**Journal of Transfer and Generalization**
Volume 1 · Number 1 · 2026

# Cleaning Can Erase Information: Median Filtering and Historical-Character Recognition

Adrian Scott[1] · Dominic Evans[2] · Gavin Bell[3]

[1] Denmark · [2] Norway · [3] Finland

Original publication: November 1, 2026 | Article 577

## Abstract

A denoising step can help damaged inputs while weakening clean ones. On ten thousand held-out Kuzushiji-MNIST images, a frozen ridge classifier achieved 59.54 percent accuracy without preprocessing and 58.67 percent after median filtering. Under one fixed salt-and-pepper corruption draw, accuracy fell to 53.73 percent and recovered to 56.90 percent with the same filter. An image-paired regression estimated a 4.04 percentage-point corruption-by-filter interaction. The contrast demonstrates a conditional preprocessing benefit rather than an unconditional improvement. Japanese cursive recognition and artificial pixel replacement do not establish restoration fidelity for oracle-bone rubbings or the performance of a modern inpainting model.

*Controlled corruption experiment on public Kuzushiji images. This draft reports a simple classifier, with human scientific and authorship review pending.*

# Cleaning Can Erase Information: Median Filtering and Historical-Character Recognition

Controlled corruption experiment on public Kuzushiji images. This draft reports a simple classifier, with human scientific and authorship review pending.


## Abstract

A denoising step can help damaged inputs while weakening clean ones. On ten thousand held-out Kuzushiji-MNIST images, a frozen ridge classifier achieved 59.54 percent accuracy without preprocessing and 58.67 percent after median filtering. Under one fixed salt-and-pepper corruption draw, accuracy fell to 53.73 percent and recovered to 56.90 percent with the same filter. An image-paired regression estimated a 4.04 percentage-point corruption-by-filter interaction. The contrast demonstrates a conditional preprocessing benefit rather than an unconditional improvement. Japanese cursive recognition and artificial pixel replacement do not establish restoration fidelity for oracle-bone rubbings or the performance of a modern inpainting model.


## 1. A restoration operation has a downstream purpose

Historical-character images contain marks that can be both noise-like and informative. A small isolated stroke may distinguish two classes, while an isolated damaged pixel may interfere with recognition. A local smoothing rule has no direct knowledge of that distinction. The empirical question is therefore not simply whether filtering makes an image look cleaner, but whether it helps the specified downstream reader under the input conditions that actually occur in the experiment.

Kuzushiji-MNIST supplies a public, labeled collection of historical Japanese cursive characters (Clanuwat et al., 2018). We use its official 60,000-image training partition and 10,000-image test partition. The task is ten-class recognition. These are small grayscale character crops, not full document pages. The source makes a controlled pixel-level experiment possible, while its writing system and acquisition context impose a clear boundary on claims about other historical scripts.

Work on oracle-bone inpainting, including Meng et al., 2025, motivates the broader problem of preserving useful inscription structure during restoration. Our experiment does not reproduce that model or use oracle-bone images. The oracle-character dataset described by Wang and Deng, 2024 represents a distinct source and task. We keep that distinction explicit: a preprocessing result on Japanese cursive is evidence about the tested recognition pipeline, not validation of reconstruction accuracy on ancient Chinese rubbings.

## 2. From visual plausibility to retained information

Context encoders (Pathak et al., 2016), partial convolutions (Liu et al., 2018), and gated convolutions (Yu et al., 2018b) exemplify learned ways to complete missing image regions. EdgeConnect (Nazeri et al., 2019), LaMa (Suvorov et al., 2021), and RePaint (Lugmayr et al., 2022) use different architectural or generative mechanisms. Their outputs can be plausible without necessarily preserving every feature needed by a fixed recognition model. We do not compute those methods; they motivate the distinction between attractive completion and task-preserving information.

Deep Image Prior (Ulyanov et al., 2017) offers another perspective on image structure, while MambaIR (Guo et al., 2024) applies state-space modeling to restoration. The underlying state-space developments in Gu and Dao, 2023 and Liu et al., 2024 concern expressive modeling rather than the local median operator

used here. Our intentionally simple design makes the preprocessing rule easy to reproduce and its limitations easy to identify. It should not be described as a competitive restoration benchmark against those systems.

Structural similarity (Wang et al., 2004) and learned perceptual distance (Zhang et al., 2018) are useful image-comparison concepts, but neither is our primary outcome. The experiment measures whether the class prediction matches the supplied test label. That outcome can worsen even when a visual similarity score improves. Conversely, a recognizer might tolerate an aesthetically unpleasant artifact. Without measuring both, one cannot infer that task accuracy and perceived restoration quality move together.

## 3. The paired experimental model

We evaluate four conditions on the same test images: clean without filtering, clean with filtering, corrupted without filtering, and corrupted with filtering. The first proposition is that the fixed filter can change recognition accuracy even when no corruption is added. The second is that its effect differs under corruption. This interaction is the central quantity: it asks whether the preprocessing effect depends on the input condition, rather than assuming that denoising is uniformly beneficial.

The model is a linear probability regression of correct classification on corruption, filtering, their interaction, and class indicators. Because each source image appears four times, covariance is clustered by image. Class indicators adjust for baseline differences across the ten categories but are not tuned to produce a favorable effect. In this balanced design, the policy contrasts can also be read directly from the four empirical accuracies.

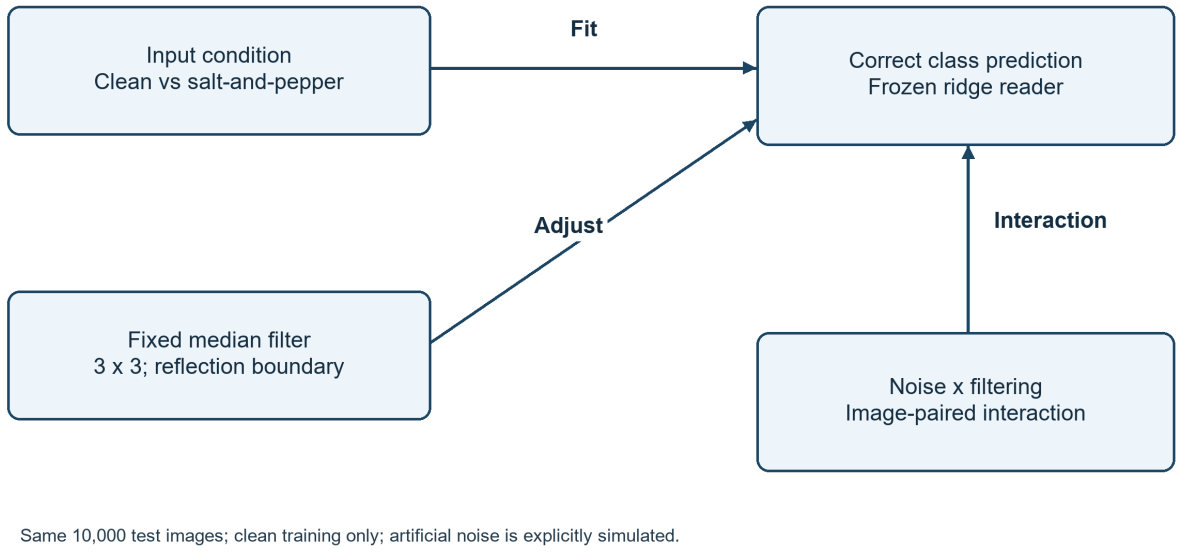



*Figure 1. A clean-trained classifier is frozen before test images receive corruption and median-filter conditions. The fitted interaction describes how filtering changes paired correctness under the controlled corruption.*

## 4. A classifier fixed before evaluation

Every 28 by 28 image is converted to 196 features by averaging nonoverlapping two-by-two pixel blocks and dividing intensities by 255. Training means and standard deviations standardize these features. We fit a linear ridge regression to ten one-hot class columns, with penalty 100 on feature coefficients and an unpenalized intercept. The predicted class is the largest fitted score. This is a regularized least-squares classifier, not a calibrated ten-class probability model.

The classifier is fitted once on clean training images. Filtered test images are passed through the same training-derived scaling and coefficient matrix. We do not retrain on filtered images or augment training with corruption. Those would be legitimate additional interventions, but they would confound our narrow question about inserting one preprocessing operation into a frozen reader. The fixed rule also avoids choosing the penalty or block size by test accuracy.

Structured evaluation concerns in Roberts et al., 2017 and selection-bias analysis in Varma and Simon, 2006 support holding the test partition outside the fitting procedure. The official test set still belongs to the same benchmark, so it does not establish performance on new archives, writers, or document conditions. Dataset-shift evidence in Ovadia et al., 2019 motivates checking such differences rather than assuming that one test accuracy transfers to every historical-image source.

## 5. Corruption is controlled, not naturally observed damage

We use random seed 20261106 and independently mark each test pixel for replacement with probability 0.20. A marked pixel is replaced by either zero or 255 with equal probability. The same resulting corrupted image is used in its filtered and unfiltered conditions. The realized mask fraction is 20.028 percent. Only 12.729 percent of pixel values actually change because some replacements equal the original intensity, especially at existing black or white pixels.

This difference between selected pixels and altered values is important for reproducibility. Calling the realized image twenty percent different would misdescribe the operation. The procedure approximates impulsive pixel contamination, not tears, missing regions, ink bleed, rubbing artifacts, or authentic conservation damage. No claim is made that the selected corruption prevalence matches a historical archive. There is only one random realization, so the reported intervals do not represent repeated draws of the corruption mechanism.

The filter is a three-by-three spatial median with reflection at image boundaries. It operates separately on each image and is never applied across the image batch dimension. It is used identically in the clean and corrupted conditions. The graph-free operator has no learned parameters, semantic model, or character dictionary. Its effect comes from modifying local intensity order statistics, which can suppress isolated extremes and can also remove thin informative details.

## 6. Inference attached to the paired design

There are 40,000 condition-image rows, but only 10,000 source-image clusters. We use a finite-sample cluster correction and t-based intervals. Cameron and Miller, 2015 explains why repeated measurements should not be treated as unrelated samples, while White, 1980 provides the broader heteroskedastic covariance foundation. Clustering by image allows its four correctness outcomes to covary; it does not account for possible common writers or documents whose identifiers are unavailable in this experiment.

The model uses correctness as a binary descriptive response. Guo et al., 2017 concerns probability calibration, which we do not assess because the ridge score vector is used only for its largest component. Gneiting and Raftery, 2007 reinforces the need to identify the evaluation target before interpreting a metric. Our target is benchmark class accuracy for a frozen pipeline. It is neither human readability, transcription completeness, restoration authenticity, nor uncertainty calibration of a deep character recognizer.

*Table 1. Class-adjusted paired correctness contrasts. Coefficients are probability differences; multiplying by one hundred gives percentage points. Covariance clusters the four conditions of each test image.*

| Model | Term | Estimate (SE) | 95% CI | N |
|---|---|---|---|---|
| M1 | Corruption, no filtering | -0.0581 (0.0039) | [-0.0657, -0.0505] | 40000 |
| M1 | Filter effect, clean input | -0.0087 (0.0024) | [-0.0133, -0.0041] | 40000 |
| M1 | Corruption x filtering | 0.0404 (0.0042) | [0.0321, 0.0487] | 40000 |
| M1: fit | R-squared = 0.0261 | - | - | 40000 |

Model labels: M1 = Paired correctness contrasts.

## 7. The clean-image cost and corrupted-image benefit

Clean accuracy is 0.5954 without the filter and 0.5867 with it. The clean filtering effect is therefore -0.0087, or a loss of 0.87 percentage points. Its standard error is 0.0024 and its interval is -0.0133 to -0.0041. This is evidence that an automatic cleanup step can damage information useful to this particular clean-trained classifier even though the operation is commonly described as denoising.

Unfiltered corrupted accuracy is 0.5373. Filtering raises it to 0.5690, a gain of 3.17 percentage points in that condition. The corruption-by-filter interaction is 0.0404, with interval 0.0321 to 0.0487. It measures the difference between the positive corrupted-image filtering effect and the negative clean-image filtering effect. It should not be confused with the total filtered-corrupted advantage over clean images, because the latter condition still has lower accuracy than the unfiltered clean baseline.

Corruption without filtering reduces accuracy by 5.81 percentage points, with interval -6.57 to -5.05 points. The full correctness model has R-squared 0.0261. That modest value is unsurprising for individual binary outcomes and does not erase the paired average contrasts. Class-level summaries in the archived output also show that a positive overall effect need not hold identically for every character category. We do not select only the favorable classes for the main result.

The simple reader's absolute accuracy is limited. A more capable model could use the available information differently and show another filtering effect. This is why the result is framed as a pipeline interaction rather than an inherent property of the dataset. The classifier's weaknesses are part of the tested system. Changing its architecture or training procedure would require a new frozen evaluation design, not a reinterpretation of the current coefficient as a general restoration law.

## 8. Implications and limits

The four conditions also clarify which comparison answers which practical question. If every incoming image is known to be corrupted by this specific mechanism, the filtered-versus-unfiltered corrupted contrast is relevant. If input condition is unknown and the filter is always applied, the clean-image loss also matters. The overall value of that automatic policy would depend on the prevalence of corruption in actual use and on the relative consequences of different recognition errors. Neither quantity is estimated from a test set in which we deliberately create both versions of every image.

Pairing prevents class composition from explaining the average filtering difference, because each condition contains the same thousand images per class. It does not ensure that damage effects are uniform across instances. A filter may correct some predictions and break others, with the reported contrast recording only the net difference. A later inspection could compare those two groups using expert-reviewed stroke properties, but it should define the inspection rules before claiming a mechanism.

The present processed predictions preserve that possibility without supplying an unsupported explanation for each changed classification.

The training standardization is another part of the tested pipeline. Filtered images can have feature distributions that differ from clean training inputs, and the frozen coefficients may react to those differences. Retraining after preprocessing could reduce that mismatch, while corruption augmentation could alter the classifier's sensitivity to extreme pixels. Those are plausible extensions, not reasons to reinterpret the current result as a universal property of median filters. Its empirical meaning is conditional on a clean-trained, low-dimensional linear reader with a fixed penalty and a fixed feature extraction rule.

Finally, downstream correctness is a narrower conservation objective than faithful reconstruction. A method could improve class recognition while erasing variation important to paleography, handwriting analysis, or provenance. Conversely, preserving an authentic unusual stroke could confuse a weak classifier. The benchmark's ten supplied classes do not adjudicate those scholarly goals. Any archival workflow should therefore define which information must survive restoration and gather appropriate expert evidence, rather than assuming that higher machine accuracy alone establishes historical fidelity.

Robust-statistics ideas in Huber, 1964 and Hampel, 1974 help explain the attraction of suppressing extreme values, yet robust preprocessing and task preservation are different objectives. The median can remove an outlying pixel without knowing whether that pixel is noise or a legitimate small stroke. Cook, 1977 provides a framework for investigating influential observations in regression, but deleting difficult test images after seeing their predictions would no longer answer the balanced paired question posed here.

Measurement considerations in Carroll et al., 2006 also apply to the class labels and the distinction between noisy images and uncertain transcription. Missing-data principles in Little and Rubin, 2002 would be relevant to absent or unreadable labels; no such imputation is performed. The inferential guidance in Wasserstein and Lazar, 2016 encourages reading effect sizes and their design context together. A narrow interval around this artificial experiment cannot establish the authenticity of restored historical content or justify unattended archival replacement.

For a real archive, the next useful experiment would combine documented damage categories with expert transcription and compare recognition before and after a prespecified restoration policy. A damage detector might decide when to apply filtering, but its own errors would become part of the end-to-end outcome. Our four-condition comparison motivates that conditional policy question without claiming to have solved damage detection, conservation assessment, or cross-script transfer.

The release record follows Gebru et al., 2021 by identifying the Japanese source, image dimensions, artificial operation, and evaluation partition. Reproduction materials follow Wilkinson et al., 2016 through preserved data provenance, fixed seeds, and processed predictions. Dataset attribution and share-alike conditions remain relevant to reuse. The empirical takeaway is the interaction itself: a preprocessing step that helps a declared damaged-input condition can impose a measurable cost when applied indiscriminately to clean inputs.

## References

1. 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.

2. 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.

3. 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.

4. 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.

5. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

6. 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.

7. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

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. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

10. 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.

11. 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.

12. 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.

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

14. 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.

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

16. 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.

17. 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.

18. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

19. 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.

20. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

21. 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.

22. 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.

23. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.

24. 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.

25. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

26. 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.

27. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

28. 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.

29. 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.

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