Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

When Background Pixels Hide Damaged Strokes: An Error-Normalization Audit of Oracle-MNIST

Jack Wallace[1] · Eric Clark[2] · Paul Lewis[3]

[1] Latvia · [2] Lithuania · [3] Luxembourg

Original publication: November 7, 2026 | Article 570

Abstract

Erasing the middle of an inscription can leave a modest average image error despite removing much of its visible stroke content. A fixed 12-by-12-pixel mask applied to 3,000 Oracle-MNIST images produced markedly different priorities under global and foreground-normalized squared error: their rank correlation was 0.449, and 1,221 images moved by at least one quartile. Within glyph classes, an additional ten percentage points of foreground increased global error by 0.0186 but reduced foreground-normalized error by 0.0146. These findings concern measurement under an artificial erasure baseline. They support retaining separate whole-image and stroke-focused audit measures before evaluating an inpainting system.

Research draft. This study examines controlled digital damage and a zero-fill baseline; it does not reproduce Orpaint or establish historical readings.

When Background Pixels Hide Damaged Strokes: An Error-Normalization Audit of Oracle-MNIST

Research draft. This study examines controlled digital damage and a zero-fill baseline; it does not reproduce Orpaint or establish historical readings.

Abstract

Erasing the middle of an inscription can leave a modest average image error despite removing much of its visible stroke content. A fixed 12-by-12-pixel mask applied to 3,000 Oracle-MNIST images produced markedly different priorities under global and foreground-normalized squared error: their rank correlation was 0.449, and 1,221 images moved by at least one quartile. Within glyph classes, an additional ten percentage points of foreground increased global error by 0.0186 but reduced foreground-normalized error by 0.0146. These findings concern measurement under an artificial erasure baseline. They support retaining separate whole-image and stroke-focused audit measures before evaluating an inpainting system.

1. A conservation decision hidden inside an average

Restoration systems are often compared through an average discrepancy between an output image and a digital reference. For a rubbing dominated by background, that average answers a different question from the proportion of visible writing that has disappeared. A relatively small central gap may cut through a character's defining intersection, whereas a larger background alteration may affect little interpretable evidence. An evaluation pipeline therefore makes a substantive choice when it decides which pixels receive equal weight. This paper asks whether that choice changes the order in which damaged oracle-character images appear to require attention.

Orpaint addresses zero-shot inpainting of oracle-bone rubbings through a diffusion framework incorporating visual state-space blocks and an efficient scanning mechanism (Meng et al., 2025). Its architectural and efficiency questions make the specification of an evaluation target consequential: faster reconstruction is useful only relative to an adequately described objective. Our contribution sits before model comparison. We examine a deliberately elementary erasure operation so that disagreements between loss definitions cannot be attributed to a learned network, training corpus, or sampling schedule. We neither run Orpaint nor reinterpret the performance claims reported in its letter.

The research question is concrete: do whole-image squared error and foreground-normalized squared error identify the same difficult cases after an identical central erasure? Oracle-MNIST provides a suitable small-image setting because its public test partition contains 3,000 grayscale images distributed equally across ten categories (Wang and Deng, 2024). These processed character crops are not complete rubbing sheets. They permit a reproducible measurement audit while placing a clear boundary on the archaeological conclusions available from the analysis. The original pixels serve as the digital comparison image, rather than as proof of an inscription's historically complete form.

2. Why architecture and measurement should be separated

Several restoration traditions address how plausible content can be inferred from a visible context. Context encoders learn representations through a missing-region task, while partial convolutions explicitly adapt convolutional operations to irregular holes (Pathak et al., 2016; Liu et al., 2018). Neither research direction makes every unweighted pixel discrepancy an adequate record of the damage that matters to a domain specialist. Their relevance here is methodological: the shape and location of absent evidence are part of the task specification, and an evaluator should preserve those details instead of reducing an image to one unexplained score.

Contextual attention and gated convolutions offer different mechanisms for using spatial information during inpainting (Yu et al., 2018a; Yu et al., 2018b). EdgeConnect places additional emphasis on edge structure (Nazeri et al., 2019). These approaches motivate a distinction between recovering an image's broad appearance and preserving particular boundaries or strokes. We do not compare their outputs. Instead, their differing emphases explain why a loss restricted to a visibly defined subset of pixels may reorder cases even when the underlying damaged image never changes. A disagreement between scores is an evaluation finding, not evidence that any one restoration architecture is superior.

Diffusion-based reconstruction creates another reason to keep the measurement problem explicit. Denoising diffusion models establish a generative route from progressively perturbed images, and implicit diffusion models offer a related sampling formulation (Ho et al., 2020; Song et al., 2020). RePaint applies diffusion to inpainting, whereas latent diffusion operates with a learned compressed representation (Lugmayr et al., 2022; Rombach et al., 2021). These works concern reconstruction mechanisms with distinct computational and representational choices. Our baseline has no denoising process at all. Its simplicity isolates the denominator of a reported error from those mechanisms, making the resulting audit relevant to future comparisons without claiming to evaluate a diffusion sampler.

The state-space literature likewise concerns how information is processed, rather than which conservation objective should be selected. Mamba develops selective state-space sequence modeling, VMamba extends state-space ideas to visual representation, and MambaIR examines restoration with such models (Gu and Dao, 2023; Liu et al., 2024; Guo et al., 2024). Large-mask inpainting with Fourier convolutions and the deep image prior provide further examples of architectural diversity (Suvorov et al., 2021; Ulyanov et al., 2017). Taken together, these sources make an argument for a common, inspectable measurement layer. A transparent damage audit can accompany an architecture, but cannot replace an experiment on that architecture.

3. Exploratory propositions and research model

Let foreground density describe the proportion of source-image pixels whose normalized intensity reaches 0.5. With a fixed mask, an image containing more bright stroke material has more opportunity to lose squared intensity. Proposition P1a therefore examines a positive density association with global squared error. Foreground-normalized error divides its selected squared loss by the image's total foreground count. That denominator changes with the explanatory variable, so P1b examines whether normalization materially alters the density relationship. It would be inappropriate to read either association as an independently identified causal effect of adding a historical stroke.

P2 concerns prioritization rather than average accuracy. Each loss orders the same 3,000 images from lower to higher damage. The difference between an image's foreground-loss percentile and its global-loss percentile describes a change in priority. A negative density coefficient in this discrepancy model would mean that denser characters tend to move downward when the evaluator switches to a foreground

denominator. Gradient energy enters this third model to distinguish density from one simple description of local intensity variation. Category indicators account for average differences among the ten supplied labels.

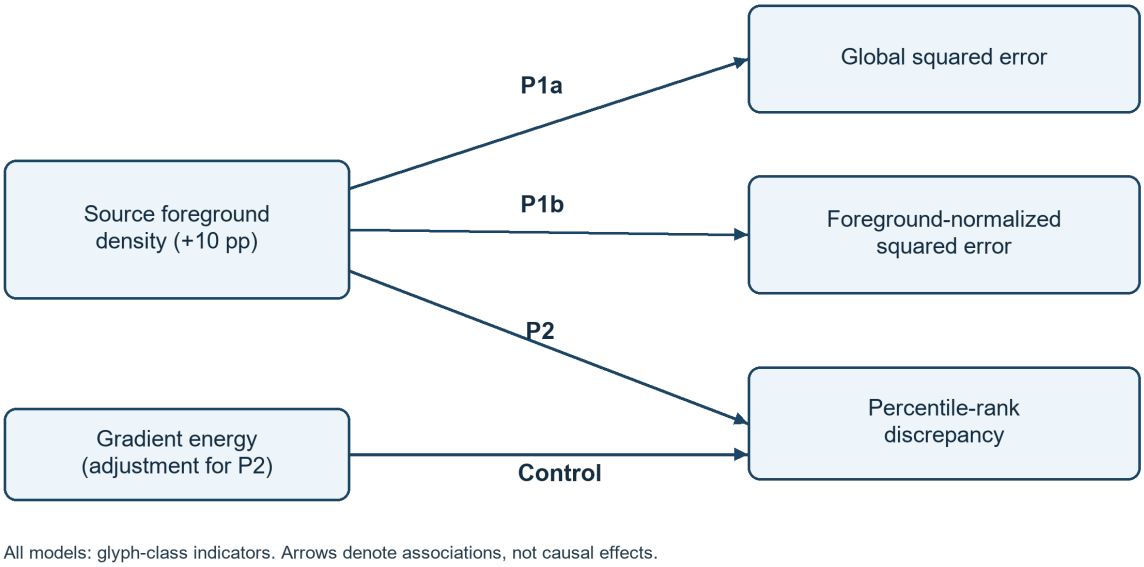

Figure 1. Research model for the exploratory measurement audit. P1a and P1b concern separate error definitions; P2 concerns their percentile-ranking difference. Arrows represent fitted associations. Category indicators enter every model, and gradient energy enters the P2 specification. The diagram makes no claim that category adjustment removes within-category dependence or establishes a historical mechanism.

4. Public images and the controlled erasure

We downloaded the two compressed test-partition files from the public Oracle-MNIST repository through a distribution mirror. Their Git object hashes matched the values returned by the author's repository listing. Decompression yielded 3,000 images of 28 by 28 pixels and 3,000 labels, with 300 observations in each category. Pixel values were divided by 255. No image or label was discarded, and every image contained at least one pixel meeting the foreground threshold. Data provenance, transport hashes, and the derived image-level measurements are retained with the analysis record.

Every image received the same square mask covering rows and columns eight through nineteen under zero-based indexing. This removes 144 of 784 pixels, or 18.37 percent of the image, and fills that area with zero. The operation is an artificial stress condition. It does not estimate the distribution of erosion, burial damage, scanning artifacts, or human intervention in an archaeological collection. Missing-data terminology is useful precisely because the observation mechanism matters; conclusions about one imposed pattern should not be silently generalized to an unknown historical process (Little and Rubin, 2002).

Global mean squared error averages squared differences over all 784 positions. Foreground-normalized error sums those differences only at pixels classified as foreground in the source image, then divides by that source foreground count. The source definition prevents the erased output from deciding which missing strokes qualify for evaluation. At the same time, it includes bright noise as well as writing and does not constitute expert segmentation. A thresholded foreground is a reproducible operational measure whose shortcomings remain visible, rather than an annotation of semantic importance.

The source-image gradient measure is the average sum of squared horizontal and vertical numerical gradients. All variables are computed before regression, with a retained image index connecting each row to its original source position. Percentile ranks use average ranks for ties. We report the number of cases whose absolute rank discrepancy reaches 0.25 as a readily interpretable audit quantity. This cutoff was not optimized to maximize disagreement, and it should not be treated as a validated conservation triage threshold. Dataset documentation remains essential for communicating such construction choices (Gebru et al., 2021).

5. Estimation and uncertainty

Three ordinary least-squares specifications summarize the measurement relationships. The first two regress their respective errors on foreground density scaled in ten-percentage-point units and nine category indicators. The third uses the percentile-rank discrepancy as its outcome and adds gradient energy to the same explanatory structure. The coefficients describe conditional linear associations in this fixed test partition. They are not out-of-sample prediction scores, and their interpretation does not require claiming that the fitted linear equation is the true biological or cultural process producing a character.

The results table reports HC3 standard errors and 95 percent intervals. Heteroskedasticity-consistent inference addresses one feature of unequal residual variability, following the general covariance-estimation rationale associated with White (White, 1980). It does not address arbitrary dependence between crops from a common source or publication. Category indicators remove class means, but are not a substitute for an appropriate sampling model. We retain this distinction because cluster-robust inference also requires a meaningful cluster definition and enough informative clusters, issues emphasized by Cameron and Miller (Cameron and Miller, 2015).

No training/test optimization is undertaken in this measurement audit. If a later study learns thresholds or predicts damage priority, the way evaluation units are separated becomes consequential; validation for structured observations requires a design consistent with their dependence (Roberts et al., 2017). Selecting a measurement rule after inspecting the same images could also create optimistic reported performance, a problem related to model-selection bias in error estimation (Varma and Simon, 2006). Our present table is exploratory and keeps all three specified outcomes. Inferential calculations accompany the estimated magnitudes; they do not turn this exercise into a preregistered hypothesis test or a causal study.

Table 1. Executed image-level regression results. Density coefficients correspond to ten percentage points of additional source foreground. HC3 standard errors appear in parentheses and 95 percent confidence intervals in brackets. All three specifications contain category indicators.

Model / outcome	Term	Estimate (HC3 SE)	95% CI	N; R2
Global pixel error	Foreground density, +10 pp	0.01860 (0.00042)	[0.01777, 0.01943]	3000; 0.621
Foreground-normalized error	Foreground density, +10 pp	-0.01459 (0.00147)	[-0.01748, -0.01171]	3000; 0.143
Ranking discrepancy	Foreground density, +10 pp	-0.24115 (0.00531)	[-0.25156, -0.23074]	3000; 0.726
Ranking discrepancy	Gradient energy	1.17864 (0.41874)	[0.35759, 1.99970]	3000; 0.726

6. The same erasure produces different priorities

Across the 3,000 images, mean global squared error was 0.05221, while mean foreground-normalized squared error was 0.26151. Their different denominators mean that the magnitude of one should not be subtracted from the other as if they shared an identical substantive scale. More informative for prioritization is the Spearman correlation of 0.449 between their case orderings. Exactly 1,221 images, or 40.7 percent of the test set, changed percentile position by at least a quarter. The evaluator's choice therefore affects which cases appear most damaged, even with no reconstruction model in the pipeline.

The class-adjusted density association changed direction across error definitions. Ten additional percentage points of source foreground corresponded to 0.01860 greater global error, with a 95 percent interval from 0.01777 to 0.01943. For foreground-normalized error, the corresponding estimate was minus 0.01459, with an interval from minus 0.01748 to minus 0.01171. These contrasts are consistent with the proposed denominator mechanism. They do not demonstrate that increasing inscription density would improve restoration, because the predictor and both outcomes are constructed from overlapping source-pixel information.

In the discrepancy specification, the density coefficient was minus 0.24115 percentile units per ten percentage points, conditional on category and gradient energy. Its interval extended from minus 0.25156 to minus 0.23074. The gradient-energy coefficient was 1.17864, with an interval of 0.35759 to 1.99970, but a one-unit change in that measure is much larger than a routine pixel-gradient change and should not be interpreted as a typical intervention. The three models had R-squared values of 0.621, 0.143, and 0.726 respectively. None of these fit statistics establishes recognition accuracy or historical fidelity.

7. Implications for evaluating restoration

This result favors a small measurement panel over an unexplained scalar ranking. Whole-image error records every pixel consistently and remains useful for a precise digital comparison. A foreground-focused denominator asks how much selected visible material was affected relative to the amount present. Both can be reported, alongside the mask, segmentation rule, and representative cases. Structural similarity provides a different account of visible structure, and perceptual feature distances offer another comparison space (Wang et al., 2004; Zhang et al., 2018). Those measures were not calculated here, so their inclusion in a future experiment remains a proposed extension.

Natural-image quality databases also illustrate the importance of the data supporting an evaluation convention. KADID-10k and KonIQ-10k cover different approaches to image-quality data construction (Lin et al., 2019; Hosu et al., 2019). A natural-scene quality estimator such as NIQE is not automatically an archaeological interpretation instrument (Mittal et al., 2013). Likewise, a distributional generative metric associated with FID does not directly answer whether one reconstructed junction corresponds to a warranted reading (Heusel et al., 2017). These references identify complementary evaluation traditions; the present numerical evidence supports only the observed disagreement between two explicitly defined squared-error rankings.

Repeated agreement between algorithms would raise another question: reliability across raters, scanning conditions, masks, and source collections. Reliability statistics require their own target and design rather than an unexamined substitution of rank correlation for every form of agreement (Koo and Li, 2016). If many alternative thresholds or mask layouts are evaluated in a follow-up, the family of comparisons should be declared and multiplicity handled transparently; false-discovery-rate methods provide one established option (Benjamini and Hochberg, 1995). Here, the fixed-mask findings are reported together,

with uncertainty and mathematical dependence made explicit, rather than filtered to retain only a preferred direction.

Several limitations constrain application. The corpus consists of normalized small crops, so sheet layout, relative character size, and contextual fragments are absent. Brightness thresholding can include surface noise or omit faint strokes. A central black square is substantially simpler than real damage. Observations may share an unrecorded source image, and the HC3 intervals do not resolve that dependence. Finally, zero filling supplies no plausible restored content. An experiment with a trained or zero-shot inpainting system must evaluate generated pixels, input conditions, computational cost, and specialist judgments before making claims about model performance or conservation utility.

8. Conclusion

An identical digital erasure did not produce a stable ordering of oracle-character cases across global and foreground-normalized error. The disagreement was large enough to move 40.7 percent of images by at least one quartile, and the conditional density association reversed direction between definitions. The finding concerns the measurement layer: an evaluator can alter apparent difficulty without changing the damaged input or running a reconstruction network. For studies motivated by Orpaint, documenting both the denominator and the selected evidence would make an eventual comparison easier to interpret. Historical interpretation still requires evidence that these small-image loss calculations do not supply.

References

1. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
2. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.
3. Wang, Mei; Deng, Weihong. (2024). A dataset of oracle characters for benchmarking machine learning algorithms. Scientific Data, 11, 87. https://doi.org/10.1038/s41597-024-02933-w.
4. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
5. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.
6. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
7. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.
8. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.
9. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
10. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
11. Guo, Hang; Li, Jinmin; Dai, Tao; Ouyang, Zhihao; Ren, Xudong; Xia, Shu-Tao. (2024). MambaIR: A Simple Baseline for Image Restoration with State-Space Model. arXiv:2402.15648. https://arxiv.org/abs/2402.15648.
12. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
13. Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; Ommer, Björn. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752. https://arxiv.org/abs/2112.10752.
14. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

15. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.

16. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.

17. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas S.. (2018a). Generative Image Inpainting with Contextual Attention. arXiv:1801.07892. https://arxiv.org/abs/1801.07892.

18. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

19. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

20. Liu, Yue; Tian, Yunjie; Zhao, Yuzhong; Yu, Hongtian; Xie, Lingxi; Wang, Yaowei; Ye, Qixiang; Jiao, Jianbin; Liu, Yunfan. (2024). VMamba: Visual State Space Model. arXiv:2401.10166. https://arxiv.org/abs/2401.10166.

21. Lin, Hanhe, et al. "KADID-10k: A Large-scale Artificially Distorted IQA Database" *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019. https://doi.org/10.1109/qomex.2019.8743252.

22. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.

23. Nazeri, Kamyar; Ng, Eric; Joseph, Tony; Qureshi, Faisal Z.; Ebrahimi, Mehran. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv:1901.00212. https://arxiv.org/abs/1901.00212.

24. Song, Jiaming; Meng, Chenlin; Ermon, Stefano. (2020). Denoising Diffusion Implicit Models. arXiv:2010.02502. https://arxiv.org/abs/2010.02502.

25. Ho, Jonathan; Jain, Ajay; Abbeel, Pieter. (2020). Denoising Diffusion Probabilistic Models. arXiv:2006.11239. https://arxiv.org/abs/2006.11239.

26. Heusel, Martin, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.

27. Gu, Albert; Dao, Tri. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752. https://arxiv.org/abs/2312.00752.

28. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

29. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

30. Hosu, Vlad, et al. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.