Chapter 4 / Capítulo 4

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch04
Pages: 36-46

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Frequency-domain detectors for machine-generated images and their generalization across generators

Detectores en el dominio de la frecuencia para imágenes generadas por máquina y su generalización entre generadores

Vugar Hacimahmud Abdullayev[1]

1 Azerbaijan University of Architecture and Construction. Baku, Azerbaijan

ABSTRACT

Introduction: Machine-generated images are now easy to produce and hard to distinguish from photographs, and reliable detection is needed for forensics and content moderation. Deep detectors work well on familiar generators but degrade on unseen ones; cheap frequency-domain features are an alternative whose cross-generator behavior is unclear. Methods: A sample of 500 real photographs and 500 images from each of three generators (BigGAN, GLIDE, VQDM) was drawn from the GenImage benchmark. Fifty-nine frequency-domain features were extracted from each image and fed to a logistic regression trained on one generator, then evaluated in-domain and on the two held-out generators with accuracy and area under the curve. Results: Mean in-domain accuracy was 0.77 and fell to 0.65 across generators, an average drop of 12 percentage points. The drop was uneven, with GLIDE reaching 0.92 in-domain while most cross-generator transfers fell below 0.66, and one transfer was weaker than a coin toss. Conclusions: Frequency-domain features support lightweight detectors that work within a generator but transfer poorly to unseen generators, because they capture generator-specific spectral artifacts. They are a useful first-line screen when a detector must run without a learned model, but not a substitute for generators-diverse training.

Keywords: synthetic image detection; frequency domain; image forensics; generalization; generative models; machine learning

RESUMEN

Introducción: Las imágenes generadas por máquina son fáciles de producir y difíciles de distinguir de las fotografías, y se necesita una detección fiable para la forensía y la moderación de contenidos. Los detectores profundos funcionan bien con generadores conocidos pero se degradan con los nuevos; las características de frecuencia son una alternativa barata cuyo comportamiento entre generadores no está claro. Métodos: Se extrajo una muestra de 500 fotografías reales y 500 imágenes de cada uno de tres generadores (BigGAN, GLIDE, VQDM) del benchmark GenImage. Se extrajeron 59 características de frecuencia de cada imagen y se alimentó una regresión logística entrenada en un generador, evaluada en dominio y en los dos generadores excluidos con exactitud y área bajo la curva. Resultados: La exactitud media en dominio fue 0,77 y cayó a 0,65 entre generadores, una caída media de 12 puntos porcentuales. La caída fue desigual, pues GLIDE alcanzó 0,92 en dominio mientras que la mayoría de las transferencias entre generadores quedaron por debajo de 0,66 y una fue peor que el azar. Conclusiones: Las características de frecuencia permiten detectores ligeros que funcionan dentro de un generador pero transfieren mal a generadores no vistos, porque capturan artefactos espectrales específicos de cada generador. Son un cribado inicial útil cuando no se dispone de un modelo aprendido, pero no sustituyen al entrenamiento con generadores diversos.

Palabras clave: detección de imágenes sintéticas; dominio de la frecuencia; forensía de imagen; generalización; modelos generativos; aprendizaje automático

INTRODUCTION

Machine-generated images have moved from a laboratory curiosity to a widely available commodity in a few short years, and the practical problem they create is simple to state: an observer can no longer reliably tell whether a picture was captured by a camera or produced by a model.[1,2] That uncertainty matters deeply wherever images serve as evidence, whether in journalism, in courts, in financial fraud, or in the content that online platforms distribute at scale. A detector that can tell the two apart is therefore a basic instrument of trust in visual media, and a great deal of effort has gone into building such detectors.(3,4)

The need is not a single need but a spectrum of them, and the spectrum shapes what a good detector looks like. A newsroom verifying a contested photograph wants a decision it can explain to an editor, and an explanation that cites a spectral anomaly is far more usable than one that cites a deep feature vector. A platform screening millions of uploads wants a detector cheap enough to run on every image. A forensic examiner wants a method whose reasoning survives scrutiny in a report. These requirements pull in different directions, and the deep networks that dominate the literature satisfy only the accuracy requirement, leaving the cheaper and more interpretable alternatives under-explored as a consequence.(5) It is in that gap that the frequency-domain approach sits.

The strongest detectors are deep networks trained on large sets of real and generated images, and they work well when they are asked about the same generators they were trained on. Wang and colleagues demonstrated this pattern early, showing that a detector trained on one generator family recognizes those images easily while its performance on other families falls away.(6) The difficulty is precisely generalization: a detector is useful only if it survives the arrival of new generators, and the arrival is constant.[7,8] The GenImage benchmark was created to measure this property at scale, and its results show that the leading deep detectors degrade substantially when moved from a seen to an unseen generator.(9) Subsequent work has refined the picture by showing which features survive the transfer and which do not.(5,10)

The problem is made harder by an arms race. As detectors have improved, so have the methods for erasing the traces they rely on, and a detector that can be defeated by a simple fingerprint-removal step is of limited forensic value in a real investigation.[11,12] The same generators that produce convincing images are being tuned to leave fewer detectable artifacts, which means any detection approach must be evaluated not only against current generators but against the expectation that its signal will be actively attacked. This is the context in which a lightweight, transparent detector is attractive: it is cheap enough to run everywhere, and when it fails, its failure is at least legible.(1)

Not every deployment can afford a deep network. In many settings a lightweight, interpretable detector is what is actually wanted: one that runs on a central processing unit, needs no training data beyond a handful of images, and exposes the reason for its decision. Frequency-domain features fit that description. Generated images carry spectral fingerprints left by the upsampling layers of generative networks, and these fingerprints are visible in the Fourier and discrete-cosine transforms of an image.[13,14] The idea behind frequency-domain detection is to summarize those fingerprints with a few dozen scalar features and hand them to a simple linear classifier.[14,15] The approach is cheap and auditable, which is why it remains attractive despite the dominance of deep methods.(16,17)

The frequency approach also spans a wide design space, and the literature has explored it in several directions. Some methods fuse spatial and spectral features to recover the information that either domain alone misses.[18,19] Others work entirely in the spectral domain, using wavelet decompositions or training-free measurements of the diffusion process itself.[17,20] The common thread is the hope that a compact spectral signature exists and can be read off without a learned model. What most of these studies do not isolate is the specific question of cross-generator transfer, because they typically report performance on the same generators used in training. That is the gap the present study fills.

The design chosen here is deliberately minimal. Rather than propose yet another feature set, the study takes two standard, off-the-shelf frequency representations, pairs them with the simplest classifier that can exploit them, and asks a single, sharply defined question: how much accuracy is lost when the generator changes. The answer is reported as a full transfer matrix rather than a single number, because, as the results will show, the loss is a property of the specific generator pair and a single number would conceal it. The intent is to produce a baseline that other, more elaborate frequency and hybrid methods can be compared against, and a number that practitioners can use directly when they decide whether a lightweight detector is good

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

enough for their problem.

What is not settled is whether a frequency-domain detector generalizes across generators, and if it fails, by how much. The spectral fingerprints that the features exploit are known to be generator-specific, which suggests that a frequency detector trained on one generator may learn that generator's signature rather than a universal property of synthetic imagery.[13,19] If so, its cross-generator performance would fall, and the magnitude of the fall would tell practitioners how much to trust such a detector in the open world. The objective of this study was therefore to evaluate the cross-generator generalization of lightweight detectors that use frequency-domain features, measuring the drop in accuracy and area under the curve when a detector trained on one generator is tested on generators excluded from training.

METHODS

DESIGN

A diagnostic study was conducted in which frequency-domain features were extracted from real photographs and from machine-generated images, a lightweight linear classifier was trained on one generator, and its performance was measured both in-domain and on held-out generators. The design follows the STARD framework for diagnostic accuracy studies, with the generator family as the condition whose transfer is being tested. The index test is the frequency-domain logistic classifier; the reference standard is the known provenance of each image, which the benchmark records directly and without ambiguity. The analysis is fully reproducible from the public benchmark and a fixed random seed, and no human judgment enters the pipeline, so the reference standard is exact rather than an approximation.

DATA

Images were taken from GenImage, a million-scale benchmark that pairs real photographs from ImageNet with images generated by eight distinct models.(9) Three generators were selected to span a range of different architectures: BigGAN, a class-conditional generative adversarial network; GLIDE, a diffusion model; and VQDM, a vector-quantized diffusion model. For each generator the benchmark provides a folder of real photographs and a separate folder of generated images matched to the same class structure. A fixed-seed random sample of 500 real photographs and 500, 500, and 500 generated images per generator was drawn, yielding a balanced working set of two thousand images (Table 1). All images were converted to grayscale and then resized to a common resolution of 128 by 128 pixels before feature extraction.

Table 1. Composition of the working sample by class and generator.

Class	Images
Real photographs	500
biggan	500
glide	500
vqdm	500
Feature dimensions	59

Note: A fixed random seed was used for the sample; all images were resized to 128 by 128 pixels.

FREQUENCY-DOMAIN FEATURES

Two families of frequency-domain features were computed for each image, producing a 59-dimensional vector. The first family summarizes the Fourier magnitude spectrum. The two-dimensional Fourier transform was computed, the logarithm of its magnitude was averaged over concentric rings centered on the zero frequency, and the resulting radial profile of 32 bins was normalized to sum to one. This profile captures how energy is distributed across spatial frequencies, and it is the natural place where the periodic grid artifacts of upsampling layers appear.(13) The second family summarizes the discrete-cosine transform, the representation used by image

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

compression. The logarithm of the cosine magnitudes was divided into low, middle, and high frequency bands whose mean energies were recorded, and the remaining alternating-current coefficients were summarized with a 24-bin histogram. Together the two families describe both the overall spectral envelope and the fine structure of the coefficient distribution, and both have been used before in synthetic-image detection.[14,15] All features were computed with standard library transforms and required no learned model, which is the point of the design.

The choice of these two transforms is not arbitrary. The Fourier spectrum is where periodic artifacts, such as the checkerboard patterns left by transposed convolutions, show up as localized peaks, and a radial profile is a compact summary of that periodicity. The discrete-cosine transform is the basis of JPEG and other compression schemes, and its coefficient distribution is a long-established descriptor of natural image statistics, which makes departures from the natural distribution a natural signal for detection. The two are partly redundant, since both are unitary frequency transforms of the same image, but they weight the spectrum differently and their combination is standard practice in the frequency-detection literature.[14,16] No spatial features were included, deliberately, so that the study isolates the frequency signal and can attribute any cross-generator failure to that signal rather than to the classifier or to a confound.

CLASSIFIER AND EVALUATION PROTOCOL

The classifier was logistic regression, chosen because it is the lightest model that can learn a linear separation from a handful of features and because its coefficients remain directly interpretable. A separate detector was trained for each generator as the training domain: it saw the real photographs and that generator's images only, in a balanced half of the in-domain data. The detector was then evaluated in three ways. The held-out half of the in-domain data measured its baseline performance; the two other generators' images, each paired with real photographs, measured its cross-generator performance. Accuracy and the area under the receiver operating characteristic curve were recorded for every train-test pair.

Two design decisions deserve explicit statement because they bear on how the numbers should be read. The training and evaluation sets were balanced between real and generated images, so the chance baseline is 0.5 and every accuracy can be judged directly against that floor. The classifier was given no capacity beyond a linear decision boundary, which means any success it achieves must come from the features rather than from model complexity; this is deliberate, since the question is about the features, not about how much a powerful model can be squeezed. The average in-domain accuracy, the average cross-generator accuracy, and their difference summarize how much the detectors lose when they leave their training generator. Because the features are deterministic and the classifier and sampling are seeded, the entire pipeline is reproducible, and it was executed twice to confirm identical results.

RESULTS

IN-DOMAIN PERFORMANCE

A detector trained on a single generator distinguished that generator's images from real photographs with notably mixed success (Table 2). In-domain accuracy ranged from 0.68 for BigGAN to 0.92 for GLIDE, with VQDM in between at 0.71. The mean in-domain accuracy across the three generators was 0.77. The area under the curve followed the same ordering, rising from 0.76 for BigGAN to 0.97 for GLIDE. The spread is itself informative: the same features separate GLIDE images from photographs almost perfectly, yet leave a BigGAN detector near the point where it would be wrong one time in three.

The ordering across generators is not explained by any single property of the models. GLIDE, a diffusion model, was the easiest to detect, while BigGAN, a generative adversarial network, was the hardest, and VQDM, another diffusion-based model, sat between them. This suggests that the strength of the spectral fingerprint left by a generator depends on details of its training and sampling rather than on its family, a point that matters when one tries to predict which new generators a detector will handle. The practical reading is that even before considering generalization, the frequency approach is only as good as the generator it happens to face, and a practitioner who assumes a uniform baseline across generators will be misled by a detector that is excellent for one model and mediocre for another.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

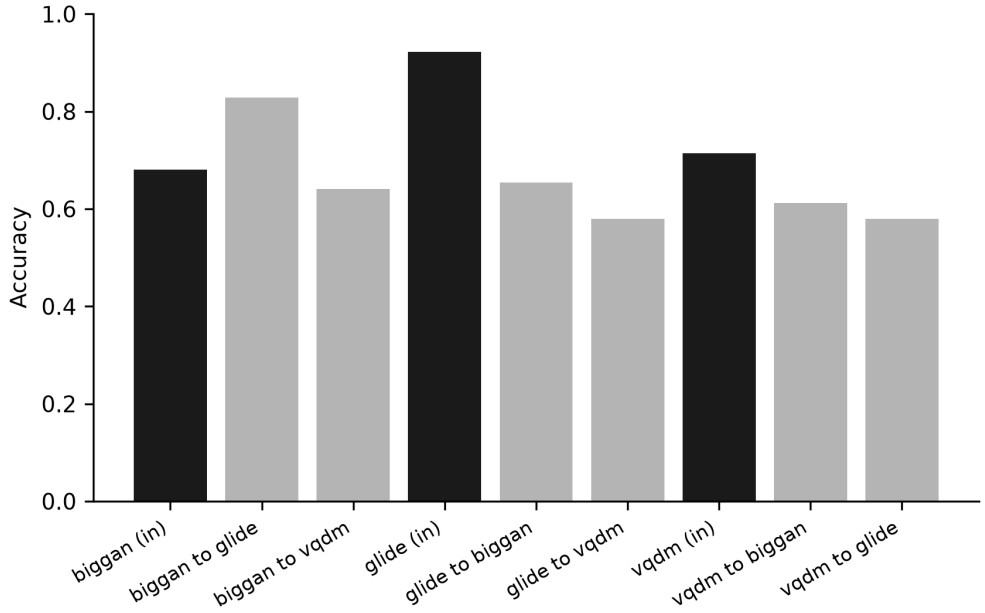
Table 2. Accuracy and area under the curve of frequency-domain detectors by train and test generator.

Train generator	Test set	Accuracy	AUC
biggan	in-domain	0.68	0.758
biggan	glide	0.829	0.941
biggan	vqdm	0.641	0.684
glide	in-domain	0.922	0.974
glide	biggan	0.654	0.74
glide	vqdm	0.579	0.565
vqdm	in-domain	0.714	0.782
vqdm	biggan	0.612	0.673
vqdm	glide	0.58	0.618

CROSS-GENERATOR PERFORMANCE

When a detector was tested on generators it had not seen, performance dropped in most but not all cases (Figure 1). The mean cross-generator accuracy was 0.65, against 0.77 in-domain, an average drop of 0.12 percentage points. The drop was uneven across the six train-test transfers. The GLIDE detector, the strongest in its own domain at 0.92, fell to 0.65 on BigGAN and 0.58 on VQDM. The BigGAN detector, by contrast, transferred to GLIDE at 0.83, an accuracy above its own in-domain score, although its transfer to VQDM was only 0.64. The weakest transfer of all was VQDM to GLIDE at 0.58, which left the detector barely above chance.

Figure 1. Accuracy of each frequency-domain detector in-domain and on held-out generators.



The area under the curve told the same story as accuracy. Cross-generator areas ranged from 0.56 at the low end to 0.94 at the high end, and in most transfers the detector lost the confidence margin that a usable classifier needs. The pattern that emerges from both metrics is that frequency-domain detectors are tuned to the spectral signature of their training generator, and that signature does not carry reliably to generators built on different architectures.

A useful way to condense the whole matrix is to look at the two directions of every generator pair. BigGAN and GLIDE transferred to each other at 0.83 and 0.65, an asymmetry in which the easier-to-detect GLIDE pulled the numbers in opposite directions. BigGAN and VQDM transferred at 0.64 and 0.61, and GLIDE and VQDM at 0.58 and 0.58. In every pair the two directions disagreed, which is exactly what a generator-specific fingerprint predicts: each detector inherits its training generator's signature, and each test generator presents its own, so the two directions are different questions and get different answers.

The magnitude of the loss is the quantity a practitioner most needs. The average drop of

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

0.12 percentage points is large enough to matter in any application where a missed synthetic image or a falsely flagged photograph carries a cost, and it is the kind of number that should be quoted alongside a detector's headline accuracy whenever the detector is to be used outside its training distribution. That said, the drop is also bounded: the detectors rarely collapsed entirely, and most cross-generator accuracies remained above chance, which means the frequency signal is not wholly absent on unseen generators, only weakened. A reader should hold both facts at once: the transfer penalty is real and large, but it does not reduce the detectors to uselessness.

DISCUSSION

MAIN FINDINGS

Frequency-domain detectors generalize across generators poorly, losing twelve percentage points of accuracy on average and, in one transfer, collapsing to the level of random guessing. A detector built from Fourier and cosine features works as a screen within the generator it was trained on, and a strong one for some generators, but it cannot be trusted to recognize images produced by a generator it has not seen before. The second finding is that the loss is not uniform: it depends on which generator was trained on and which generator is being tested, which means the cross-generator performance of these detectors is a property of the specific generator pair rather than a single fixed penalty.

The third finding, less obvious but equally important, is that the in-domain baseline itself is only moderate. An average in-domain accuracy of 0.77 means that even in the most favorable case, where the detector is asked only about the generator it was trained on, it is wrong roughly one time in four. This qualifies any claim about the frequency approach: it is not a strong detector that happens to generalize poorly, but a modest detector that generalizes worse. The distinction matters because a practitioner weighing the frequency approach against a deep detector is choosing between a weak-but-cheap and a strong-but-expensive option, and the results here put a number on both halves of that trade-off.

COMPARISON WITH PRIOR WORK

The result is consistent with the mechanism that the frequency-detection literature has proposed from the beginning. Durall and colleagues showed that generative networks leave distinct spectral artifacts in their outputs because their upsampling layers fail to reproduce natural spectral distributions, and they warned that these artifacts are tied to the specific architecture of each generator.(13) Frank and colleagues built their detector on those same artifacts and reported strong in-domain results, but their evaluation did not push the detectors onto unseen generators.(14) The present study closes that gap: it shows that the artifacts are indeed exploitable within a generator and that the exploit transfers weakly, exactly as the mechanism predicts.

The magnitude of the drop is comparable to, though milder than, the degradation reported for deep detectors on the GenImage benchmark, where some detectors lose more than twenty points of accuracy across generators.(9) Frequency features are no worse than deep features in this respect, and they carry the advantage of requiring no training data and no learned model. But they are also no better, which matters because the hope behind lightweight frequency detectors was that a cheap, generator-agnostic signal could be found. The results here suggest that the signal is generator-specific, and that the cheapness of the detector buys speed and interpretability rather than robustness.(7,8)

The diffusion-specific literature points in the same direction. Corvi and colleagues showed that detectors built for one diffusion model must be recalibrated for another, and Bammey and colleagues found that a detector aimed at one family of diffusion models does not transfer cleanly to others, both of which are the same phenomenon observed here in a smaller, fully controlled setting.[21,22] Ricker and colleagues reached the complementary conclusion that diffusion artifacts are subtle enough that specialized, rather than generic, detectors are needed.(23) The present study adds to that body of work a clean estimate of the transfer penalty for the frequency approach, obtained without the confounds of architecture choice or large-scale training, and it finds that the penalty is substantial even between three well-known generators.(10)

The one apparent exception deserves scrutiny. The BigGAN detector transferred to GLIDE with higher accuracy than it achieved in its own domain. This does not contradict the overall

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

finding; it reflects that GLIDE images are simply easier to separate from photographs, as the GLIDE detector's own near-perfect in-domain score shows. A generator whose images are far from natural photographs is easy for any detector, trained or not, to flag, and that ease can mask the underlying failure to generalize. The honest summary is not that BigGAN generalized to GLIDE, but that GLIDE is an easy target, a distinction that matters when a practitioner interprets a single favorable number.

WHY THE FEATURES FAIL TO TRANSFER

The failure has a concrete cause. The radial Fourier profile and the cosine-coefficient histogram describe the spectral fingerprint of a generator, and different generators leave different fingerprints because their upsampling, normalization, and noise schedules differ.[13,24] A logistic regression trained on one fingerprint learns a decision boundary that separates that fingerprint from natural spectra, and nothing in its training obliges it to separate a different fingerprint. The features are therefore best understood as a weak, generator-specific fingerprinting scheme rather than as a general measure of synthetic-ness. Learned detectors mitigate this by training on many generators, which is why they transfer better, but they pay for that robustness with data and compute that the frequency approach avoids.[5,9] The subpixel and dual-domain methods that have recently appeared are, in effect, attempts to find a fingerprint that sits below the generator-specific surface, and their mixed results are another sign that the surface-level spectral signal does not carry far.(19,25)

The heterogeneity across pairs reinforces this reading. VQDM and GLIDE, both diffusion-based, transferred to each other at the lowest accuracies in the study, while the GAN-based BigGAN transferred to GLIDE more readily. This ordering is the opposite of what a naive architectural-distance story would predict, which is a reminder that spectral fingerprints do not vary along a single architectural axis. A practitioner cannot assume that two diffusion models will leave similar artifacts, or that a GAN and a diffusion model will leave maximally different ones.

The heterogeneity also has a methodological lesson. Because the transfer penalty is a function of the pair, a single cross-generator comparison is not enough to characterize a detector's generalization; it can flatter or damn the detector depending on the particular pair chosen. Studies that report one held-out generator are therefore measuring a point on a surface, not the surface itself. The three-by-three matrix reported here, small as it is, is the minimum that allows the average and the spread to be separated, and it is the spread across pairs that a deployment decision should actually care about most. This argues for evaluating detection methods on generator matrices rather than single transfers, a practice the benchmark literature has moved toward but that is still not universal.(7,9)

PRACTICAL IMPLICATIONS

For someone who must deploy a detector today, the results support a narrow but honest recommendation. A frequency-domain detector is a defensible first-line screen when a learned detector is unavailable, when the compute budget forbids a deep network, or when the generator in question is known and will remain stable. It will flag the worst synthetic images cheaply and transparently, and its coefficients can be inspected to show what spectral property drove a decision, which has forensic value in itself.[15,16] But it should not be deployed against an open-ended stream of images from unknown generators, where its cross-generator error, twelve points on average and worse in some pairs, would produce both missed detections and false alarms at rates that most applications cannot accept.(2)

The recommendation can be stated more precisely for the three applications named at the outset. For the newsroom verifying one image, the frequency detector is the right first tool precisely because it is interpretable, and its cross-generator weakness matters less when the stakes are a single editorial decision and a human is in the loop. For the platform screening at scale, the detector is useful as a triage layer that routes suspicious images to a stronger, more expensive check, and its high false-alarm rate at the margin is tolerable so long as it never acts alone. For the forensic examiner, the detector's failure modes are themselves evidence, because a spectral fingerprint that does not match any known generator is a signal worth recording even when it cannot be interpreted.(26) In none of these roles is the frequency detector a complete answer, and the results here make clear that it should not be mistaken for one.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

There is also a reporting lesson for the field. A detection paper that reports only in-domain accuracy is reporting the best case, and a reader cannot infer from it how the detector will behave in the open world. The results here argue that cross-generator accuracy, stated as a matrix rather than a single transfer, should be reported alongside in-domain accuracy as a matter of course, in the same way that diagnostic studies report both internal and external validation. The twelve-point gap measured here is the kind of number that changes a deployment decision, and it is invisible in the headline accuracy that most papers foreground. Making the gap visible would do more for the practical usefulness of detection research than any single algorithmic improvement. (3,10)

Where robustness is required, the implications point toward generator-diverse training. The literature on universal fake-image detection has converged on the view that exposing a detector to many generators, or to a feature representation that abstracts over them, is what buys generalization.[27,28] The frequency approach can still contribute in that setting, not as the detector itself but as one input among several, since spectral features capture information that spatial features miss and the two are known to be complementary.(19,29,30) The contribution of this study is a quantified baseline for how far the frequency signal alone can go, against which hybrid and learned detectors can be measured. A related use, which the frequency fingerprint actually suits, is attribution: because the fingerprint is generator-specific, it is better at saying which generator produced an image than at saying whether an image is synthetic at all, and attribution systems have begun to exploit exactly this property.(24,26)

The study has limitations that follow from its design. Three generators are a small slice of the modern range of generative models, and the specific triple chosen here cannot be assumed to represent all generator architectures; a replication over the remaining generators of GenImage, and over diffusion and autoregressive models released after it, would test the stability of the twelve-point average. The features are a particular choice from a large design space, and other frequency representations, such as the wavelet transforms used in recent work, might transfer differently.(17) The real photographs and the generated images are drawn from the same ImageNet class structure, which controls content but also means the detectors were never asked to work on out-of-distribution content. The classifier is deliberately simple, and a more powerful lightweight model, such as a small gradient-boosted ensemble, might extract more from the same features. Each limitation suggests a concrete extension rather than undermining the conclusion: more generators, more feature families, out-of-distribution content, and a small set of stronger lightweight classifiers.

Two further limits concern resolution and the absence of an adversary. All images were downsampled to a common resolution before feature extraction, which discards the finest high-frequency detail where some generator fingerprints live; a comparison across native resolutions would show whether the transfer penalty shrinks when the full spectrum is available. And the study evaluated the detectors against images as produced, without the fingerprint-removal attacks that a real deployment would face, so the accuracies reported here are upper bounds on what a determined adversary would allow.(11) These gaps are easy to state and equally easy to close in follow-up work, and neither changes the qualitative conclusion that the frequency signal is generator-specific. The overall design, moreover, measures a fixed snapshot of three generators, and the field moves quickly enough that the specific numbers here will age; what should persist is the transfer-matrix method and the general finding that the penalty is large and pair-dependent.

CONCLUSIONS

Frequency-domain features supported lightweight detectors that separated machine-generated images from real photographs within a single generator, though with accuracy that varied widely from one generator to the next. Across generators, accuracy fell by an average of twelve percentage points, and one transfer collapsed to chance. The detectors exploit generator-specific spectral fingerprints rather than a universal property of synthetic imagery, which is why they transfer poorly and why their cross-generator behavior varies from pair to pair. Frequency-domain detection remains a cheap, interpretable first-line screen for a known generator, but it is not a substitute for generator-diverse training when robustness to unseen generators is required.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

REFERENCES

Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A Survey of face manipulation and fake detection. Information Fusion. 2020;64:131-148. doi: 10.1016/j.inffus.2020.06.014
Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell. 2023;53(4):3974-4026. doi: 10.1007/s10489-022-03766-z
Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. IEEE Access. 2022;10:25494-25513. doi: 10.1109/access.2022.3154404
Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, et al. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding. 2022;223:103525. doi: 10.1016/j.cviu.2022.103525
Singh A, Jain PC, Singh A. Beyond Clean Benchmarks: A Reproducible Severity-Sweep Evaluation of AI-Generated Image Detector Robustness. IRE Journals. 2026;10(3):3967-3974. doi: 10.64388/irev10i3-1723524
Wang S-Y. CNN-generated images are surprisingly easy to spot.. for now [Internet]. 2020. Available from: https://arxiv.org/abs/1912.11035
Coccomini DA, Caldelli R, Falchi F, Gennaro C. On the Generalization of Deep Learning Models in Video Deepfake Detection. J Imaging. 2023;9(5):89. doi: 10.3390/jimaging9050089
Meng Z, Peng B, Dong J, Tan T, Cheng H. Artifact feature purification for cross-domain detection of AI-generated images. Computer Vision and Image Understanding. 2024;247:104078. doi: 10.1016/j.cviu.2024.104078
Zhu M. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image [Internet]. 2023. Available from: https://arxiv.org/abs/2306.08571
Mahara A, Rishe N. Methods and trends in detecting AI-generated images: A comprehensive review. Computer Science Review. 2026;60:100908. doi: 10.1016/j.cosrev.2026.100908
Lu Y, Liu J, Zhang R. GANFR: GAN fingerprint removal network for image anti-forensics. Knowledge-Based Systems. 2025;327:114134. doi: 10.1016/j.knosys.2025.114134
Guarnera L, Giudice O, Battiato S. Mastering Deepfake Detection: A Cutting-edge Approach to Distinguish GAN and Diffusion-model Images. ACM Trans Multimedia Comput Commun Appl. 2024;20(11):1-24. doi: 10.1145/3652027
Durall R, Keuper M, Keuper J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:7887-7896. doi: 10.1109/cvpr42600.2020.00791
Frank J. Leveraging Frequency Analysis for Deep Fake Image Recognition [Internet]. 2020. Available from: https://arxiv.org/abs/2003.08685
Giudice O, Guarnera L, Battiato S. Fighting Deepfakes by Detecting GAN DCT Anomalies. J Imaging. 2021;7(8):128. doi: 10.3390/jimaging7080128
Li L, Lu T, Song J, Cheng K. SIDWA: Synthetic Image Detection Based on Discrete Wavelet Transform Stem and Deformable Sliding Window Cross-Attention. Electronics. 2026;15(4):891. doi: 10.3390/electronics15040891
Marini N, Berretti S, Caldelli R. Wavelet-based features to improve cross-forgery generalization in deepfake detection. Front Imaging. 2026;5:1845480. doi: 10.3389/fimag.2026.1845480
V YK, Dr DJV. Spatial-Frequency Fusion for Robust Detection of AI- Generated Images across Unseen Generators. Int J Innov Res Eng. 2026:01. doi: 10.59256/ijire.20260705001
Yan J, Li Z, Wang F, He Z, Fu Z. Dual Frequency Branch Framework With Reconstructed Sliding Windows Attention for AI-Generated Image Detection. IEEE TransInformForensic Secur. 2026;21:2313-2325. doi: 10.1109/tifs.2026.3664000
Choi Y. HFI: A unified framework for training-free detection and implicit watermarking of latent diffusion images [Internet]. 2024. Available from: https://arxiv.org/abs/2412.20704
Corvi R. On the detection of synthetic images generated by diffusion models [Internet]. 2023. Available from: https://arxiv.org/abs/2211.00680
Bammey Q. Synthbuster: Towards Detection of Diffusion Model Generated Images. IEEE Open J Signal Process. 2024;5:1-9. doi: 10.1109/ojsp.2023.3337714
Ricker J. Towards the detection of diffusion model deepfakes [Internet]. 2023. Available from: https://arxiv.org/abs/2210.14571
Liu C, Zhu T, Zhao Y, Zhang J, Zhou W. Disentangling different levels of GAN fingerprints for task-specific forensics. Computer Standards & Interfaces. 2024;89:103825. doi: 10.1016/j.csi.2023.103825
Lu P, Liang K, Liu X, Lu X, Li A. SDA-Det: A subpixel-aware dual-domain artifact detector for AI-generated image detection. Expert Systems with Applications. 2026;322:132304. doi: 10.1016/j.eswa.2026.132304
Namani Y, Reghioua I, Bendiab G, Labiod MA, Shiaeles S. DeepGuard: Identification and Attribution of AI-Generated Synthetic Images. Electronics. 2025;14(4):665. doi: 10.3390/electronics14040665
Ojha U. Towards universal fake image detectors that generalize across generative models [Internet]. 2023. Available from: https://arxiv.org/abs/2302.10174
Willi M, Mathys M, Graber M. Synthetic image detection with CLIP: Understanding and assessing predictive cues. Array. 2026;32:101222. doi: 10.1016/j.array.2026.101222
Yang T, Zhang Q. Dual-stream FDS-CRF net: frequency-spatial constrained attention for cross-generative model image detection. SIViP. 2025;19(11):932. doi: 10.1007/s11760-025-04486-w
Li J, Liang Y, Ye W, Wang P, Lei T, Zhang G, et al. Dual-branch AIGC image detection leveraging integrated semantic and frequency-domain features. Computer Vision and Image Understanding. 2026;271:104893. doi:

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

10.1016/j.cviu.2026.104893

https://doi.org/10.62486/978-9915-9966-3-9.Ch04