Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

Quality Control and Interannotator Agreement in Manual Annotation of Damage in Chinese Paintings and Calligraphy

Chao Deng, Fei Chen

Industry Yibin Polytechnic College, Yibin, Sichuan, China

Abstract: Manual annotation of damage in images of Chinese paintings and calligraphy involves locating damaged areas, assigning damage categories, and delineating their boundaries. Annotators may disagree over the category or extent of the same feature, affecting both AI training data and documentation made before conservation. Drawing on standards for paper-based cultural heritage and image-annotation practice, this methods paper proposes a workflow comprising common annotation rules, independent annotation by two people, discrepancy detection, expert adjudication, and retention of revision records. Using labelme as an example, it describes how to retain two independent annotation files, check for omissions and incorrect categories, and assess agreement with category agreement, intersection over union (IoU), and the Dice coefficient. Five working labels, four operational tables, and three illustrative figures support implementation. The figures and numerical examples are explanatory and are not measured outcomes from collection objects.

Keywords: Chinese Paintings and Calligraphy; Damage Annotation; Quality Control; Independent Double Annotation; Labelme.

1. Introduction

After imaging a historical painting or calligraphic work, annotators must decide where damage occurs and which category it belongs to. Standards define and illustrate damage to paper-based artifacts, but applying those definitions to the boundaries of features in a particular image still requires operational rules.[1]

A national standard also calls for correspondence between a damage diagram and its written description. Drawing regions on an image alone therefore does not produce a complete damage record suitable for direct use.[2]

Studies on automated damage detection in paintings and calligraphy commonly train models using human-labeled images. Multiple checks and expert review are needed because annotation quality affects the subsequent recognition task.[3] The results may also enter conservation records, which makes it necessary to link each original image, its annotations, and the revision history to the same object.[4]

This article develops a method rather than reporting a measured evaluation of a batch of cultural objects. Here, a label is the name of a damage category; an instance is a feature recorded separately; and a boundary is the region drawn around that feature in an image.

2. Establishing Annotation Rules

2.1. Deciding What to Annotate and What to Defer

Before assigning the formal task, the project lead should specify what each damage category looks like, how shadows, brushwork, and earlier repairs will be handled, and whether very small holes will be recorded individually. Data-annotation guidance likewise calls for the task, rules, and quality requirements to be established before annotation and acceptance checks begin. [5]

Table 1. Operational criteria for five working damage labels

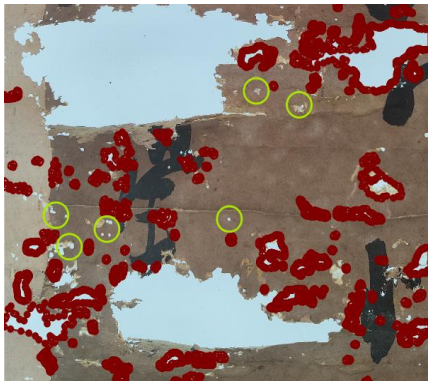
Working label	Image evidence	Common confusion	Recommended action
Material loss	A local part of the paper or silk is missing.	Insect holes, transmitted light, old patches	If the cause is unclear, record loss or mark as uncertain.
Fracture	A distinct split or crack in the support is visible.	Creases, lines in the artwork	Check raking-light images or the object.
Stain	An anomalous colored area unrelated to the original image is present.	Ink bleeding, shadows	Compare original coloration and images under other lighting.
Insect damage	Holes, channels, or visible feeding marks are present.	Pinholes, binding holes, other loss	Do not infer an insect cause without sufficient evidence.
Crease	A fold or ridge results from deformation of paper or silk.	Brushstrokes, seams, raking-light shadows	Check raking-light or surface-topography information.

As shown in Figure 1, a small hole with a characteristic feeding edge may be labeled insect damage. If only a hole is

visible and its cause cannot be determined, the label should remain uncertain until a conservator examines the object or additional raking-light images. Annotators should not force every feature into one of the five labels.



(a) Image before annotation.



(b) Red indicates identifiable insect damage; green circles indicate features whose cause remains uncertain.

Figure 1. An illustrative decision about insect damage in a painting

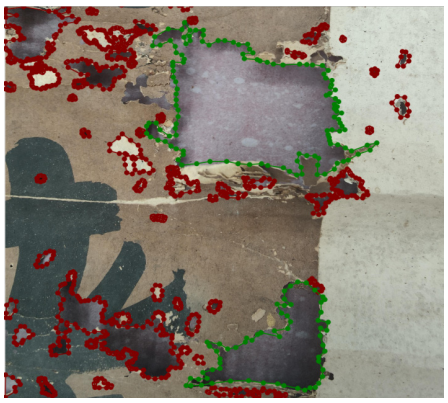
Image provenance, substrate, magnification, lighting conditions, and annotation version should be recorded. Research on other annotated image datasets likewise treats location, completeness, and metadata as dimensions of sample quality in addition to category.[6]

Table 1 presents project-level working labels rather than a verbatim reproduction of the national standard. If one area shows both insect damage and material loss, a primary label and a co-occurring condition may be retained, but their use in statistics and model training must be specified.

2.2. Trial Annotation and Independent Double Annotation

All annotators should first label the same practice images independently. The project lead can then classify disagreements as missed instances, use of different category names, or different interpretations of the boundary. Work on damage mapping at the Mogao Grottoes demonstrates the value of a purpose-built recording scheme, although mural and painting damage categories are not interchangeable.[7] Automated delineation of damaged mural regions likewise illustrates that identifying the boundary is a distinct technical task.[8]

For formal double annotation, annotators A and B use the same original image and rule version, save separate files, and do not inspect one another's results before completing their work. Expert review begins only after both versions have been frozen. A single jointly negotiated mask cannot be used to quantify their original disagreement (Figure 2).



(a) Annotator A calls the large damaged opening material loss



(b) Annotator B calls it insect damage. The example describes a possible adjudication as combined insect damage and material loss; the illustration is not a measured inter-annotator result.

Figure 2. A category discrepancy between two independent annotations

Deep-learning tools may help locate candidate damage, but automatically proposed regions still require human checking.[9] This study evaluates the manual annotation process and does not treat a model suggestion as ground truth. Research on damage extraction and virtual restoration also represents damaged areas as masks. A mask simply marks the pixels inside a selected region, making area measurement and overlap comparisons possible.[10]

3. Independent Annotation and Review in Labelme

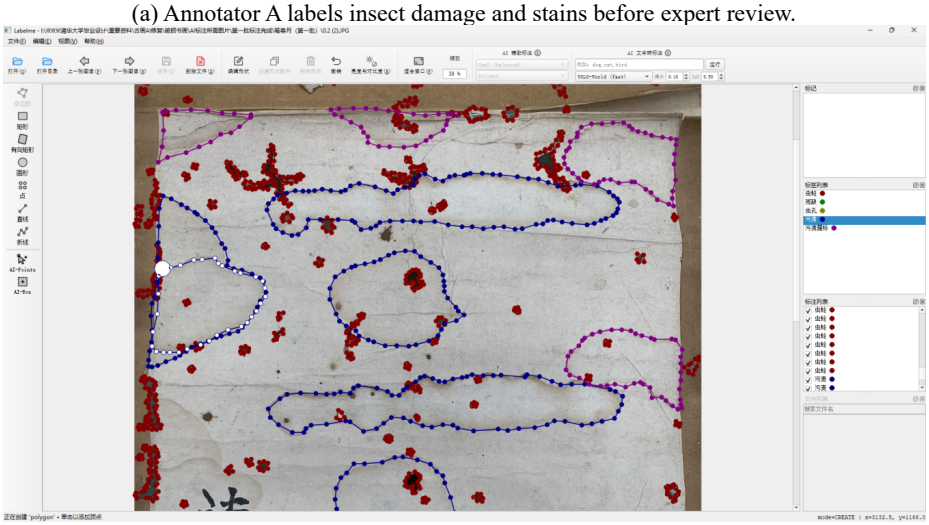
3.1. Two Independent Files for Each Image

The Python-based labelme program supports polygon annotations and stores category names and vertex coordinates in JSON files. JSON is a text-based format that can record image names, labels, and coordinates; annotators need only open the image, draw the region, and save the file, while technical staff can check the data programmatically.[11] The earlier browser-based LabelMe system also organized images and objects in structured annotation records, but it is a different program with a similar name.[12]

A practical filename contains the object identifier, image identifier, annotator identifier, and rule version, for example 001_023_A_v1.json. Annotators' files should be kept in separate folders and the original images stored read-only. Expert corrections should generate new adjudicated files rather than overwrite either original file (Figure 3).



(a) Annotator A labels insect damage and stains before expert review.



(b) The expert records the previously missed stain regions in purple in a separate adjudicated version. The interface shown in the screenshots is localized in Chinese

Figure 3. Annotator A and expert revision in the labelme interface

3.2. Matching Instances before Comparing Annotations

Before comparing A and B, decide whether the regions they marked refer to the same physical instance. Once instances have been matched, compare their assigned damage categories and the extent of their regions. If only one person marked a region, list it as a possible omission rather than assuming the other annotator was wrong; an expert should check it first.

The adjudicator should record the final category, final boundary, and reason for each change, such as "a crack confirmed under raking light" or "the object shows an earlier repair." Such records help annotators interpret the rules and make later auditing possible.

Table 2. Handling disagreements in independently annotated images

Disagreement	What it means	Priority check	Recorded outcome
Possible omission	Only one annotator marked a region.	Very small lesion, oversight, or non-damage feature	Confirm whether the region should be added or deleted.
Category conflict	Both marked the same instance but chose different categories.	Working definitions and evidence from the object	Record final category and supporting evidence.
Boundary conflict	The categories agree but region extents differ.	Magnification, original image, raking-light image	Retain both boundaries and the adjudicated one.
Duplicate	One annotator marked the same feature twice.	Instance identifier and region overlap	Merge and record the reason.
Uncertain	The image alone is inconclusive.	Additional photographs, lighting, or the object	Retain uncertain status.

If the adjudicated records will later be converted into a training format such as COCO, retain the original labelme JSON files before mapping categories and validating coordinates. The COCO dataset established a widely used object-annotation and evaluation format, but conversion does not repair annotation errors. [13]

4. Measuring Annotation Agreement

4.1. Comparing Categories and Boundaries Separately

Category agreement is straightforward: first match the same set of instances, then divide the number assigned the same category by the number compared. If 8 of 10 matched instances have matching categories, agreement is 80%. The values 8 and 10 are an arithmetic example, not observed data from this project.

Table 3. Four essential quality checks

Question	Metric or record	Interpretation	Caution
Do categories agree?	Category agreement; Cohen's κ if needed	Compare labels for matched instances.	Expert review is needed to resolve category errors.
Were instances missed?	Expert-confirmed omissions / eligible instances	Tabulate by category and annotator.	Confirm first that a feature is damage.
Do regions overlap?	IoU and a visual overlay	Measure the overlap of marked regions.	Interpret small objects separately.
Are revisions traceable?	Completeness of adjudication log	Record who changed what, when, and why.	Do not overwrite original annotations.

When category frequencies are strongly imbalanced, Cohen's κ may also be reported because it adjusts observed agreement for agreement expected by chance.[14] A single κ value cannot replace per-category counts, omission lists, or expert judgment.

Intersection over union (IoU) compares the areas marked by A and B. The intersection is the area both marked; the union is the area marked by at least one person. Thus, IoU = intersection area / union area.[15]

The Dice coefficient is another overlap measure: Dice = 2 × intersection area / (area A + area B). If each annotator selects 100 pixels and 80 pixels overlap, IoU = 80 / 120 ≈ 0.667 and Dice = 160 / 200 = 0.800. These numbers explain the calculation; they are not project results.[16]

For a small insect hole, a shift of only a few pixels can markedly change IoU. Small holes and large stains should therefore be reported separately and visual overlays examined alongside summary values. Segmentation metrics should be interpreted in relation to object size, geometry, noise, and the purpose of the evaluation.[17]

For annotations intended for model training, checking omissions and wrong categories may take priority. For a test set, category decisions, boundaries, and a locked version require stricter control. A single agreement threshold is unlikely to be appropriate for every damage-recognition task.[18]

5. Retaining a Traceable Revision History

Usable annotations must answer three questions in addition to showing the final region: who drew it, which rule version was used, and why it was changed. Mixing original and final files hinders later review and increases the risk that images of the same object will appear in both training and test sets.

Research on data cascades in high-stakes AI shows that weaknesses in data collection and documentation can create downstream problems in deployed models. Quality checks and change records should therefore be built into annotation from the outset.[19]

A study of ten widely used benchmark datasets estimated an average of at least 3.3% erroneous test-set labels; its estimates included about 6% in the ImageNet validation set, 0.15% in MNIST, 0.54% in CIFAR-10, and 5.85% in CIFAR-100.[20] Once a test set is released, edit permissions should be restricted and any necessary corrections should produce a documented new version and identify affected results.

Table 4. Minimum records for an annotation task

Record	Recommended fields	Purpose
Original image	Object and image IDs; provenance; permitted use; imaging conditions	Identifies each image and its authorized use.
Independent A and B files	Annotator; category; instance ID; coordinates; rule version	Preserves the two original judgments.
Discrepancy list	Instance ID; disagreement type; both regions; timestamp	Directs expert review to disputed regions.
Expert adjudication	Final category and boundary; evidence; reviewer; date	Explains how the final decision was reached.
Released dataset version	Version; train/validation/test assignment; change summary	Prevents mixing of different versions.

A practical sequence is trial annotation, rule revision, formal independent annotation, expert adjudication, and version release. The initial batch should include paper- and silk-based objects and both simple and complex damage. New annotators may require more intensive review; after the rules stabilize, sampling can be stratified by category and annotator. Sampling rates and acceptance criteria should be specified before a batch begins, not lowered after its results are known.

6. Discussion and Conclusion

No single score captures annotation quality. If category labels match but region boundaries differ substantially, the boundaries need review. If regions match but both annotators mistake an intact feature for damage, a conservator must

examine the object. A tidy final overlay is also insufficient as test-set evidence if its revision history cannot be reconstructed.

The proposed workflow makes independent double annotation and expert review operational. It first defines five working labels and an uncertain category, retains two independent originals, and then checks omissions, category conflicts, and boundary differences. Category agreement and, when appropriate, Cohen's κ quantify disagreement in classification. For localization or segmentation, category agreement should be accompanied by IoU and, where useful, Dice or overlays; pairwise IoU can be averaged across annotators if the object-matching and aggregation rules are specified in advance.

This article does not calculate category agreement, κ, IoU, or Dice for a real batch of historic paintings or calligraphy, and it does not demonstrate that a particular review rate achieves a specified quality level. Future work with authorized images should lock the protocol before evaluation, report the number of instances per category and observed disagreements, document expert revisions and review time, and then assess both the workload and effectiveness of the procedure.

Acknowledgments

This work was supported by the Yibin Industrial Vocational and Technical College research project "Development and Application Demonstration of a Chinese Painting and Calligraphy Extended Depth of Field Scanning Device and Intelligent Damage Recognition System" (XJKY2025-01ZN).

References

[1] National Cultural Heritage Administration. (2010). WW/T 0026—2010, Classification and Illustrations of Deterioration in Museum Collections of Paper-Based Cultural Heritage. (In Chinese).

[2] State Administration for Market Regulation, & Standardization Administration of China. (2024). GB/T 44208—2024, Basic Elements for Description and Illustration of Deterioration in Museum Collections. (In Chinese).

[3] Deng, X., Li, Z., Zhang, Y., et al. (2025). Application of AI-based object detection to damage surveys of painting and calligraphy artifacts. Journal of Northwest University (Natural Science Edition), 55(1), 98–105. https://doi.org/10.16152/j.cnki.xdxbzr.2025-01-008. (In Chinese).

[4] National Cultural Heritage Administration. (2010). WW/T 0027—2010, Specifications for Conservation and Restoration Records of Paper-Based Cultural Heritage in Museum Collections. (In Chinese).

[5] State Administration for Market Regulation, & Standardization Administration of China. (2023). GB/T 42755—2023, Artificial Intelligence: Data Annotation Procedures for Machine Learning. (In Chinese).

[6] Le, P., Liu, R., Shangguan, B., et al. (2023). GeoAI samples: Models, quality, and services. Geomatics and Information Science of Wuhan University, 48(10), 1616–1631. https://doi.org/10.13203/j.whugis20230125. (In Chinese).

[7] Zhang, N., Zhang, Q., Feng, W., et al. (2017). A damage identification system for ancient murals and its application at the Mogao Grottoes. Dunhuang Research, (2), 135–140. (In Chinese).

[8] Deng, X., Yu, Y., Yang, L., et al. (2023). Automatic delineation of damaged areas in Ming-dynasty murals at Zhilin Temple in Jianshui. Journal of Computer-Aided Design and Computer Graphics, 35(10), 1554–1565. https://doi.org/10.3724/SP.J.1089.2023.19738. (In Chinese).

[9] Zhao, H., Yu, Y., Chen, A., et al. (2025). Detection of damage in ancient murals using a residual dual-channel attention U-Net. Journal of Computer-Aided Design and Computer Graphics, 37(6), 1040–1052. https://doi.org/10.3724/SP.J.1089.2023-00430. (In Chinese).

[10] Zhang, Z., Shui, W., Zhou, M., et al. (2021). Damage extraction and virtual restoration algorithms for digital murals. Application Research of Computers, 38(8), 2495–2498, 2504. (In Chinese).

[11] Wada, K. (n.d.). labelme: Image annotation with Python. Retrieved September 25, 2026. https://github.com/wkentaro/labelme.

[12] Russell, B. C., Torralba, A., Murphy, K. P., et al. (2008). LabelMe: A database and web-based tool for image annotation. International Journal of Computer Vision, 77(1–3), 157–173. https://doi.org/10.1007/s11263-007-0090-8.

[13] Lin, T. Y., Maire, M., Belongie, S., et al. (2014). Microsoft COCO: Common objects in context. In Computer Vision—ECCV 2014 (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48.

[14] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104.

[15] Jaccard, P. (1912). The distribution of the flora in the alpine zone. New Phytologist, 11(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[16] Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), 297–302. https://doi.org/10.2307/1932409.

[17] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Medical Imaging, 15, 29. https://doi.org/10.1186/s12880-015-0068-x.

[18] Maier-Hein, L., Reinke, A., Godau, P., et al. (2024). Metrics reloaded: Recommendations for image analysis validation. Nature Methods, 21(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z.

[19] Sambasivan, N., Kapania, S., Highfill, H., et al. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM. https://doi.org/10.1145/3411764.3445518.

[20] Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.