and detections outside the original page extent.

#### 4.2.5 Parameter Selection and Design Rationale

The parameters $(T = 640, S = 320, p = 320)$ are selected to minimize computational cost while guaranteeing complete page coverage:

- $S = T/2$ is the largest stride that still ensures every page pixel falls within at least one tile's central region.
- Padding $p = S$ ensures full coverage of border regions.
- Central-region acceptance limits each object to a single tile, preventing over-suppression during NMS.

### 4.3 Algorithm Summary

**Algorithm 1** Tiled inference pipeline for high-precision character detection

**Input:** Document with $N$ pages
**Output:** Final detection set $\mathscr{B}$ for each page
**for** each page **do**
  $P \leftarrow \text{pad}(page, p = 320)$
  $\mathscr{B} \leftarrow \emptyset$
  **for** each window $(i, j)$ in sliding_windows$(P)$ **do**
    $tile \leftarrow \text{crop}(P, i, j, 640, 640)$
    $\hat{\mathscr{B}} \leftarrow \text{YOLOX}(tile)$
    **for** each box $b$ in $\hat{\mathscr{B}}$ **do**
      **if** top-left of $b$ is inside the central region **then**
        $\mathscr{B} \leftarrow \mathscr{B} \cup \text{to\_page}(b, i, j)$
      **end if**
    **end for**
  **end for**
  $\mathscr{B} \leftarrow \text{NMS}(\mathscr{B}, 0.5)$
  $\mathscr{B} \leftarrow \text{post\_filter}(\mathscr{B})$
  **return** $\mathscr{B}$
**end for**

**Algorithm Description.** Each page is padded and divided into overlapping $640 \times 640$ tiles. YOLOX performs detection on each tile independently. Detections are accepted only if they lie within the central region of each tile, ensuring complete coverage with minimal redundancy. Tile-level predictions are converted back to page coordinates and merged with hard NMS. Final post-processing removes implausible boxes. This pipeline produces accurate and high-recall detection results for dense manuscript pages.

## 5 Experimental Results and Discussion

**Table 2:** Ablation study of the proposed pipeline conducted on 11 historical manuscript pages.

| Method | Precision | Recall | AP |
|---|---|---|---|
| Baseline YOLOX (page) | 0.948 | 0.798 | 0.720 |
| Tile Inference (no center) | 0.624 | 0.984 | 0.867 |
| Tile + Center (proposed) | **0.976** | **0.989** | **0.982** |

**Quantitative Results** The detection performance of different inference methods is summarized in Table 2. Overall, the proposed divided-tile inference method ("Tile + Center") achieved an average detection accuracy of **98.7%** across 11 manuscript pages, significantly outperforming the baseline YOLOX approach (B1–B3). In the ablation study, our full method also showed higher precision and recall than both the non-tiled baseline and a tiled variant without center merging, leading to a substantially higher AP (0.982 vs. 0.720). While the conventional YOLOX model achieved high accuracy on some pages, its performance dropped below 50% on others (e.g., manuscript IDs 200017458 and 200021063) that contained many small characters. Resizing entire pages to $640\times640$ pixels for training and inference in the YOLOX baseline caused a considerable loss of fine detail, degrading detection accuracy on densely written or finely detailed manuscripts. In contrast, the proposed tiling strategy preserved local image resolution and maintained consistently high accuracy across all tested manuscripts.

**Qualitative Analysis** Figure 4 illustrates example detection outputs on a representative manuscript page (ID 200003803), comparing our method with the conventional YOLOX approach. The left image in Fig. 4 shows the result of our **divided-tile inference method**, where each page is split into overlapping tiles to retain fine character details. The right image shows the output of the **conventional YOLOX method**, which processes the full page resized to $640\times640$. As highlighted by the colored boxes in Fig. 4, our method successfully detects the vast majority of character regions (green boxes denote ground truth and red boxes indicate correctly detected characters), whereas the non-tiled YOLOX baseline misses many small characters or merges closely spaced characters (blue boxes mark missed detections or confused regions) due to the resolution loss from downsampling.

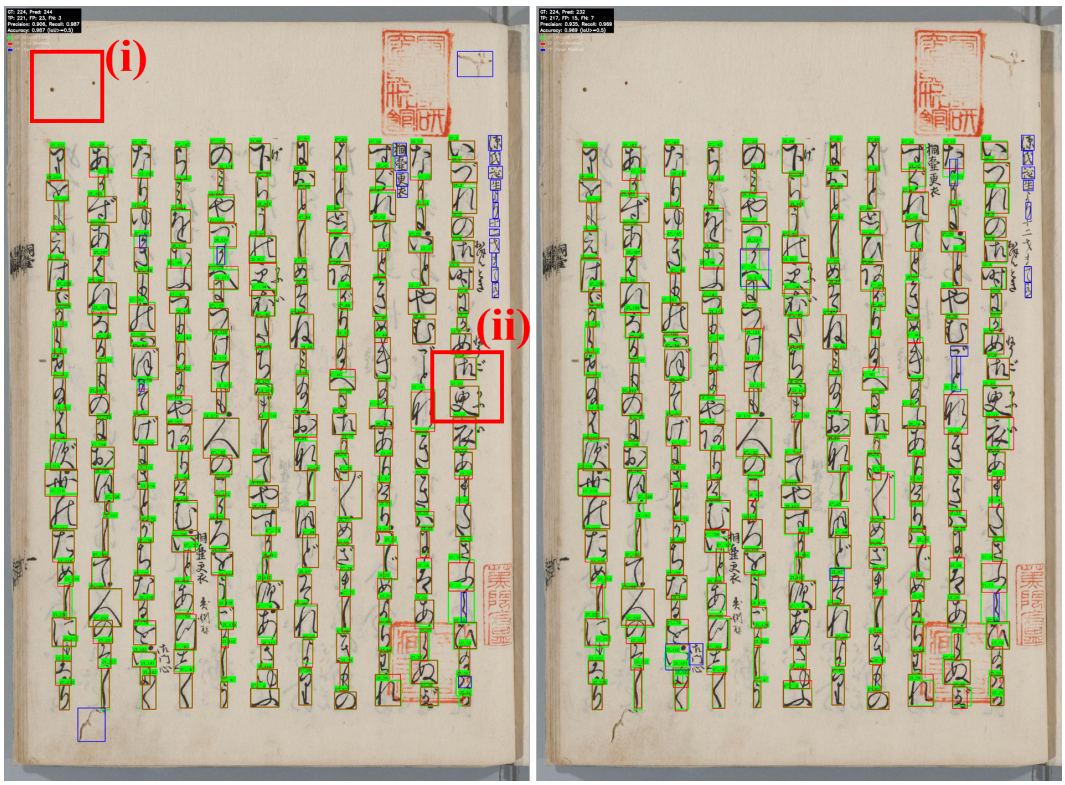



**Fig. 4.** Detection results for manuscript ID 200003803. Green boxes indicate ground truth (GT), red boxes denote correctly detected predictions, and blue boxes represent missed detections (including annotation characters). Two typical failure cases are highlighted: (i) circular insect damage mistaken for text, and (ii) annotation characters located near the main text region, which are not included in the ground-truth annotations.

The few detection errors by our method were mainly caused by two factors: (i) circular insect damage holes that were mistakenly identified as characters, as shown in Fig. 4(i), and (ii) annotation or red-ink marginalia located near the main text, which were not included in the ground-truth annotations (Fig. 4(ii)). This example underscores the effectiveness of the proposed tiling strategy in maintaining small-scale feature visibility, as well as its remaining challenges with certain types of noise.

**Error Discussion** Circular insect damage often exhibits textures and brightness patterns similar to those of paper stains or punctuation marks, making such holes difficult to distinguish from actual characters. This similarity is one of the main causes of the residual false positives observed in our results. A potential countermeasure is to train the detector explicitly to recognize these “false alarm” patterns — for example, by augmenting the training data with challenging background artifacts

(stains, holes, etc.) and teaching the model that these should not be detected as characters. In future work, such targeted retraining with difficult examples could help the model avoid mistaking noise for text. Regarding the missed annotation characters (e.g., editorial notes or rubricated text), detection was unstable because these elements were not labeled in the training data. To address this, an additional class for annotations could be introduced during the annotation process, or a two-stage detection pipeline could be employed (first detecting main text, then handling secondary markings separately). Either approach would likely improve the model's ability to discriminate primary text from other extraneous markings.

**Efficiency Considerations** Finally, although the tiled inference method greatly improves detection accuracy for small characters, it does come with increased computational cost. The overlapping tile approach means more forward passes and some redundant processing at tile boundaries. For practical or real-time applications, further optimizations — such as adjusting the tile stride, reducing test-time augmentation overlaps, or leveraging batch inference — could be explored to speed up the pipeline without sacrificing accuracy.

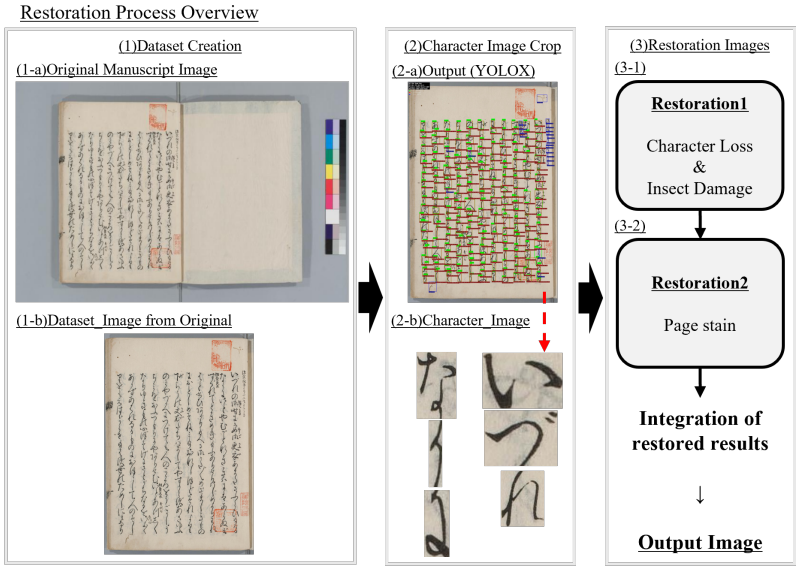



**Fig. 5. Restoration pipeline.** The YOLOX detector extracts character and damage regions, which serve as input for restoration models that fill missing text or repair degradation. Final outputs integrate both stages.

## 6 Conclusion and Future Work

This study presented a high-precision character detection pipeline for historical Japanese manuscripts using YOLOX-x and tiled inference. The method achieved 0.982 AP across diverse pages, effec-

tively separating main text from annotations and damage while demonstrating robustness against the high character density and low-contrast strokes typical of such documents.

The detector is integrated into a broader restoration system (Fig. 5) to reconstruct deteriorated areas like missing characters and stains. This effort is crucial for preserving cultural heritage, ensuring that historical documents remain digitally accessible and legible to future generations despite physical decay.

To enhance flexibility, we plan to introduce separate classification modules for characters and degradation types. Since historical texts contain over 3,000 distinct *kuzushiji* characters, a modular approach enables more accurate restoration tailored to specific damage patterns, thereby improving the overall fidelity of digital reconstructions.

Restoration quality will be evaluated using image metrics (PSNR, SSIM, LPIPS) alongside comprehensive OCR accuracy and human readability to ensure practical usability in real-world archival workflows.

Finally, hybrid CNN-ViT models may be explored to balance high-resolution local feature extraction with global semantic understanding of the entire page layout. Ultimately, this research seeks to bridge the gap between computer vision and traditional philology, providing intelligent tools that enhance the interpretability of our shared historical record.

## References


[1] Yoshizu Y, Kaneko H, Ishibashi R, Meng L. Evaluation of Missing Image Restoration with a Binary Character Image Diffusion Model. In: Proceedings of ICAMechS; 2024. .

[2] Kaneko H, Ishibashi R, Meng L. Deteriorated Characters Restoration for Early Japanese Books using Enhanced CycleGAN. Heritage. 2023;6(5):4345-61.

[3] Lyu B, Yue X, Meng L. Early Japanese Books Organization and Spatiotemporal Database System Creation for Natural Disaster Analysis. Heritage Science. 2024;6(14):1-20.

[4] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO Series in 2021. arXiv preprint arXiv:210708430. 2021.

[5] Ghaith F, Al-Sayyed R. AI Integration in Cultural Heritage Conservation: A Review. Heritage Science. 2023.

[6] Vavoula G, Dallas C. Artificial Intelligence in Heritage Studies: Opportunities and Challenges. Digital Applications in Archaeology and Cultural Heritage. 2023.

[7] Tiribelli G, Baraldi A. AI Integration in Cultural Heritage: Ethical Considerations and Future Directions. Heritage. 2024.

[8] Marinos G, Karydis T. Ethical Aspects of AI in Cultural Heritage Preservation. Journal of Cultural Heritage. 2024.

[9] Stiglec I. AI in the Context of Cultural Heritage; 2023. `https://teachwitheuropeana.eun.org/updates/ai-in-the-context-of-cultural-heritage/`.

[10] Fiorucci M, et al. Machine Learning for Cultural Heritage: Challenges and Perspectives. Journal of Cultural Heritage. 2020.

[11] Stoean C, et al. Deep Learning in Digital Restoration of Cultural Heritage: A Comprehensive Review. Pattern Recognition Letters. 2024.

[12] Lombardi F, Marinai S. Deep Learning for Historical Document Analysis and Recognition: A Survey. Journal of Imaging. 2020;6(8):1-27.

[13] Chen K, Liu Y. CNN-Based Layout Detection for Historical Documents. In: Proceedings of ICDAR; 2019. .

[14] Bhatt J, Sathe S. A Survey of Graphical Page Object Detection with Deep Learning. Applied Sciences. 2021.

[15] Hirata N, Sugimura R, Nishida K. Layout Analysis of Historical Document Images using Deep Learning Methods. In: CVPR Workshops; 2023. .

[16] Nikolaidou K, Seuret M, Liwicki M. A Survey of Historical Document Image Datasets. International Journal on Document Analysis and Recognition (IJDAR). 2022.

[17] Clausner C, Antonacopoulos A. Logical Layout Understanding for Historical Documents. Pattern Recognition. 2024.

[18] Cheng H, Yang Z, Liu Y. M6Doc: A Large-Scale Dataset for Multi-Layout Document Analysis. CVPR. 2023.

[19] Subramani A, et al. BaDLAD: A Benchmark Dataset for Layout Analysis in Historical Documents. Neural Computing and Applications. 2023.

[20] Unel FO, Ozkalayci BO, Cigla C. The Power of Tiling for Small Object Detection. In: CVPR Workshops; 2019. .

[21] Nguyen ST, Tulabandhula T. Dynamic Tiling for Efficient Small Object Detection. arXiv preprint arXiv:230911069. 2023.

[22] Wang I. EdgeDuet: Tiling Small Object Detection for Edge-Assisted Vision. In: IEEE INFOCOM; 2021. .

[23] Xiao Y. Group Evidence Matters: Tiling-Based Semantic Gating for Dense Object Detection. arXiv preprint arXiv:250910779. 2025.

[24] Feng Q, Xu X, Wang Z. Deep Learning-based Small Object Detection: A Survey. Mathematical Biosciences and Engineering. 2023;20(4):6551-90.

[25] Zhou X. A Review of Small Object Detection Based on Deep Learning. Knowledge-Based Systems. 2024.

[26] Hua W, Zhang Y, Liu L. A Survey of Small Object Detection Techniques Using Deep Learning. Artificial Intelligence Review. 2025.

[27] MathWorks. Detect Small Objects Using Tiled Training of YOLOX Network; 2024. https://www.mathworks.com/help/vision/ug/detect-small-objects-using-tiled-training-yolox.html.