Calli-Tongji: A Dataset of Historical Calligraphy Styles

Yan Liu, Yinsheng Yao, Chen Ye, Youzhong Zhi
Tongji University, Yiguan Shufa

Email: y_an@tongji.edu.cn

01. VISION & MISSION

In the era of AI + Humanities, we are moving digital preservation from simple "static recording" to "semantic understanding." This project constructs a standardized, large-scale calligraphy style dataset, providing the foundational support needed for advanced deep learning models in learning micro-level representations of renowned masters' brushwork, structure, and composition and the intelligent preservation and revitalization of Chinese aesthetics.

02. THE CHALLENGE VS. OUR SOLUTION

- **Current Data Limitations**
  - **Raw Rubbings:** Suffer from noise, black-and-white inversion, and erosion.
  - **Standard Digital Fonts:** Lack the original texture, ink variation (flying white/fading), and the artistic "spirit" of hand-written strokes.

- **Our Solution**
  - **High-Fidelity Authenticity:** Sourced from authoritative stele rubbings and authentic works.
  - **Preserved Nuance:** Perfectly retains ink flow, stroke dryness/wetness, and stone texture.
  - **Comprehensive Scale:** Covers 5 Major Scripts (Seal, Clerical, Regular, Running and Cursive) and 400+ Masters (e.g., Wang Xizhi, Yan Zhenqing) across dynasties.
  - **Fine-Grained System:** A structured "Author-Style" taxonomy.

Figure 1. Schematic diagram of the dataset.

03. TECHNICAL INNOVATION

Adaptive Automated Data Production Pipeline
We developed a proprietary pipeline transforming raw scans into high-quality datasets using **Visual Feature Classification**:

1. **Intelligent Sorting (SVM-Based):**
   Extracts features (Edge density, LBP, Stroke width) to automatically classify images as "Stone Inscription" or "Ink on Paper."
2. **Adaptive Normalization:**
   - Inscriptions: Color inversion to standard "Black text on White background."
   - Ink: Otsu's method for binarization to remove yellowed paper backgrounds.
3. **Morphological Denoising (CCA):**
   Uses **Connected Component Analysis** to remove isolated noise (salt-and-pepper) and mold spots, ensuring a high signal-to-noise ratio.
4. **Dual Verification:**
   "AI Pre-recognition + Manual Review" ensures minimal artifacts and high integrity.

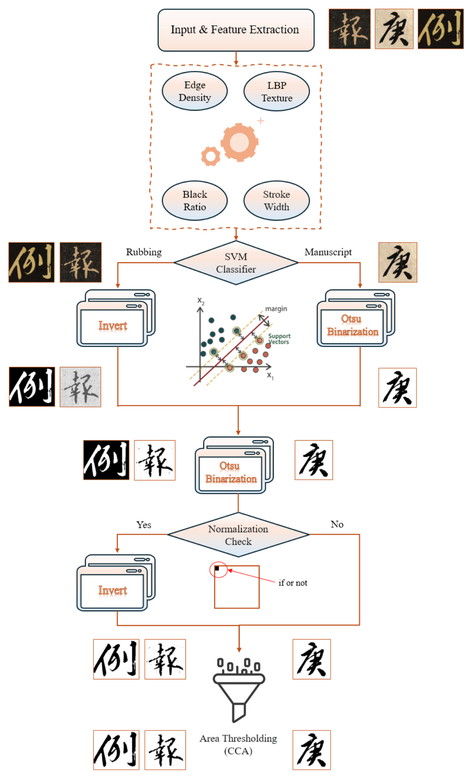
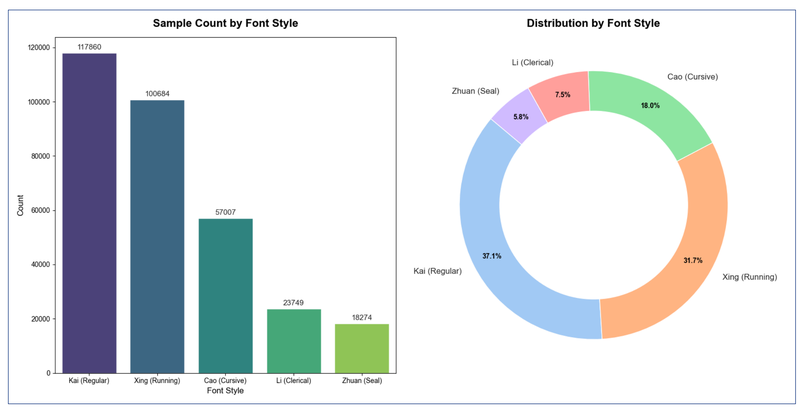


Figure 2. Data processing flowchart.

04. DATASET OVERVIEW

Modalities:
Image: High-Fidelity Binary PNGs (Preserving Edge Details).
Text: Structured Metadata (Unicode, Author, Style, Dynasty).

Scale (Post-Processing):
Total Calligraphers: 310
Font Styles: 5 **Author-Style Pairs**: 427
Single-Character Samples: 317574



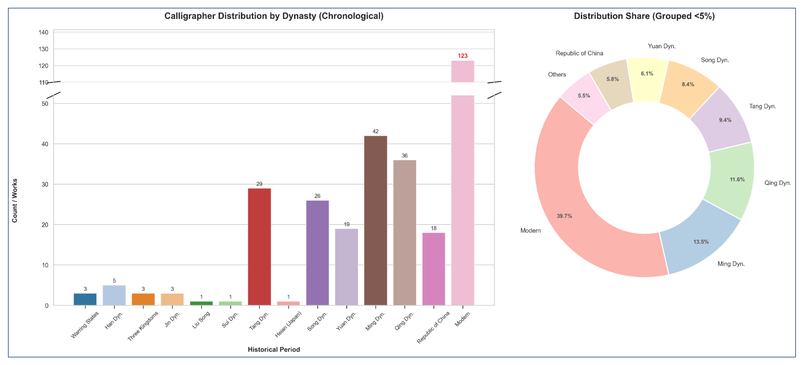


Figure 3. The distribution of fonts and dynasties in the dataset.

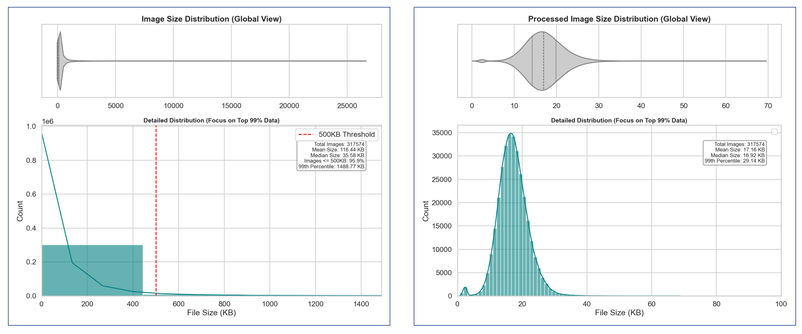


Figure 4. The size distribution of images before and after processing.

05. CORE APPLICATIONS

➢ **Achievements in Digital Humanities Research and Applications**

Fine-Grained Style Recognition
Serves as a benchmark corpus for training encoders to distinguish subtle stylistic differences, setting new standards in Digital Humanities research. According to the benchmark evaluation results shown in the appendix, dedicated encoders trained on this dataset set new performance standards for fine-grained style classification methods, establishing a new benchmark for intelligent research in the field of digital humanities.

Table 1. Model performance comparison under strong enhancement strategy.

Model	Accuracy	Precision	Recall	F1-Score
ResNet-50	0.7380	0.7391	0.7443	0.7394
DenseNet-121	0.7710	0.7739	0.7759	0.7732
EfficientNet-B0	0.7767	0.7787	0.7811	0.7783
ConvNeXt-Tiny	**0.8256**	**0.8238**	**0.8273**	**0.8239**
ViT-Small	0.7944	0.7928	0.7965	0.7932
Swin-Tiny	0.8104	0.8081	0.8120	0.8087

Table 2. Model performance comparison under simple enhancement strategy.

Model	Accuracy	Precision	Recall	F1-Score
ResNet-50	0.7325	0.7395	0.7413	0.7386
DenseNet-121	0.7526	0.7541	0.7576	0.7543
EfficientNet-B0	0.7612	0.7633	0.7667	0.7636
ConvNeXt-Tiny	**0.8100**	**0.8106**	**0.8125**	**0.8103**
ViT-Small	0.7732	0.7718	0.7770	0.7695
Swin-Tiny	0.7944	0.7934	0.7972	0.7939

➢ **Core Applications in the Future**

1. Generative AI (Brush Intent Reproduction)
Supports Diffusion Models in learning micro-level habits (start, move, end strokes). Enables AI to leap from "imitating shape" to "reproducing artistic intent."

2. Intelligent Restoration
Overcomes physical damage (weathering/erosion). The model infers missing strokes based on the author's specific logic, offering scientifically backed "virtual restoration" for cultural relics.

06. ACCESS & ECOSYSTEM

➢ **Sustainable & Expandable**
- **One-Click Preprocessing:** New data can be automatically cleaned using the pipeline.
- **Future Roadmap:** Integrating phonetic and semantic info (Shuowen Jiezi) and biographies.

➢ **Balancing Academic Accessibility and Copyright Protection:**
- **Open Access:** Representative subsets of mainstream styles (e.g., "Wang Xizhi - Running Script") are available as standard benchmarks (limit: 100 samples per category).
- **Full Access:** Available for institutions via real-name application and deep collaboration.