ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887712

Forensics Analysis of Residual Noise Texture in digital Images for Detection of Deepfake

Arthur Méreur
Troyes University of Technology
France
arthur.mereur@utt.fr

Antoine Mallet
Troyes University of Technology
France
antoine.mallet@utt.fr

Rémi Cogranne
Troyes University of Technology
France
remi.cogranne@utt.fr

Minoru Kuribayashi
Center for Data-driven Science and Artificial Intelligence
Tohoku University, Japan
kminoru@tohoku.ac.jp

***Abstract*—This paper proposes an original approach for the automatic detection of AI-generated images, using features derived from noise residuals artefacts. Contrary to most current research that leverages sophisticated deep learning models to further improve performance, this study highlights the distinct noise residual characteristics in deepfakes, facilitating the identification of AI-generative images. Our findings highlight some limitations of image models, which can be used for forensic analysis and for future AI-based text-to-image generative models. Broad numerical results on a large and diverse dataset show the interest of the identified features as well as the relevance of the present method.**

***Index Terms*—DeepFakes, Noise residual, Explainable method, Machine learning, Statistical detection.**

I. Introduction

The rapid development of AI has made the generation of multimedia content accessible to all thanks to intuitive tools offering new opportunities in terms of realism, personalization, and speed. However, this digital revolution has also given rise to new challenges, particularly on social networking platforms where disinformation and deepfakes can circulate quickly and widely, often for political purposes. With the proliferation of deepfakes came a pressing need for reliable methods that can distinguish genuine photographs from fabricated images [1], [2].

In this context, many deep learning models have been proposed for detecting AI-generated content and deepfakes [3-7] often reaching impressive results. However, the common limitations of those models, often referred to as "black boxes" are well known: their internal workings remain opaque hence their lack of explainability and interpretability preventing their widespread adoption [7], [8]. Developing transparent and interpretable alternative methods is crucial for understanding deepfakes' limitations and enabling more explainable detection tools.

In this study, we focus on a specific characteristic of AI-generated images: noise texture. We propose a rather simple method, which does not rely on deep learning techniques, aimed at analysing residual noise patterns, or texture, and identifying distinctive features that would differentiate deepfake from natural photographs. By focusing on the analysis of noise texture, we seek to demonstrate that current AI models, while impressive, are still unable to faithfully reproduce the complexity and variability of noise present in real photographic images.
The results of this research could contribute to a better understanding of the current limitations of generative AI methods, enabling their improvement and facilitating the more accurate and interpretable detection of deepfakes.

The present paper is organized as follows: Section II explains the rationale of the proposed method, supported by an illustrative example. Section III details practical and methodological aspects of the proposed approach. Then Section IV shows the relevance and limitations of the proposed methodology through numerical results on large and diverse image databases. Eventually, Section V summarizes the contribution of the proposed method and presents an outline of future works.

II. State-of-the-art and position of the method

This paper is in line with our previous work, which is based on multivariate Gaussian modelling of noise in digital photographs with applications to source identification [9], data hiding [10]–[12], or to characterize *cover-sources* [13].

The central objective of our work is to expose the limitations of current AI-generative models, specifically regarding the realism of noise in digital images and its statistical properties. To this end, we want to show that it is possible to design a simple and effective detector for distinguishing deepfakes from natural photographs using only residual noise pattern or texture. It is important to acknowledge that the specific noise texture observed in deepfake images has already been identified and discussed in several previous studies pointing out its specificity [14], [15]. For instance, such observation

This work has been funded by the EU's Horizon 2020 program under grant agreement No. 101021687(UNCOVER), and the French ANR PACeS project No. ANR-21-CE39-0002.This work was conducted while A. Mallet and R. Cogranne were invited, by Prof. Kuribayashi, at Tohoku University.

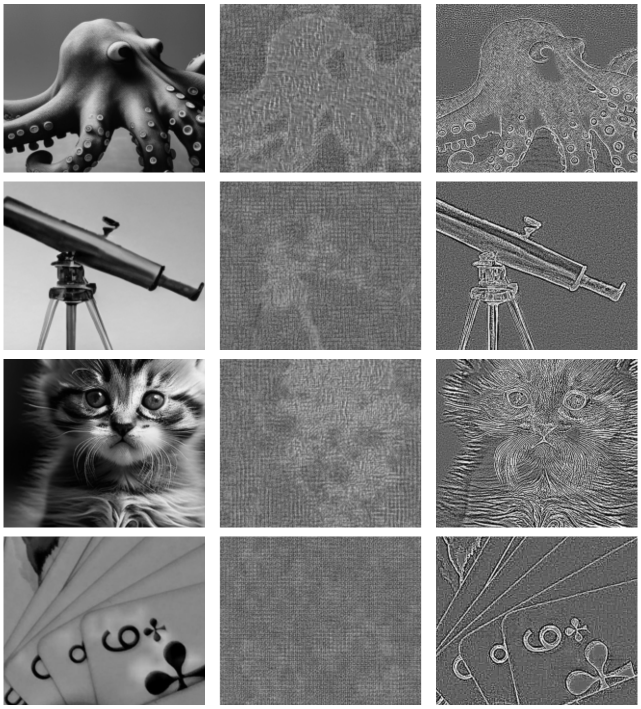

Fig. 1. Illustration of images and their noise residuals ; The second column are Noiseprint [23] residuals and the third column are Laplacian filter residuals. In rows images from: Kandinsky, MegaDall-E, Pixart-α and a real photograph.

was reported in [16], [17] which attempted to design a deepfake detector based on the well-known Photo-Response Non-Uniformity (PRNU). These works, however, concluded that noise pattern is very reliable for device identification, but not a good forensic feature for Deepfake detection.

To exploit the "noise camera fingerprints" it was proposed in [18] adding an unsupervised classification to the Peak to Correlation Energy (PCE) originally proposed in [19]. A fundamentally different approach has been proposed in [20] designing a specific deep learning model based on the well-know Local Binary Pattern in order to exploit the local noise residual texture. A similar specific and novel approach was proposed in [21] using a Siamese architecture together with the RIDnet denoising model. This method analyze the consistency between faces from the background with the assumption that only faces are modified using AI-generative models. Eventually, another similar approach based on Multi-head Attention network has been proposed in [22] with the underlying rationale that residual textural patterns differ in different regions.

These prior works show that exploiting the peculiarity of deepfake images noise texture, or pattern, has some potential but it is far from being obvious.

In this context, the principal objective of this paper is to present a proof-of-concept demonstrating that residual noise texture is a valuable forensics analysis offering insights for deepfake detection.

III. Presentation of the Proposed Method

A. Extracting Noise Residuals

The observations explained in Section II are illustrated in Figure 1, which depicts a representative visual comparison of the noise residuals extracted from various generative AI models with a natural photograph. This figure especially emphasized that the noise extraction method used does impact significantly the ensuing residuals.

In order to design the proposed method as a proof-of-concept, we have deliberately chosen to use two complementary noise estimation techniques. On the one hand, we used Noiseprint [23], an advanced deep-learning method based on Siamese networks and designed originally for camera model fingerprinting. On the other hand, we used the well-known discrete Laplacian second-order differential operator [24] which can be implemented as a simple 2D convolution filter with the following kernel:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \quad (1)$$

The interest of these complementary approaches is illustrated in Figure 1, which offers a representative visual comparison of the noise residuals extracted from various generative AI models with a natural photograph (last row).

The middle column corresponds to the residuals from Noiseprint. The presence of much more complex textures in the deepfakes images (first three rows) can be observed from the uniform values in the photograph. Similarly, the results of the discrete Laplacian operator (right-hand column) show clear and sharp edges in the photograph, with significantly smaller residuals in the homogeneous areas.

B. Measuring Texture Complexity

Undoubtedly, additional noise residuals can be extracted; again, with the aim of designing a proof of concept, and given the size constraints of present submission, we limited our analysis to two complementary residuals and, for each, two complexity measures.

On the one hand, we used the fractal dimension which is a rather common concept for image surfaces description and classification [25]. The fractal dimension is based on the concept of self-similarity, which is straightforward for mathematically defined surfaces and curves, but is difficult to measure directly on image data, which generally do not possess self-similarity in the strict sense [25]. To this end, it is popular to substitute the fractal dimension with the efficient and accurate box-counting approach which operates on tiles, or "boxes", of $D \times D$ pixels as follows:

$$F_D = \lim_{D \to 0} \frac{-logN(D)}{log(D)} \quad (2)$$

Here, F_D stands for the estimated fractal dimension, $N(D)$ is the number of "boxes" of side D containing pixels with intensity greater than a given threshold τ.

In our case we used the absolute value of noise residuals and set the threshold to a very small value (typically $\tau = 0.01$). In addition, the estimation as the size of the boxes D tends to 0 is often replaced with several (small) values and a linear regression is carried out. Since we wanted to use the fractal dimension as a global measure of noise texture, we used 44 values of $N(D)$ with $2 < D < 112$ scaled according to a logarithmic scale.

As a complementary texture complexity measure [26], we used a local approach proposed in [27]. It is based on non-overlapping blocks of $D \times D$ pixels extracted from absolute values of noise residuals, denoted $\mathbf{R} = \{r_{m,n}\}$. For a block of residual noise $\mathbf{R}$, this measure of Texture Complexity, $TC(\mathbf{R})$, is computed as:

$$TC = log(t(\mathbf{R})/(1 - t(\mathbf{R}))) + 4, \quad (3)$$

$$with \quad t(\mathbf{R}) = 1 - \sum_{m,n} 2^{-r_{m,n}} \frac{r_{m,n}}{D^2}. \quad (4)$$

In our case, the size $D = 16$ pixels was chosen as it provides the best tradeoff.

Once again, there exists many approaches to characterize and classify images based on noise texture [26]. In the present paper, we retain two complementary approaches: Fractal dimensions being a common measure of global texture complexity while local texture complexity is usual for weak signal detection in image forensics.

Each being used with the two aforementioned different methods for noise extraction, we eventually ended up with 4 clearly distinguish feature sets.

IV. Experimental Setup and Numerical Results

A. Common Benchmark for all Experiences

In order to show the relevance of the noise texture complexity for forensics analysis of deepfake images, we have conducted rather large-scale numerical experimentation using four deepfake generators. To maximize diversity we included StyleGAN 3 [28] as a representative of Generative Adversarial Network (GAN) and one transformer-based model, namely "DALL·E Mega" [29] which "attempted to reproduce OpenAI DALL·E results with an open-source model". We also included in the numerical results two diffusion-based models from the state of the art, namely Kandinsky 2 [30] and Pixart-α [31].

In order to maximize diversity, we randomized the prompts as well as some hyperparameters.

For real images we used images from ALASKA dataset [32], [33] and, once again, to increase variability, included images downloaded from Flicker.

We used images of size 512×512 pixels and used only the luminance channel, which contains the most visual information. Our dataset consists of $20,000$ real images and $5,000$ images from each AI-generative model, for a total of $20,000$ deepfakes.

All the results we present were obtained through k-fold cross-validation, with $k = 5$ hence the use of 80% of data for training.

TABLE I
Detection and identification performance for all proposed features sets and for different scenarios.

		Features Set				
Scenario		NP-F_D	Lap. F_D	NP-TC	Lap. TC	Merge
Known	Acc	85.98	77.40	81.55	87.34	96.01
	AUC	0.932	0.845	0.876	0.937	0.987
Holistic	Acc	84.78	76.33	79.47	84.97	93.38
	AUC	0.923	0.835	0.855	0.918	0.978
Atomistic	Acc	46.52	62.88	68.52	75.44	82.65

B. Numerical Results

All classification results were obtained using a Linear Support Vector Machine (SVM). Table I summarizes the performance of the presented features set. For clarity, let us recall that F_D represents the box-counting fractal dimension (2) whose dimension is 44 and TC stands for the Texture Complexity as defined in (3) whose dimension is 1024, for the image of size 512×512 ; similarly, 'NP' and 'Lap.' corresponds respectively to the noise residuals extracted with NoisePrint and Laplacian discrete differential operators.

It is worth mentioning that the accuracy (in %) is given using the Balanced Accuracy which is defined as the average of the True-Positive and True-Negative rates such that the numbers can always be compared even in the case of imbalanced classes (match case).

In order to emphasize better the limit of the proposed method, three "scenarii" were considered:

- the '*Known*' case is when a single AI-generative model is considered: the binary classifier is therefore trained to detect a specific AI-generative model;
- the '*Holystic*' case considers all four AI-generative models as a single class hence training and testing are out merging all "deepfake images";
- the '*Atomistic*' case uses another approach: it uses a multi-class classifier to identify the most likely AI-generative model.

Several interesting conclusions can be made from Table I. First, one can note that the fractal dimension works very well with NoisePrint while, on the opposite, the local texture complexity works better with the Laplacian operator residuals. This was generally true in all our experimentations.

One can also note that the generalization to four very different text-to-image AI-generative models seems very efficient with the holistic strategy. That is when the training is carried out using all deepfakes as one single class. Indeed, the loss of detection performance is rather limited as compared to the most favourable '*Known*' case in which the model is known to the detector.

On the opposite, it is striking that the atomistic approach fails to identify the generative model since the results reported in the Table I show that the correct class can be identified in only half of the case. Even though it is worth noting the multi-class case, our results also confirmed that even for binary detection (deepfakes / photographs) this approach perform much worse as compared to the 'holistic' approach. Our results point out that our method is rather limited in terms of identification of a specific text-to-image model. This observation and our results tend to point out that text-to-image models are limited in terms of replication of real images.

We also would like to emphasize that the small dimension of fractal dimension, or box-counting features, with respect to the very large size of the texture complexity make the merging of these features far from being straightforward. In addition, these characteristics provide distinct information and our experience shows that concatenating them does not significantly improve the performance of the ensuing detection.

Two important information should be provided here. On the one hand, we have increased the feature space of the fractal-dimension by using the well-known Nyström approximation. This was, however, not sufficient for a linear classifier to take proper accounts of the different features altogether. Therefore, the resulting presented as ‘merge’ is in fact the simple average of soft-output of the four linear classifier.

We are aware that our classification method is far from optimal and that additional noise-texture-related features could be added. We acknowledge that there is room for improvement as the present paper mostly aims at presenting a POC.

Last but not least, Figure 2 presents the same results as those reported in Table I under the form of a ROC curve. Note that this figure also presents a numerical comparison with [34] evaluated only on images generated by diffusion model and with [35] evaluated only on images generated by GANs. For readability and comparison, this figure present only the results obtained with our detection method in the most interesting and realistic case, namely, the holistic scenario.

The green curve plots the ROC curve when merging the decision from all four linear classifiers. While the results are interesting, it shows that the current methods perform better at identifying deepfakes than real photographs. Indeed, one can note that it is possible to reach almost 100% True-positive rate for 20% of photographs falsely labelled as deepfakes. Unfortunately, the opposite is not as encouraging and for even a 50% detection rate of deepfakes, a few percent of false positives have already been achieved.

V. Conclusion and Future Works

The methodology presented in this paper shows a rather high degree of accuracy in detecting deepfakes by solely analysing the residual noise textures within images. Although our approach does not match the efficiency of latest deep-learning models, it offers unique merits in terms of providing valuable insights into result interpretability. To the best of our knowledge, it is the first method for deepfake detection that does not rely on deep learning models and, therefore, it offers valuable insights into the interpretability of results.

More importantly, our results underscore the limitations of current generative AI models which are specifically trained for content replication. However, these models often overlook the intricate details and statistical properties of real-world images. These findings can either be exploited to improve AI-generative models focusing on capturing the statistical properties and noise characteristics of real-world images. Our future work will be to improve the analysis presented in the paper, by looking at colour components and additional noise residuals and texture complexity measurements, in order to be able to improve detection performance and locate AI-based content generation while preserving explainability of the results.

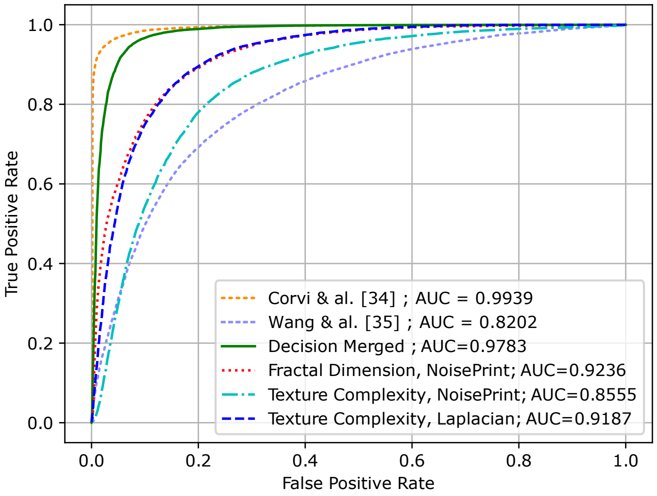

Fig. 2. Graphical representation of the performance of the proposed method and numerical comparison with [34] on images generated by diffusion model and with [35] on images generated by GANs.

References

[1] Avril Haines, “An update on foreign threats to the 2024 elections, senate select committee on intelligence,” May 2024.

[2] “Roadmap for researchers on priorities related to information integrity research and development,” the National Science and Technology Council, 2022.

[3] L. Bondi *et al.*, “Training strategies and data augmentations in cnn-based deepfake video detection,” in *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2020, pp. 1–6.

[4] M. Masood *et al.*, “Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward,” *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, jun 2022.

[5] D. Pan *et al.*, “Deepfake detection through deep learning,” in *2020 IEEE/ACM BDCAT*, 2020, pp. 134–143.

[6] P. Korshunov and S. Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” *arXiv preprint arXiv:1812.08685*, 2018.

[7] W. Samek and K.-R. Müller, “Towards explainable artificial intelligence,” *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.

[8] Y. Zhang*et al.*, “A survey on neural network interpretability,” *IEEE Trans. on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[9] A. Mallet, P. Bas, and R. Cogranne, “Statistical correlation as a forensic feature to mitigate the cover-source mismatch,” in *Proc. ACM IH&MMSEC’24*, 2024.

[10] T. Taburet, & *al.* “Jpeg steganography and synchronization of dct coefficients for a given development pipeline,” in *Proc. of the 2020 ACM Workshop IH&MMSEC*, New York, NY, USA, 2020.

[11] E. Giboulot, R. Cogranne, and P. Bas, “Detectability-based jpeg steganography modeling the processing pipeline: The noise-content trade-off,” *IEEE Trans. on Information Forensics and Security*, vol. 16, pp. 2202–2217, 2021.

[12] E. Giboulot, P. Bas, and R. Cogranne, “Multivariate side-informed gaussian embedding minimizing statistical detectability,” *IEEE Trans. on Information Forensics and Security*, vol. 17, 2022.

[13] A. Mallet, R. Cogranne, and P. Bas, “Linking intrinsic difficulty and regret to properties of multivariate gaussians in image steganalysis,” in *Proc. ACM IH&MMSEC’24*, 2024.

[14] J. Yang *et al.*, "Mtd-net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. on Information Forensics and Security*, vol. 16, pp. 4234–4245, 2021.

[15] C. Miao *et al.*, "F 2 trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. on Information Forensics and Security*, vol. 18, pp. 1039–1051, 2023.

[16] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *Proc. Irish machine vision and image processing conference (IMVIP)*, 2018, pp. 133–136.

[17] C. De Weever, *et al.*, "Deepfake detection through prnu and logistic regression analyses," Tech. Rep., Technical report, University of Amsterdam, 2020.

[18] J. Pu *et al.*, "Noisescope: Detecting deepfake images in a blind setting," in *Proc. ACM ACSAC*, p. 913–927, 2020.

[19] M. Chen *et al.*, "Determining image origin and integrity using sensor noise," *IEEE Trans. on information forensics and security*, vol. 3, no. 1, pp. 74–90, 2008.

[20] S. Kingra, N. Aggarwal, and N. Kaur, "Lbpnet: Exploiting texture descriptor for deepfake detection," *Forensic Science International: Digital Investigation*, pp. 301452, 2022.

[21] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14548–14556, Jun. 2023.

[22] H. Zhao *et al.*, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF CVPR*, 2021, pp. 2185–2194.

[23] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[24] V. Berzins, "Accuracy of laplacian edge detectors," *Computer Vision, Graphics, and Image Processing*, vol. 27, no. 2, pp. 195–210, 1984.

[25] J. M Keller, S. Chen, and R. M Crownover, "Texture description and segmentation through fractal geometry," *Computer Vision, Graphics, and image processing*, vol. 45, no. 2, pp. 150–166, 1989.

[26] M. P. Petrou and S.-E. Kamata, *Image processing: dealing with texture*, John Wiley & Sons, 1st edition edition, 2006.

[27] D. Hu *et al.*, "Study on the interaction between the cover source mismatch and texture complexity in steganalysis.," in *Multimed Tools Appl*, 2019.

[28] T. Karras*et al.*, "Alias-free generative adversarial networks," *Proc. NeurIPS*, vol. 34, pp. 852–863, 2021.

[29] B. Dayma *et al.*, "Dall·e mini," 2021.

[30] A. Razzhigaev *et al.*, "Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion," *arXiv preprint arXiv:2310.03502*, 2023.

[31] J. Chen *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.

[32] R. Cogranne, E. Giboulot, and P. Bas, "The alaska steganalysis challenge: A first step towards steganalysis," in *Proc. ACM IH&MMSEC'19*, 2019, pp. 125–137.

[33] R. Cogranne, E. Giboulot, and P. Bas, "Alaska# 2: Challenging academic research on steganalysis with realistic images," in *Proc. IEEE WIFS*, 2020.

[34] R. Corvi *et al.*, "On the detection of synthetic images generated by diffusion models," *In Proc. ICASSP*, 2023.

[35] S.-Y. Wang *et al.*, "CNN-generated images are surprisingly easy to spot... for now," *In Proceedings of the IEEE/CVF CVPR*, 2020.