Temporal Systems and Forecasting
Volume 1 · Number 1 · 2026

Frozen Binarization Across Two Handwritten-Document Benchmark Cohorts

Dylan Fisher[1] · Blake Harper[2] · James Harrison[3]

[1] France · [2] United States · [3] Switzerland

Original publication: November 4, 2026 | Article 606

Abstract

A local binarization rule selected on H-DIBCO 2016 improved foreground F1 on nine of ten H-DIBCO 2018 pages, yet one large failure made its average advantage uncertain. The frozen rule had mean F1 0.648, compared with 0.515 for Otsu, and a paired difference of 0.133 with a 95% interval from −0.060 to 0.327. Average recall was lower. Reuse across benchmark cohorts therefore preserved much of the rule's usefulness without establishing a uniform benefit or identifying an effect of document age.

Exploratory public-benchmark draft. Earlier-cohort parameter selection and later-cohort evaluation; no document-aging effect or generative restoration experiment.

Frozen Binarization Across Two Handwritten-Document Benchmark Cohorts

> Exploratory public-benchmark draft. Earlier-cohort parameter selection and later-cohort evaluation; no document-aging effect or generative restoration experiment.

Abstract

A local binarization rule selected on H-DIBCO 2016 improved foreground F1 on nine of ten H-DIBCO 2018 pages, yet one large failure made its average advantage uncertain. The frozen rule had mean F1 0.648, compared with 0.515 for Otsu, and a paired difference of 0.133 with a 95% interval from -0.060 to 0.327. Average recall was lower. Reuse across benchmark cohorts therefore preserved much of the rule's usefulness without establishing a uniform benefit or identifying an effect of document age.

1. Reusing a rule beyond its selection cohort

A researcher can select a document-processing parameter on an available benchmark and later reuse it without further tuning. That is an understandable workflow, particularly when labeled historical material is scarce. However, a setting that works well on the selection pages may not suit the backgrounds, handwriting, resolution, or contrast of a later collection. Testing a frozen rule on another cohort gives a more informative answer than reporting only the maximum score achieved during its original parameter search.

The research question asks whether a local binarization rule chosen on an earlier handwritten-document benchmark cohort outperforms an untuned global rule on a later cohort. The official H-DIBCO 2016 release Pratikakis et al., 2016 provides ten source pages and their binary references. H-DIBCO 2018 Pratikakis et al., 2018 supplies ten later evaluation pages. The year ordering refers to competition cohorts. The source does not establish the creation or capture year of each document, so the experiment cannot identify a historical aging process.

The comparison is relevant to broader document restoration, while remaining a different task. Meng et al., 2025 develop zero-shot inpainting for oracle-bone inscription rubbings using a visual Mamba component. That work motivates preserving meaningful document content when processing degraded imagery. Our experiment instead classifies existing pixels as foreground or background. It neither fills missing content nor executes the proposed inpainting architecture. A binarization result therefore cannot be relabeled as evidence of that architecture's restoration or recognition performance.

Two analytical propositions organize the study. Local intensity adaptation may retain an advantage over a whole-image threshold on many later pages. At the same time, a frozen spatial scale can fail on particular documents, making a positive average insufficient for a universal reuse claim. We evaluate ordinary foreground F1 and recall together, retain every held page, and report uncertainty at the page level. These choices allow a favorable majority of pages to coexist with an important failure rather than averaging away the tradeoff.

2. Global and local threshold rules

The global comparator is Otsu's histogram-based threshold selection Otsu, 1979. Our implementation uses the grayscale histogram to choose the first threshold maximizing between-class variance. Pixels at or below that value are foreground. The rule adapts to the image histogram but has no parameter chosen from the held labels. It is therefore a fixed algorithmic comparator, not a single intensity cutoff imposed on every page. This distinction matters when describing what is frozen in the experiment.

The local comparator uses Sauvola's adaptive document-binarization construction Sauvola and Pietikäinen, 2000. For each pixel, a neighborhood mean and standard deviation determine the threshold, with a fixed reference standard-deviation scale of 128 on the eight-bit grayscale range. The parameter grid contains window sizes twenty-five, fifty-one, and one hundred one pixels, crossed with coefficients 0.1, 0.2, and 0.3. All nine candidates are evaluated only on the 2016 source pages before the later-page metrics are computed.

The local rule can respond to background variation that a global histogram does not represent spatially. That is a mechanism suggested by its computation, not a measured explanation for every page in this study. A neighborhood that contains text and background in a suitable balance can produce a useful threshold; a different stroke scale or image pattern can change that balance. We do not infer the cause of the held failure without a separate image-level diagnostic. The empirical contrast tests the frozen rule's output, not a causal account of document degradation.

Document collections also differ in their representation. The oracle-character resource Wang and Deng, 2024, Kuzushiji-MNIST Clanuwat et al., 2018, and the Omniglot challenge Lake et al., 2019 illustrate other forms of character evidence. Their cropped symbols and intended tasks are not interchangeable with full handwritten pages. We do not add them to the parameter search or test sample. Their relevance is to explain why a document-processing rule should be evaluated on its intended input support rather than assumed to transport because every source contains written marks.

3. Official images, paired references, and source integrity

Both cohorts are downloaded from their official benchmark directories with their paired binary ground truths. Archive byte counts, source URLs, and SHA256 values are retained. ZIP integrity checks pass, and every page is matched to its reference by its numerical basename. Each cohort has exactly ten image-reference pairs. The implementation verifies equal spatial dimensions within every pair and binary reference values of zero and 255. Black reference pixels define the foreground to be recovered.

The input is converted to grayscale using the recorded Pillow procedure, without resizing. Native resolution is retained throughout. Local means and second moments use reflection at image boundaries; numerical variance is clamped at zero when floating subtraction would otherwise produce a tiny negative value. These details define the actual rule. They are not interchangeable with a resized image, a different padding convention, or a window measured in physical millimeters rather than pixels.

The source mean F1 is computed with equal weight per page. The selected candidate uses a fifty-one-pixel window and coefficient 0.2, with source mean F1 0.8453903. A deterministic tie rule prefers the smaller window and then smaller coefficient. The selected pair is written to a source-rule record before running the 2018 metrics, and later executions assert that it has not changed. This is a transparent source-only selection step, without a claim of preregistration.

The official evaluation references have their own scope. They support a reproducible comparison with supplied foreground labels, but do not establish what a lost or uncertain historical mark originally meant.

The 2018 archive also contains evaluation weights that are not used here. We calculate ordinary foreground precision, recall, and F1 locally rather than running the full competition evaluation software. The resulting numbers are not presented as official entrant scores, pseudo-F-measure, or distance-reciprocal distortion results.

4. Research model and separation of selection from evaluation

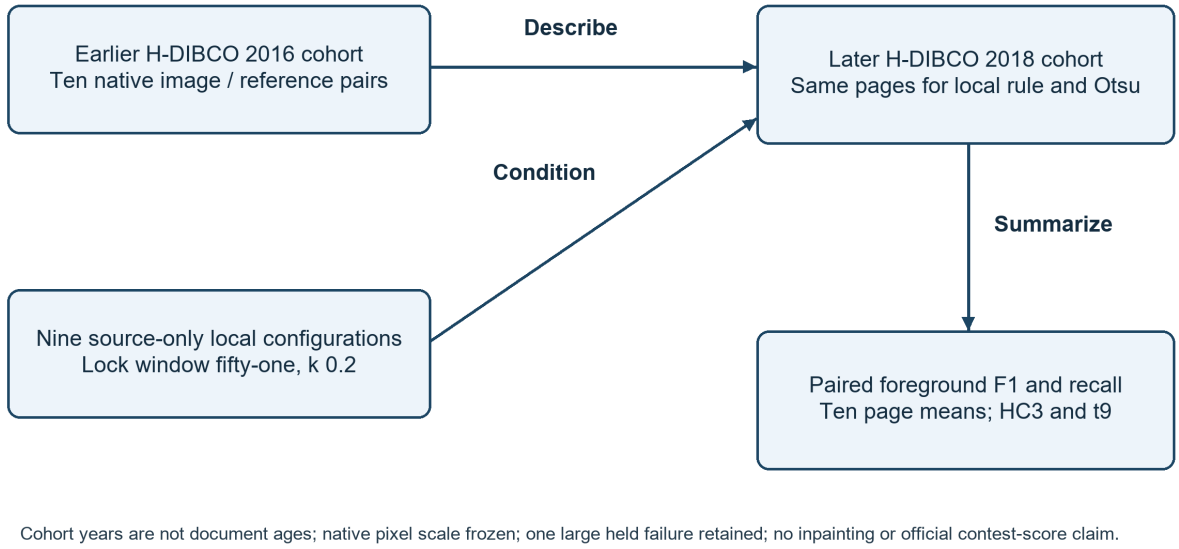


Figure 1. Earlier-cohort parameter selection and frozen later-cohort foreground evaluation.

Figure 1 distinguishes the source selection branch from the held evaluation branch. The older references select one local rule; both frozen algorithms then process the same later inputs without consulting the later binary labels. Those labels enter only metric calculation. A comparison between their held outputs therefore conditions on the realized source search. It does not treat the nine source candidates as nine independently confirmed discoveries, and it does not optimize an exception specifically for the difficult held page.

Structured evaluation guidance Roberts et al., 2017 supports keeping related inputs and selection information in the appropriate partitions. Model-selection bias discussed by Varma and Simon, 2006 explains why the maximum source score should not be interpreted as an external accuracy estimate. Here, the selected source value is useful for recording the choice, while the later cohort supplies the evidence about reuse. Differences between 0.845 on source pages and the later score cannot be attributed solely to elapsed time because the page collections differ in many uncontrolled respects.

False-discovery procedures Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 address a different statistical problem from choosing one source configuration and evaluating it once on held data. We do not attach hypothesis-test claims to the nine source grid values or pretend that those procedures certify the selected rule. If the held cohort were repeatedly reused to revise the grid, it would cease to provide the present separation. The locked parameter record and retained source sweep make that boundary inspectable for subsequent revisions.

The spatial window is frozen in pixels. That supplies a reproducible algorithm, but not invariance to scanner resolution or stroke width. A physically normalized window would require additional scale information and a different rule. No such normalization is applied after seeing the held results. This limitation is particularly relevant to transport between full-page cohorts, where similar semantic content

can occupy very different numerical supports. It offers a concrete reason to avoid equating successful source tuning with universal parameter sufficiency.

5. Foreground outcomes and page-level uncertainty

For each page, precision is the share of predicted foreground pixels that match the binary reference, and recall is the share of reference foreground pixels recovered. F1 combines those two page-level quantities through their harmonic mean. We average the resulting page F1 values, rather than pooling all pixels across the cohort. Consequently, a small page has the same weight as a large one, and macro-average F1 is not generally the harmonic mean of the separately averaged precision and recall.

The metric construction is part of the scientific question. Threshold-performance discussion Fawcett, 2006 and scoring principles Gneiting and Raftery, 2007 motivate naming the outcome rather than treating any favorable number as generic image quality. Work on the Matthews correlation coefficient Matthews, 1975 also shows why another binary summary could emphasize different aspects of classification. We do not calculate that alternative and infer no result for it. Likewise, structural similarity Wang et al., 2004 is a distinct image comparison, not a synonym for foreground F1 or readability.

The two primary contrasts subtract Otsu's held-page value from the selected local rule's F1 and recall. Intercept-only OLS models estimate their equal-page means. Standard errors use HC3 and a t reference with nine degrees of freedom, rather than counting millions of pixels as independent replicates. Robust covariance reasoning White, 1980 and dependence guidance Cameron and Miller, 2015 motivate that declared unit. No control variables or fixed effects are included, and intercept-only R-squared is zero.

The ten held pages are curated benchmark cases, not a documented random sample of all historical documents. Their intervals are conditional model-based summaries of heterogeneous page contrasts. Scanner, writer, archive, and damage-process populations are not independently sampled here. The estimates also condition on the single earlier parameter selection. A broader transport claim would require additional cohorts or a design that represents those populations, rather than a narrower interval obtained by changing the unit from pages to pixels.

6. Most pages improve, but one important failure remains

On the held cohort, the selected local rule has mean F1 0.6479992, precision 0.6311002, and recall 0.7507092. Otsu has mean F1 0.5145482, precision 0.4221818, and recall 0.7963851. These are equal-page averages. The local rule's higher mean F1 accompanies higher average precision and lower average recall, showing that the comparison is not a uniform increase in recovered foreground. Some of its apparent quality gain comes with a different balance of false positives and missed reference pixels.

The paired mean F1 difference is 0.1334510, with standard error 0.0853820 and 95% interval from negative 0.0596967 to positive 0.3265986. The recall difference is negative 0.0456759, with standard error 0.0893368 and interval from negative 0.2477697 to positive 0.1564180. Neither interval establishes a uniformly positive effect in a broader document population. The descriptive mean F1 advantage remains a valid observation, but its uncertainty should travel with the point estimate.

Table 1. Frozen local-rule differences from Otsu on ten held document pages

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.1335 (0.0854)	[-0.0597, 0.3266]	10
M1: fit	R-squared = 0.0000	-	-	10
M2	Mean outcome/contrast	-0.0457 (0.0893)	[-0.2478, 0.1564]	10
M2: fit	R-squared = 0.0000	-	-	10

Model labels: M1 = Earlier-cohort-selected Sauvola minus Otsu foreground F1; M2 = Earlier-cohort-selected Sauvola minus Otsu foreground recall.

Table 1 presents intercept-only OLS means, HC3 standard errors, t9 intervals, N10, no controls, and no fixed effects. M1 uses ordinary foreground F1 differences; M2 uses foreground recall differences. Both compare the source-selected Sauvola configuration with Otsu on the same native-resolution held pages. The unit is a page, and fit is the zero R-squared of a mean-only model. This is not the full contest metric suite or a test of image inpainting.

The page record makes the uncertainty understandable. F1 improves on nine pages, but page ten falls from 0.732904 under Otsu to 0.222261 under the selected local rule, a difference of about negative 0.510643. Its recall difference is about negative 0.757426. We retain that page without a rule change or a favorable exclusion. The majority of positive signs does not determine the average effect's uncertainty because the magnitudes are very uneven. No post hoc sign-test claim is added to replace the reported mean comparison.

7. Reuse is a local result, not a restoration guarantee

The experiment supports a qualified account of transport. A rule chosen on the older cohort remains useful on many later pages and has a higher realized mean F1, but it fails badly on one page and has imprecise average contrasts. That is more informative than either claiming universal success or discarding the method because of the failure. The retained per-page outputs identify where further investigation is needed without pretending that the current analysis has already isolated the responsible visual mechanism.

Restoration architectures address other parts of the document-processing problem. U-Net Ronneberger et al., 2015 supplies a segmentation architecture, while Context Encoders Pathak et al., 2016, EdgeConnect Nazeri et al., 2019, gated convolution Yu et al., 2018b, and partial convolution Liu et al., 2018 represent different learned image-completion constructions. They are not additional baselines in our table. A trained segmentation or completion model would need its own data partition, fitting record, and foreground or restoration outcome before it could be compared fairly with the two executed threshold rules.

Other restoration priors also remain outside the evidence. Large-mask Fourier inpainting Suvorov et al., 2021, diffusion-based RePaint Lugmayr et al., 2022, and the deep image prior Ulyanov et al., 2017 use mechanisms beyond local threshold statistics. State-space restoration Guo et al., 2024 and visual state-space modeling Liu et al., 2024 provide further architectural context. Their existence does not imply that they would solve the failed page or preserve more authentic text. The current result concerns extracting foreground from observed intensities, which cannot by itself recover content that is genuinely absent.

A useful next study would test more benchmark cohorts, document source scale more carefully, and define whether its intended loss prioritizes false foreground or missed strokes. It could compare a scale-adaptive window with the frozen pixel window under an independent selection protocol. Those proposals are not completed robustness checks. The main limitation remains small and curated page

coverage, with no verified document ages and no human readability endpoint. Temporal ordering of releases is a defensible evaluation sequence, but not a causal theory of historical change.

8. Data record and conclusion

The reproducibility record contains official archive URLs, file hashes, integrity checks, page-to-reference mappings, the nine-candidate source sweep, locked parameters, held metrics, and structured estimates. Processing uses local open-source numerical tools and does not require the benchmark's external evaluation executable. Dataset documentation Gebru et al., 2021 and FAIR stewardship Wilkinson et al., 2016 motivate preserving those provenance details. Original attribution and source-specific conditions remain attached; public benchmark availability is not treated as an unrestricted image-redistribution license.

The answer to the stated question is conditional. Source-selected local binarization improves held F1 on most pages and in the observed mean, while the full page-level evidence does not establish a uniform or precisely estimated advantage. Lower recall and one substantial failure are part of that answer. A frozen rule should therefore be reported with its selection cohort, target cohort, metric, and page-level exceptions, rather than described as generally validated by a favorable average alone.

References

1. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.
2. 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.
3. 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.
4. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
5. 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.
6. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *Lecture Notes in Computer Science*, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
7. 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.
8. 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.
9. 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.
10. Sauvola, J.; Pietikäinen, M.. (2000). Adaptive document image binarization. Pattern Recognition. https://doi.org/10.1016/s0031-3203(99)00055-2.
11. 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.
12. Matthews, B.W.. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. https://doi.org/10.1016/0005-2795(75)90109-9.
13. 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.

14. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

15. 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.

16. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

17. Lake, Brenden M.; Salakhutdinov, Ruslan; Tenenbaum, Joshua B.. (2019). The Omniglot challenge: a 3-year progress report. https://arxiv.org/abs/1902.03477.

18. 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.

19. Otsu, Nobuyuki. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/tsmc.1979.4310076.

20. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

21. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.

22. 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.

23. 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.

24. Pratikakis, Ioannis; Zagoris, Konstantinos; Barlas, George; Gatos, Basilis. (2016). ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016). 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr.2016.0118.

25. Clanuwat, Tarin; Bober-Irizar, Mikel; Kitamoto, Asanobu; Lamb, Alex; Yamamoto, Kazuaki; Ha, David. (2018). Deep Learning for Classical Japanese Literature. https://arxiv.org/abs/1812.01718.

26. 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.

27. 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.

28. 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.

29. Pratikakis, Ioannis; Zagori, Konstantinos; Kaddas, Panagiotis; Gatos, Basilis. (2018). ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr-2018.2018.00091.

30. 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.