

Forensic Detection of Generated MRI Imagery Using Autoregressive Modeling and Frequency Analysis

Arpan Mahara, Naphtali Rishé, and Malek Adjouadi

Florida International University

Miami, FL 33199, USA

amaha038@fiu.edu, rishen@cs.fiu.edu, adjouadi@fiu.edu

Abstract

Magnetic Resonance Imaging (MRI) is a widely adopted technology for acquiring detailed internal images of the human body for diagnostic purposes. Given its sensitive nature and diagnostic importance, the integrity of MRI imagery must be preserved in both clinical and research settings. However, the rapid advancement of generative AI technologies poses risks of adversarial manipulation, potentially compromising diagnostic accuracy and jeopardizing the entire medical imaging pipeline. To address this emerging threat, we present a systematic investigation of state-of-the-art generative methods, including Generative Adversarial Networks (GANs), diffusion models, and autoregressive models, on MRI imagery, providing a comparative analysis of their generative performance. We introduce MRI-Forensics, a curated benchmark dataset, and show that generative manipulations leave distinct and quantifiable signatures in the frequency domain. With this assertion, we develop a new forensic detection framework that combines Autoregressive Image Models and Discrete Wavelet Transform (AIM-DWT) analysis that reliably detects AI-generated manipulations. By integrating frequency-based decomposition with autoregressive visual modeling, we demonstrate that AIM-DWT effectively extracts unique generative fingerprints from synthesized MRI images. In extensive evaluations using MRI-Forensics and an independent brain MRI dataset, experimental results demonstrate the efficacy of our approach, highlighting its potential to ensure diagnostic accuracy and patient trust in medical imaging in the era of generative AI. The code, pretrained weights for reproducibility, and the MRI-Forensics dataset are available at <https://github.com/amaha7984/AIM-DWT>.

1. Introduction

MRI represents a significant advancement in medical imaging, offering detailed insights into the internal structures of

the human body. Unlike X-rays, which have limited utility in visualizing soft tissues (e.g., muscles, ligaments, organs), and Computed Tomography (CT) or Positron Emission Tomography (PET) scans, which involve exposure to potentially harmful ionizing radiation, MRI provides highly accurate images with minimal or no health risks. Consequently, MRI remains the preferred modality for critical diagnostic evaluations. However, one persistent limitation of MRI is patient movement during scans, which can introduce artifacts and complicate accurate diagnoses. To address this issue, deep learning methods have been incorporated into MRI reconstruction processes to mitigate motion artifacts, thus enhancing diagnostic accuracy [18].

The potential for intentional image manipulation exacerbates concerns about diagnostic accuracy. Rapid advancements in generative artificial intelligence (AI) pose the risk of malicious alterations to MRI imagery that might evade detection by human observers or existing diagnostic systems. Generative AI technologies, including GANs [15], Diffusion Models [39], and Autoregressive Models [41], although primarily intended for benevolent purposes, can be exploited for unethical manipulation. Prior studies have explored the generation of synthetic medical imagery to address data scarcity or restrictive data-sharing policies. For example, Pinaya et al. [32] synthesized 3D brain MRI imagery using latent diffusion, achieving better performance compared to GANs. Similarly, Mueller-Franzes et al. [28] proposed Medfusion, a latent DDPM model demonstrating enhanced performance over GANs in synthesizing Color Fundus, X-ray, and colorectal imagery. Dhinagar et al. [8] trained latent diffusion models for Alzheimer’s disease diagnoses through brain MRI generation. Additionally, Chang et al. [4] employed probabilistic diffusion models for generating high-resolution MRI images from low-resolution inputs using the BraTS2020 T2-FLAIR dataset of brain tumors. Studies by Lai et al. [22] and Goncalves et al. [14] demonstrated the effectiveness of StyleGANv2-ADA [20] in generating brain and abdominal MRI images, respectively. Other approaches include the use of Single

Natural Image GANs (SinGANs) by Xu et al. [45] for synthesizing prostate MRI images, and advanced stable diffusion models by Saeed et al. [37] for generating multi-sequence prostate MRI conditioned on text. Recent applications of autoregressive models include generating MRI sequences [23] and synthesizing contrast-enhanced brain tumor images from non-contrast scans [16]. Given these advancements and their associated adversarial vulnerabilities, it is crucial to develop forensic detection methods to differentiate synthesized from authentic MRI imagery. Although AI-driven manipulations have been reported in healthcare, no incidents specifically involving MRI imagery manipulation have yet been documented. Nevertheless, the growing prominence of generative AI indicates imminent risks that must be addressed.

Due to the challenges in accessing MRI datasets, restrictions on publicly available data, and the lack of existing forensic detection methods, we present the following key contributions:

- **MRI-Forensics dataset:** We introduce MRI-Forensics, a new forensic benchmark dataset comprising 6,306 authentic and 37,836 synthetic prostate MRI images, generated using six state-of-the-art generative models. The dataset is accompanied by extensive evaluations of generative performance, including frequency-domain analysis, which reveal model-specific spectral fingerprints.
- **AIM-DWT framework:** We develop AIM-DWT, a novel forensic detection framework that combines autoregressive modeling with Discrete Wavelet Transform (DWT) analysis to extract subtle generative fingerprints from synthesized MRI imagery.
- **Cross-model robustness:** Through cross-model generalizability evaluations, we demonstrate that AIM-DWT outperforms existing forensic detection methods when tested across unseen generative model families.
- **Cross-dataset generalization:** Extending our evaluation to a publicly available brain MRI dataset by generating synthetic brain MRI images, we show that AIM-DWT, trained solely on prostate MRI data, generalizes effectively across datasets in detecting synthesized brain MRI imagery without retraining.

2. Related Work

As mentioned previously, while several generative methods have been applied to synthesize MRI imagery, there is no prior work specifically addressing the forensic detection of synthesized MRI images. In the general forensic detection domain, several methods based on deep learning principles have recently shown success. In the early era of GAN development and the need to address their adversarial effects, Marra et al. [26] discovered that GANs leave distinctive fingerprints in images. These can be extracted by applying a denoising filter to the generated image, computing the resid-

ual difference, and averaging over multiple samples. This concept was extended by Yu et al. [48], who used an autoencoder instead of a denoising filter to extract GAN-specific fingerprints for detecting generated content. More recently, Corvi et al. [7] adapted this approach to detect images generated by Diffusion models by extracting their unique fingerprints. While numerous methods exist for detecting generated images, they often lack generalizability across different generators. To address this, Wang et al. [43] proposed a CNN-based detector trained on ProGAN-generated images. They showed that pretraining on large datasets and applying post-processing strategies such as JPEG compression and Gaussian blur can improve the detection of images generated by more advanced GANs. However, with the rapid rise of diffusion models, Wang et al. [44] demonstrated that prior methods failed to detect diffusion-generated content effectively. To mitigate this, they introduced the DIRE representation, computed as the difference between an input image and its reconstruction via DDIM [40], which substantially improves detection performance for diffusion-generated images.

Recent studies [2, 10, 47] have shown that while advanced deep learning methods are effective in many cases, they can fail when synthetic images closely resemble real ones. To address this limitation, frequency analysis methods such as the Fourier Transform [5], Discrete Wavelet Transform (DWT) [25], and Discrete Cosine Transform (DCT) [1] have been proposed to extract subtle frequency-based noise signatures for forensic discrimination. A recent forensic survey [24] highlights that vision-language models like CLIP [34] are among the most powerful approaches in the current forensic landscape. For instance, Ojha et al. utilized CLIP’s ViT [9] to extract visual features and performed classification using cosine similarity in the feature space. Similarly, AIDE [46] and RINE [21] are recent advanced forensic frameworks that leverage CLIP’s ViT architecture. AIDE combines frequency-aware low-level features extracted via DCT [1] with semantic features from OpenCLIP, fusing them through an MLP for classification. In contrast, RINE concatenates CLS tokens from multiple ViT [9] layers and refines them using a Trainable Importance Estimator (TIE), applying supervised contrastive and binary classification losses.

In line with these advancements and recognizing the lack of forensic detection research in MRI imagery, the present study proposes a novel forensic detection framework that integrates autoregressive modeling with frequency analysis. The details of the proposed method are presented in the following section.

3. The Proposed Approach

The contributions of the present study are twofold: (1) the introduction of a new MRI-based forensic dataset, and

(2) the development of a novel forensic detection framework utilizing advanced autoregressive models. The subsequent subsections detail the preparation of the forensic dataset, provide relevant background on autoregressive methods, and present the architecture of our proposed detection methodology.

3.1. MRI-Forensics Dataset

We construct a dataset using selected state-of-the-art generative models from three predominant model families: Generative Adversarial Networks (GANs), Diffusion Models, and Autoregressive Models, under an unconditional generation setup. We utilize an MRI dataset (owned by us), primarily focused on prostate imagery, originally stored in DICOM format. To obtain a standardized set of images for generative model training, we converted the DICOM files to high-quality JPEG images using the `dcmj2pnm` utility from the `DCMTK` toolkit [11]. To ensure a clean MRI dataset, we applied image hashing based on image pixel contents to remove duplicates, resulting in 6,306 unique images. These images were subsequently resized to 256×256 size to maintain consistency and generalizable training. From the GAN family, we selected ProGAN [19] and StyleGANv2-ADA [20], due to their proven effectiveness and widespread adoption in generative forensics. ProGAN [19] incrementally trains GANs by progressively adding layers to generator and discriminator networks, facilitating the stability and quality of generated images. Similarly, StyleGANv2-ADA [20] improves data efficiency through adaptive discriminator augmentation, enabling stable training even with limited datasets.

For diffusion-based generative modeling, we adopt DiT [31], which integrates diffusion models with Vision Transformers (ViT) [9], replacing the conventional U-Net architecture [35]. DiT leverages transformer-based self-attention mechanisms to efficiently capture long-range dependencies in image data, improving both the scalability and the fidelity of generated images. We specifically selected DiT variants with XL/2 and L/2 architectures due to their generative performance in our experimental evaluation.

Autoregressive generative models have increasingly gained attention for their computational efficiency and robust generative capabilities. We utilize the recently proposed Randomized Autoregressive Visual Generation (RAR) model [49], which uniquely combines autoregressive generation with randomized token sampling strategies. This randomization allows the model to efficiently generate high-quality images by reducing memory consumption and accelerating inference speed. We selected the XXL and XL backbone architectures for their suitability in synthesizing MRI imagery.

In total, we generated 37,836 synthetic MRI images using these six model configurations. Consequently, the MRI-

Forensics Dataset comprises 44,142 images: 6,306 authentic MRI images and 37,836 synthetic images. A comprehensive experimental analysis of these generative models is presented in Section 4.

3.2. Background

Our method, termed AIM-DWT, is based on the recent autoregressive vision frameworks AIM [12] and AIMv2 [13], and the frequency analysis method, Discrete Wavelet Transform (DWT) [25]. AIM formulates image modeling as a causal token prediction task. Unlike masked reconstruction, AIM tokenizes input images into non-overlapping patches and processes them using a GPT-style Transformer decoder [33]. AIMv2 [13] introduces a scalable autoregressive framework for multimodal modeling, using vision and text encoders followed by a causal decoder. The AIMv2 vision encoder adopts a ViT backbone [9] with patch tokenization and prefix attention masking, allowing strong image representation learning without global tokens. In this work, we adapt the vision encoder architecture from AIMv2 for an unimodal image classification task, using its design as a strong backbone for subband processing.

Similarly, 2D DWT facilitates multi-resolution image decomposition, which is essential for identifying frequency-specific features. Mathematically, the decomposition at scale j is represented as:

$$x = \sum_k c_{j,k} \phi_{j,k}(t) + \sum_{j' \geq j} \sum_k d_{j',k} \psi_{j',k}(t), \quad (1)$$

where $c_{j,k}$ represent low-frequency approximation coefficients, $d_{j',k}$ are high-frequency detail coefficients, and $\phi_{j,k}(t), \psi_{j',k}(t)$ denote scaling and wavelet functions respectively.

3.3. Method: AIM-DWT

Our proposed AIM-DWT framework, illustrated in Figure 1, employs the 2D DWT [25] using the Haar wavelet at decomposition level 1 with symmetric padding. DWT is selected over DCT and FFT due to its ability to localize features in both spatial and frequency domains [36], making it suitable for capturing fine-grained and localized anomalies in medical imagery. This DWT process decomposes the input MRI image into four distinct frequency subbands: a low-frequency approximation band (LL) and three high-frequency detail bands, LH (horizontal), HL (vertical), and HH (diagonal). Each subband is then patchified (as shown in Figure 1, with patches depicted as color-coded rounded rectangles) to enable token-level modeling and feature extraction. These patch embeddings from each subband are independently processed through a shared prefix-attention vision encoder based on the AIMv2 architecture [13]. We hypothesize that this isolated subband processing enhances

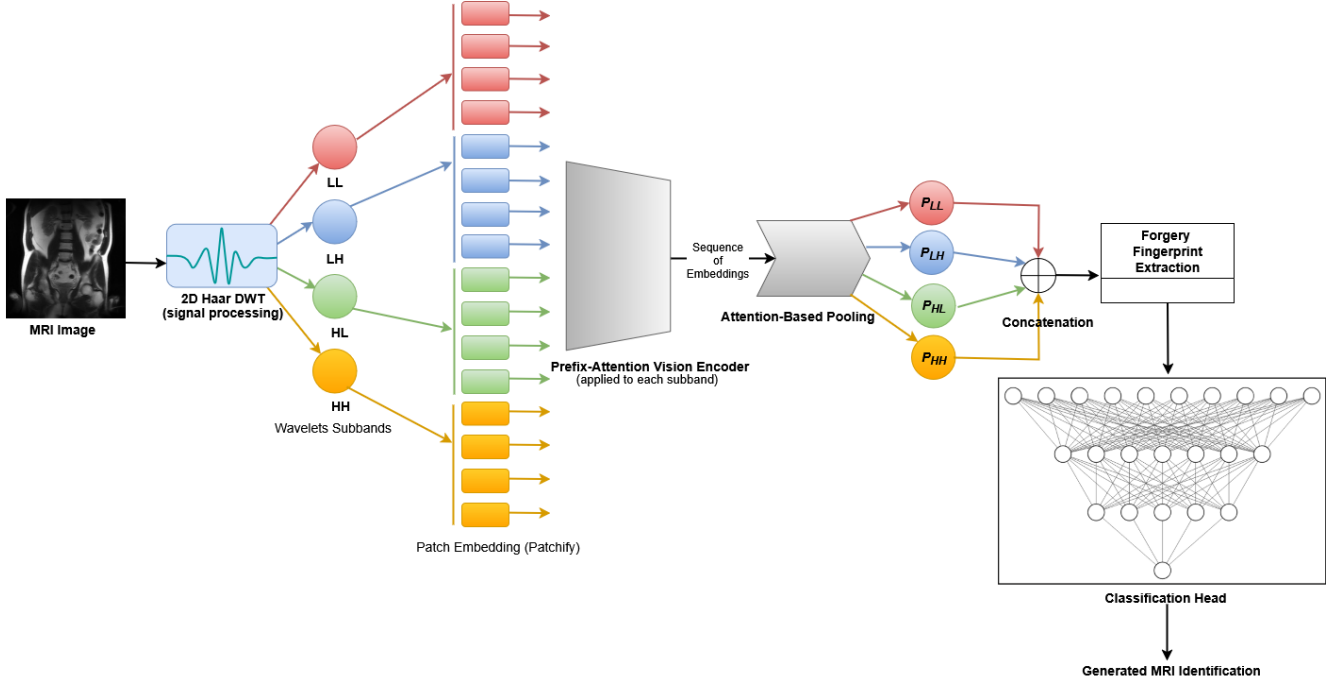


Figure 1. Overview of the proposed AIM-DWT architecture. The input MRI image is decomposed into four frequency subbands, LL (low-frequency), LH (horizontal), HL (vertical), and HH (diagonal), using 2D Haar DWT. Each subband is patchified and processed independently through a shared prefix-attention Vision Encoder (AIMv2). Attention-based pooling is applied to each subband’s embeddings, followed by concatenation for forgery fingerprint extraction. The final classification head identifies generated MRI images.

the model’s ability to capture subtle artifacts present in generated content. This encoder is based on the ViT architecture [9], and incorporates prefix attention, SwiGLU [38] feedforward blocks, and RMS normalization [50]. The encoder does not utilize a class token or static pooling; instead, it outputs patch-level embeddings for each subband.

To aggregate these embeddings into compact representations, we apply a learnable multi-head attention-based pooling module, inspired by AIM [12]. For each subband $s \in \{LL, LH, HL, HH\}$, the output embedding p_s is obtained by:

$$p_s = \text{Concat}(\text{head}_1, \dots, \text{head}_H),$$

$$\text{head}_h = \sum_{i=1}^N \frac{\exp(q_h^\top W_h^k p_i)}{\sum_{j=1}^N \exp(q_h^\top W_h^k p_j)} W_h^v p_i, \quad (2)$$

where p_i are the patch embeddings, q_h is a learnable query, and W_h^k, W_h^v are key and value projection matrices.

The resultant feature representations from each frequency subband are concatenated to form a comprehensive generative fingerprint embedding. Mathematically, it can be written as:

$$p_{\text{final}} = [p_{LL} \parallel p_{LH} \parallel p_{HL} \parallel p_{HH}], \quad (3)$$

where \parallel denotes concatenation. This embedding is passed to a lightweight classification head for final prediction.

This design enables the model to capture and fuse multi-scale spatial-frequency representations from each subband, guided by autoregressive feature learning and dynamic attention.

4. Experiments

This section presents two experimental evaluations: Subsection 4.1 evaluates the generative quality of selected generative models on the MRI dataset, while Subsection 4.2 assesses the forensic detection performance of the AIM-DWT method compared to relevant baselines on the MRI-Forensics dataset.

4.1. Evaluation of Generative Models

4.1.1. Baselines and Experimental Setup

As detailed previously in Section 3.1, we selected ProGAN [19], StyleGANv2-ADA [20], DiT XL2 and DiT L2 [31], and RAR XXL and RAR XL [49]. Each model was trained using its respective official codebase and environment configuration. For quantitative evaluation, we adopted the Fréchet Inception Distance (FID) [17] and Kernel Inception Distance (KID) [3]. FID quantifies the statistical similarity between real and generated image distributions using Inception features, and KID provides an unbiased kernel-based alternative. For visual analysis, we employed Princi-

pal Component Analysis (PCA) [30], t-distributed Stochastic Neighbor Embedding (t-SNE) [42], and Uniform Manifold Approximation and Projection (UMAP) [27]. PCA emphasizes global variance, t-SNE focuses on local clustering structure, and UMAP preserves both global and local data structures.

4.1.2. Results

Fig. 4 presents example images of real and generated MRIs from each generative model in the MRI-Forensics dataset, and Table 1 summarizes the performance across generative metrics. Consistent with the broader generative AI literature, DiT-XL2 [31] achieved the lowest FID (15.3322), indicating the highest overall distributional similarity to real MRI images. ProGAN obtained the best KID (0.0028 ± 0.00028), and RAR-XL achieved the highest IS (3.2852 ± 0.1146). The high IS but comparatively elevated FID of RAR-XL indicates that despite producing visually sharp and diverse images, their overall statistical distribution differs from real MRI imagery. These observations are further visualized using the dimensionality reduction techniques shown in Fig. 2: PCA (subfigure (a)), t-SNE (subfigure (b)), and UMAP (subfigure (c)). While PCA shows limited separation between real and generated samples, t-SNE and UMAP offer clearer clustering. For instance, ProGAN-generated images closely resemble real magnetic resonance images, as corroborated by their strong FID and KID scores relative to StyleGANv2-ADA. UMAP (subfigure (c)) shows a slightly better ability to separate generated images from real ones, particularly highlighted with StyleGANv2-ADA-generated imagery.

To investigate whether generative models imprint distinctive spectral fingerprints in MRI imagery observable in the frequency domain, we follow the methodology proposed by Corvi et al.[6]. Specifically, we apply a pretrained denoising filter[51] to 1,000 generated images from each model in the MRI-Forensics dataset to extract noise residuals. We then compute the normalized power spectra of these residuals using the 2D Fourier transform [5] and average them across samples. Fig. 3 presents the resulting 2D power spectra for each generative model, along with those of the real MRI imagery. Consistent with prior findings, each generative model leaves distinct spectral patterns, supporting the hypothesis that generation artifacts, though visually subtle, manifest in the frequency domain. Notably, DiT-XL/2 and DiT-L/2 exhibit nearly identical spectral curves, reflecting their architectural similarity, whereas models such as ProGAN produce more pronounced and distinctive spectral profiles. It is noteworthy that the higher the generative fidelity, as indicated by lower FID scores, the more similar the spectral curves between real and generated images. As shown in Fig. 3, DiT-XL/2 [31], which achieves the best FID score (15.33), displays spectral patterns that closely resemble those of real images, with only subtle deviations.

Table 1. Quantitative evaluation of generated images using Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). Lower FID and KID values indicate better quality.

Model	FID ↓	KID ↓
ProGAN [19]	17.7662	0.0028 ± 0.00028
StyleGANv2-ADA [20]	40.1445	0.0070 ± 0.00046
DiT-XL2 [31]	15.3322	0.0052 ± 0.00043
DiT-L2 [31]	18.5579	0.0061 ± 0.00038
RAR-XXL [49]	27.4211	0.0121 ± 0.00056
RAR-XL [49]	27.6076	0.0123 ± 0.00058

In contrast, other models exhibit visibly larger discrepancies. These observations further underscore the relevance of our AIM-DWT framework, which leverages frequency-aware subband representations to capture such subtle generative fingerprints for effective forensic detection.

4.2. Evaluation of Forensic Detection

To assess the effectiveness of our proposed AIM-DWT framework, we conducted comprehensive experiments on the MRI-Forensics dataset, comparing against current state-of-the-art forensic detection methods. The evaluation aims to measure not only the accuracy of detecting synthesized MRI images but also the robustness and generalizability of each method across different generative models. By examining cross-model and cross-dataset scenarios, we provide insights into how well forensic detection techniques capture the distinct artifacts left by various generative approaches. The following subsections detail the experimental setup, baselines, and the results of these evaluations, highlighting AIM-DWT’s performance relative to established methods.

4.2.1. Baselines and Experimental Setup

For comparative analysis, we selected the most relevant and state-of-the-art forensic detection methods: CNNSpot [43], DIRE [44], AIDE [46], and RINE [21]. To the best of our knowledge, none of these methods have been previously evaluated on synthesized MRI imagery. We trained all models, including our proposed AIM-DWT, under a binary classification setting to enable the detection of synthesized MRI images. All baseline implementations were sourced from their official public repositories. Interested readers are encouraged to refer to the original repositories for detailed setup information.

For AIM-DWT, experiments were conducted using Python 3.9.20 and PyTorch 2.5.1. The model was trained using AdamW optimizer with a learning rate of 0.0001, betas set to (0.9, 0.99), a weight decay of 0.006, and warm annealing cosine restarts. All computations were performed on a node equipped with eight NVIDIA A100-PCI GPUs, each with 80 GB of HBM2 memory. The dataset was split into

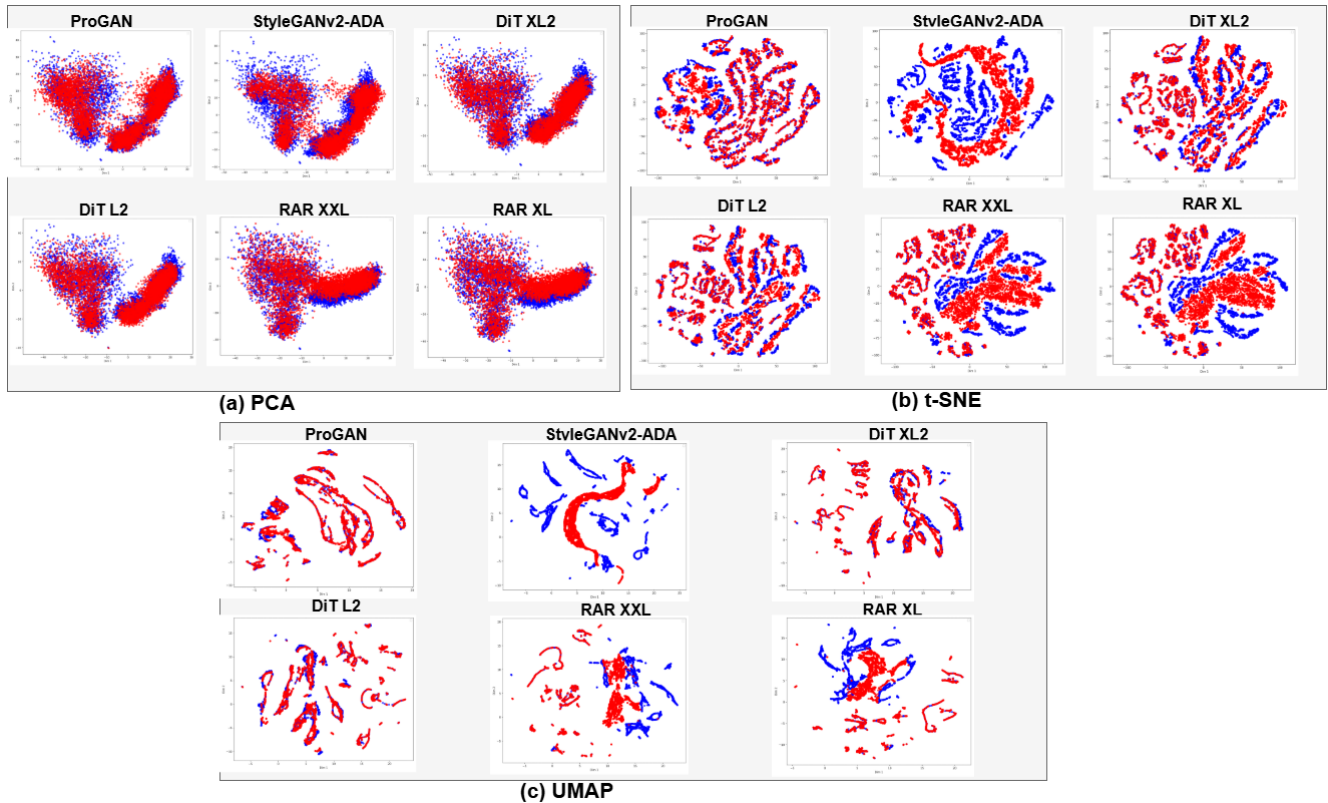


Figure 2. Dimensionality reduction analysis on synthetic MRI images generated by selected generative models. Red points represent generated MRI images, and blue points represent real MRI images

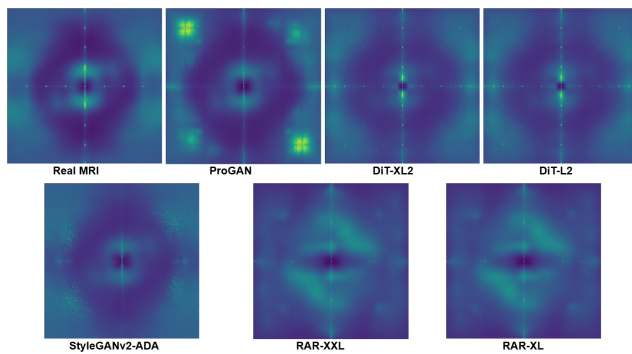


Figure 3. Average 2D power spectra of noise residuals from synthetic MRI images generated by different models in the MRI-Forensics dataset. Each spectrum reveals model-specific frequency fingerprints, computed from the squared FFT magnitude of 1,000 residuals per model.

4,806 real and 4,806 fake images for training, and 1,500 real and 1,500 fake images for validation. For evaluation, we report accuracy on detecting synthesized images, measuring how effectively each method identifies fake images among 6,306 generated samples. To evaluate generalizability, cross-generator experiments were performed: models

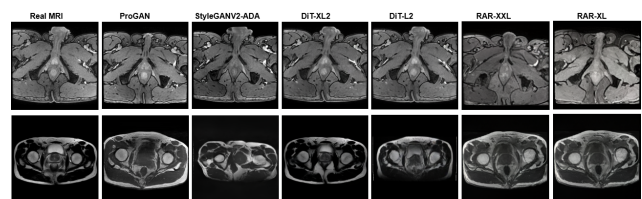


Figure 4. Example images from the MRI-Forensics dataset.

were trained on images generated by one model or family and tested on images from another.

4.2.2. Results

Table 2 reports cross-model generalizable detection performance on the MRI-Forensics dataset. AIM-DWT outperforms all baseline methods in most cases, achieving 100% accuracy when trained and validated on ProGAN [19] (4,806 training and 1,500 validation samples) and tested on the full set of 6,306 images generated by all other models. This strong performance can be attributed to AIM-DWT’s hybrid design, which combines autoregressive modeling with DWT-based frequency analysis, enabling it to capture both global spectral patterns and subtle genera-

tive artifacts that other methods often miss. The few exceptions highlight interesting nuances in model-specific strengths. For instance, CNNSpot [43] slightly outperforms AIM-DWT on RAR-XL-generated images, likely because CNNSpot’s convolutional architecture is highly sensitive to local pixel-level inconsistencies, which may be more pronounced in RAR-XL outputs. Similarly, RINE [21] performs marginally better when detecting ProGAN-generated images in one setting, possibly due to its reliance on residual noise analysis, which aligns well with the distinct artifacts left by ProGAN’s older architecture. In contrast, diffusion-based models and newer autoregressive generators tend to produce more subtle or less structured artifacts, where AIM-DWT’s frequency-domain analysis provides a clear advantage.

A key finding is that training on images generated by specific models influences generalization. For example, models trained on ProGAN exhibit higher generalization, while models trained on other generators perform worse when tested on ProGAN-generated images. This may be attributed to differences in generative quality: lower KID scores may indicate fewer detectable artifacts, making detection harder. We hypothesize that training forensic models with lower KID score generative output can enhance generalizability. Even in non-optimal cases, AIM-DWT consistently delivers better performance across evaluations. It is noteworthy that blank entries in Table 2 indicate cases where all images generated from a given model were used for training and validation, leaving no samples available for testing from the same model.

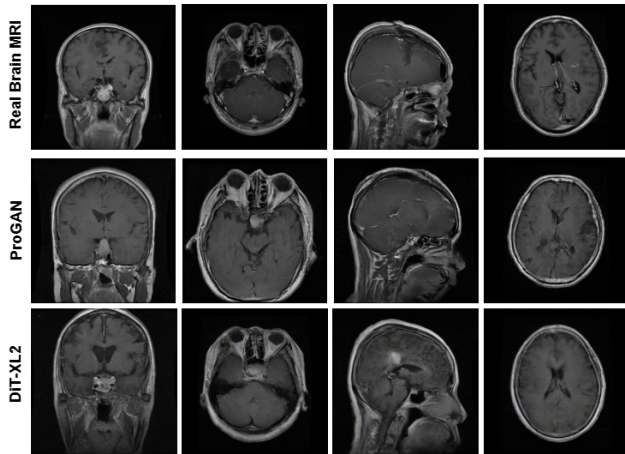


Figure 5. Example images of Brain MRI Imagery Generated from Publicly Available Dataset [29]

4.2.3. Cross-Dataset Evaluation on Brain MRI

To further assess generalizability, we evaluated AIM-DWT on a different dataset. Due to the lack of publicly available generated MRI imagery, we used a brain MRI dataset

released by [29], containing 7,023 real images. We trained ProGAN and DiT-XL/2 to generate an equal number of synthetic images (7,023 each). Example outputs are shown in Figure 5. On this dataset, ProGAN achieved an IS of 2.4064 ± 0.0213 , FID of 36.0138, and KID of 0.0245 ± 0.0008 . DiT-XL/2 obtained an IS of 2.4004 ± 0.0354 , FID of 51.7182, and KID of 0.0482 ± 0.0013 . Table 3 presents the detection performance of all methods on the brain MRI dataset. We used the same model weights trained on the MRI-Forensics dataset (prostate MRI only). AIM-DWT outperforms nearly all baselines in identifying generated brain MRI images. Since we now have both real and fake brain MRI imagery, we report accuracy and Area Under the ROC Curve (AUC). AIM-DWT achieves the best AUC overall, and its accuracy on real image detection remains comparable to top baselines.

Notably, all baseline methods experience a significant drop in detecting ProGAN-generated brain MRI images when trained on DiT-XL/2 images. In contrast, AIM-DWT achieves near-optimal performance with 99.26% accuracy in this setting, demonstrating its robust cross-dataset generalization capability.

5. Conclusion

The growing capabilities of generative AI pose significant risks to medical imaging, particularly MRI, where adversarial manipulations can lead to misdiagnosis. In response to the lack of existing forensic methods for this domain, we introduce MRI-Forensics, a new dataset comprising real and synthesized prostate MRI images generated using advanced models from GAN, Diffusion, and Autoregressive families. Through frequency-domain analysis of noise residuals in the MRI-Forensics dataset, we show that different generative models leave distinct spectral patterns in MRI imagery. We then proposed AIM-DWT, a forensic detection framework that integrates autoregressive modeling with DWT-based frequency analysis. This integration proved effective in capturing a wide spectrum of generative fingerprints, from visually apparent to deeply subtle artifacts, produced by advanced generative models. Extensive experiments demonstrated that AIM-DWT achieves state-of-the-art performance on the MRI-Forensics dataset and consistently outperforms existing detection methods in both cross-model and cross-dataset evaluations. Notably, despite being trained solely on prostate MRI data, AIM-DWT generalizes robustly to brain MRI imagery, highlighting its adaptability. In conclusion, our study establishes a foundation for future research in the forensic detection of MRI imagery and contributes to the trustworthiness and integrity of AI-assisted diagnostics.

Despite these contributions, some limitations remain in our study. Our evaluations are conducted under an unconditional generation setup. The generalization of AIM-DWT to

Table 2. Cross-Generator Detection Accuracy on MRI-Forensics. The table shows the classification accuracy for detecting generated images as fake, evaluating the generalization capability of each detection method when trained on one type of generative model and tested on others.

Model	Training	Testing on Generative Models Individually					
		ProGAN	StyleGANv2	DiT XL2	DiT L2	RAR XXL	RAR XL
CNN. [43]	ProGAN [19]	–	99.95	99.98	99.97	100.0	100.0
	StyleGANv2 [20]	5.38	–	99.59	99.49	99.30	99.38
	DiT XL2 [31]	3.04	90.85	–	99.97	100.0	100.0
	RAR XXL [49]	0.13	5.17	74.12	80.57	–	100.0
DIRE [44]	ProGAN [19]	–	99.98	99.98	99.96	100.0	100.0
	StyleGANv2 [20]	0.92	–	99.69	99.60	99.87	99.93
	DiT XL2 [31]	0.13	97.14	–	99.96	100.0	100.0
	RAR XXL [49]	0.09	0.37	70.18	76.64	–	100.0
AIDE [46]	ProGAN [19]	–	80.99	71.95	71.58	73.76	73.28
	StyleGANv2 [20]	1.73	–	1.44	1.47	7.28	7.20
	DiT XL2 [31]	14.15	36.62	–	97.51	83.63	83.44
	RAR XXL [49]	13.08	24.52	24.94	27.10	–	99.86
RINE [21]	ProGAN [19]	–	75.20	64.40	66.70	100.0	100.0
	StyleGANv2 [20]	2.30	–	53.50	60.40	100.0	100.0
	DiT XL2 [31]	1.60	17.20	–	99.90	100.0	100.0
	RAR XXL [49]	0.20	0.50	27.70	31.90	–	100.0
AIM-DWT	ProGAN [19]	–	100.0	100.0	100.0	100.0	100.0
	StyleGANv2 [20]	75.07	–	100.00	99.95	99.98	99.98
	DiT XL2 [31]	78.43	99.81	–	99.97	100.0	100.0
	RAR XXL [49]	3.50	39.12	98.26	98.57	–	99.98

Table 3. Cross-generator detection on Brain MRI dataset. Models are trained on prostate MRI (MRI-Forensics) and tested on a distinct domain. Metrics: real/fake accuracy and AUC.

Method	Training	Testing on ProGAN [19]			Testing on DiT XL2 [31]		
		Real Acc.	Fake Acc.	AUC	Real Acc.	Fake Acc.	AUC
CNNSpot [43]	ProGAN	99.35	99.83	99.99	99.35	99.99	99.99
	DiT XL2	99.15	1.89	95.92	99.15	99.99	99.67
DIRE [44]	ProGAN	99.18	99.91	99.75	99.18	100.0	99.87
	DiT XL2	99.38	2.56	96.63	99.38	99.90	99.77
AIDE [46]	ProGAN	81.43	64.43	82.46	81.43	60.56	79.85
	DiT XL2	80.58	41.42	68.27	80.58	92.60	95.24
RINE [21]	ProGAN	87.10	100.0	99.98	87.10	96.80	97.23
	DiT XL2	99.20	38.80	94.80	99.20	99.80	99.95
AIM-DWT	ProGAN	97.38	100.0	100.0	97.38	100.0	100.0
	DiT XL2	99.35	99.26	99.96	99.35	99.99	100.0

conditional generation settings, higher-resolution inputs, or other MRI modalities, for instance, contrast-enhanced sequences, remains to be explored. While AIM-DWT consistently outperforms baseline methods, we observe that it does not consistently achieve the best result in every cross-model scenario, suggesting room for improvement in de-

tecting certain model-specific artifacts.

Acknowledgment

This work was supported in part by the National Science Foundation under Grants CNS-2018611 and CNS-1920182.

References

- [1] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1): 90–93, 1974. 2
- [2] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. 2
- [3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018. 4
- [4] Chih-Wei Chang, Shaoyan Pan, Junbo Peng, Elahheh Salari, Justin Roper, Richard Qiu, Yuan Gao, Tian Liu, Hui-Kuo Shu, Hui Mao, et al. Using diffusion model to generate high-resolution MRI. In *Medical Imaging 2024: Clinical and Biomedical Imaging*, pages 444–449. SPIE, 2024. 1
- [5] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965. 2, 5
- [6] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 973–982, 2023. 5
- [7] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2
- [8] Nikhil J Dhinagar, Sophia I Thomopoulos, and Paul M Thompson. Generative AI improves mri-based detection of Alzheimer’s disease by using latent diffusion models and convolutional neural networks. *Alzheimer’s & Dementia*, 20: e089958, 2024. 1
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4
- [10] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in Neural Information processing systems*, 33: 3022–3032, 2020. 2
- [11] Marco Eichelberg, Joerg Riesmeier, Thomas Wilkens, Andrew J Hewett, Andreas Barth, and Peter Jensch. Ten years of medical imaging standardization and prototypical implementation: the dicom standard and the offis dicom toolkit (dcm2tk). In *Medical Imaging 2004: PACS and Imaging Informatics*, pages 57–68. SPIE, 2004. 3
- [12] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024. 3, 4
- [13] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilhaume Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 3
- [14] Bernardo Gonçalves, Pedro Vieira, and Ana Vieira. Abdominal MRI synthesis using stylegan2-ada. In *2023 IST-Africa Conference (IST-Africa)*, pages 1–9. IEEE, 2023. 1
- [15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information processing systems*, 27, 2014. 1
- [16] Lujun Gui, Chuyang Ye, and Tianyi Yan. Cavm: Conditional autoregressive vision model for contrast-enhanced brain tumor MRI synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 161–170. Springer, 2024. 2
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information processing systems*, 30, 2017. 4
- [18] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for under-sampled MRI reconstruction. *Physics in Medicine & Biology*, 63(13):135007, 2018. 1
- [19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3, 4, 5, 6, 8
- [20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information processing systems*, 33:12104–12114, 2020. 1, 3, 4, 5, 8
- [21] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2025. 2, 5, 7, 8
- [22] Matteo Lai, Chiara Marzi, Mario Mascialchi, and Stefano Diciotti. Brain MRI synthesis using stylegan2-ada. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 1
- [23] Guanxiong Luo, Shoujin Huang, and Martin Uecker. Autoregressive image diffusion: Generation of image sequence and application in MRI. *arXiv preprint arXiv:2405.14327*, 2024. 2
- [24] Arpan Mahara and Naphtali Rische. Methods and trends in detecting generated images: A comprehensive review. *arXiv preprint arXiv:2502.15176*, 2025. 2
- [25] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 2, 3
- [26] Francesco Marra, Diego Gagnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019. 2

- [27] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5
- [28] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarbuerger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023. 1
- [29] Msoud Nickparvar. Brain tumor MRI dataset, 2021. 7
- [30] Karl Pearson. Liii. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 5
- [31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 4, 5, 8
- [32] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022. 1
- [33] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, part III 18*, pages 234–241. Springer, 2015. 3
- [36] Anamitra Bardhan Roy, Debasmita Dey, Bidisha Mohanty, and Devmalya Banerjee. Comparison of FFT, DCT, DWT, WHT compression techniques on electrocardiogram and photoplethysmography signals. In *IJCA Special Issue on International Conference on Computing, Communication and Sensor Network CCSN*, pages 6–11, 2012. 3
- [37] Shaheer U Saeed, Tom Syer, Wen Yan, Qianye Yang, Mark Emberton, Shonit Punwani, Matthew John Clarkson, Dean Barratt, and Yipeng Hu. Bi-parametric prostate mr image synthesis using pathology and sequence-conditioned stable diffusion. In *Medical Imaging with Deep Learning*, pages 814–828. PMLR, 2024. 2
- [38] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4
- [39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [41] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016. 1
- [42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008. 5
- [43] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 2, 5, 7, 8
- [44] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 2, 5, 8
- [45] Isaac RL Xu, Derek J Van Booven, Sankalp Goberdhan, Adrian Breto, Joao Porto, Mohammad Alhousseini, Ahmad Algohary, Radka Stoyanova, Sanoj Punnen, Anton Mahne, et al. Generative adversarial networks can create high quality artificial prostate cancer magnetic resonance images. *Journal of Personalized Medicine*, 13(3):547, 2023. 2
- [46] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 2, 5, 8
- [47] Bilal Yousaf, Muhammad Usama, Waqas Sultani, Arif Mahmood, and Junaid Qadir. Fake visual content detection using two-stream convolutional neural networks. *Neural Computing and Applications*, 34(10):7991–8004, 2022. 2
- [48] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7556–7566, 2019. 2
- [49] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024. 3, 4, 5, 8
- [50] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 4
- [51] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 5