Computers in Biology and Medicine 183 (2024) 109248

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

MITS-GAN: Safeguarding medical imaging from tampering with generative adversarial networks

Giovanni Pasqualino, Luca Guarnera, Alessandro Ortis *, Sebastiano Battiato

Department of Mathematics and Computer Science, University of Catania, Viale Andrea Doria 6, Catania, 95126, Italy

ARTICLE INFO

Keywords:
Medical image
Generative adversarial network
Adversarial attacks
Image tampering

ABSTRACT

The progress in generative models, particularly Generative Adversarial Networks (GANs), opened new possibilities for image generation but raised concerns about potential malicious uses, especially in sensitive areas like medical imaging. This study introduces MITS-GAN, a novel approach to prevent tampering in medical images, with a specific focus on CT scans. The approach disrupts the output of the attacker's CT-GAN architecture by introducing finely tuned perturbations that are imperceptible to the human eye. Specifically, the proposed approach involves the introduction of appropriate Gaussian noise to the input as a protective measure against various attacks. Our method aims to enhance tamper resistance, comparing favorably to existing techniques. Experimental results on a CT scan demonstrate MITS-GAN's superior performance, emphasizing its ability to generate tamper-resistant images with negligible artifacts. As image tampering in medical domains poses life-threatening risks, our proactive approach contributes to the responsible and ethical use of generative models. This work provides a foundation for future research in countering cyber threats in medical imaging. Models and codes are publicly available.[1]

1. Introduction

In recent years, advancements in generative models have ushered in a new era of image generation and manipulation, showcasing remarkable capabilities in rendering images increasingly indistinguishable from their original counterparts [1–3]. This progress, driven by deep learning techniques, has found applications in various domains, from creative artistry [4] to medical imaging [5], among others.

In medical imaging, GANs have been instrumental in addressing the challenge of data scarcity. They are used to augment datasets by generating synthetic images or translating images between different modalities. For instance, GANs have been employed to convert MRI images into CT images [6], generate realistic 2D brain MRI images [7], and even enhance image resolution [8]. These applications not only improve the quality and availability of medical images but also support advancements in diagnostic processes. Islam et al. [9] proposed a GAN-based method to generate PET images of the brain. This new dataset could be used to create new artificial intelligence methods to help doctors make an early diagnosis of Alzheimer's disease. Due to the absence of Arterial Spin Labeling (ASL) data, Li et al. [10] proposed a GAN architecture in order to synthesize such images. ASL measures cerebral blood flow, which is useful for making diagnoses for dementia diseases. Pang et al. [11] proposed a semi-supervised GAN architecture to perform data augmentation on 'breast ultrasound mass' images, in order to significantly improve the performance of the TCGAN classifier, created to discriminate the presence or absence of breast cancer. Liu et al. [12] proposed a multi-cycle GAN to generate CT images from MRI images, overcoming the limitations of MRI in that no information about the patient's bones is obtained. The technique reduces patients' exposure to radiation, improving the safety of radiotherapy. In general, MRI images contain noise that can be removed with the conditional GAN proposed by Tian et al. [13]. This work exceeds state-of-the-art methods in both noise reduction and the preservation of robust anatomical structures and defined contrast. A very interesting approach was proposed by Dong et al. [14], in which a GAN architecture was used to automatically segment CT images of the thorax, using a U-Net architecture as generator and FCN as discriminator, in order to improve radiotherapy treatment planning. The proposed architecture achieved better segmentation results than state-of-the-art approaches.

However, alongside positive applications, researchers have demonstrated the malicious use of GANs [15] for tasks such as malware obfuscation [16] and the creation of deepfakes [17]. The key idea behind GANs involves training two neural networks, a generator, and a discriminator, in an adversarial setting. The generator aims to produce synthetic data, such as images, that is indistinguishable from

* Corresponding author.
E-mail address: alessandro.ortis@unict.it (A. Ortis).

[1] https://iplab.dmi.unict.it/MITS-GAN-2024/

https://doi.org/10.1016/j.compbiomed.2024.109248
Received 27 March 2024; Received in revised form 19 September 2024; Accepted 2 October 2024

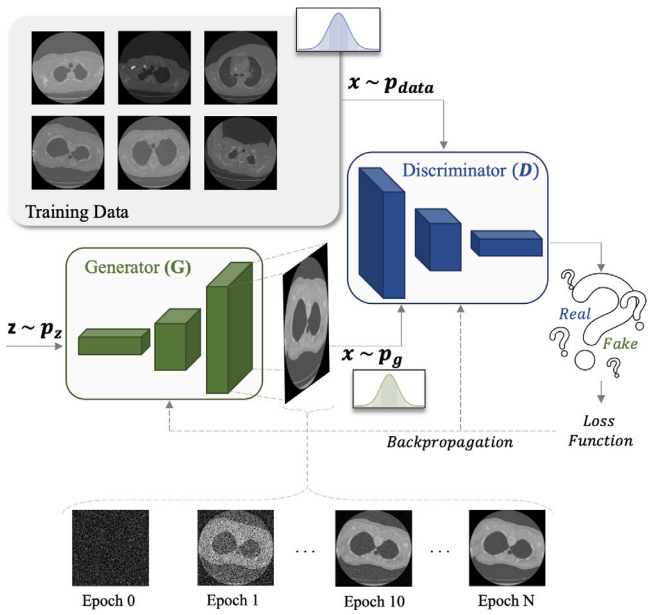

Fig. 1. Overview of the GAN architecture and training process.

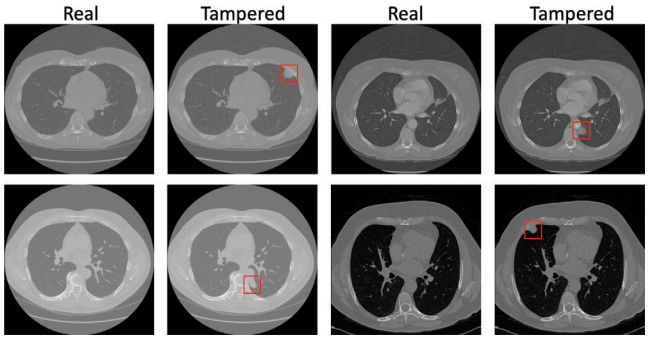

Fig. 2. Qualitative results comparison between real and tampered CT scans. Columns 1 and 3 show the original images, whereas Columns 2 and 4 depict the manipulated images. The red bounding boxes highlight the manipulations introduced by CT-GAN, wherein tumors have been added to the scans. This visual representation underscores the impact and detectability of manipulations within the medical imaging context.

real data, while the discriminator's task is to differentiate between real and generated data. This adversarial training process results in the generator continually improving its ability to create realistic data, making GANs highly effective in image generation tasks (Fig. 1). Within the medical domain, the potential consequences of malicious tampering are critical, as the integrity and authenticity of images can have life-or-death implications as shown in Fig. 2 manipulating the images provided by the authors of [17]. Image tampering techniques [18] have raised concerns by highlighting the potential for malicious manipulation of medical images, such as computed tomography (CT) scans and radiographs. This introduces a new dimension of cyber attacks, with image manipulation being employed to deceive medical professionals and compromise patient care, potentially leading to misdiagnoses. To address this challenge, the research community has focused on developing automated detection systems for image manipulation, treating it as a classification task. Various learning-based approaches have shown promise, achieving excellent classification accuracy [19–22]. Alternatively, another strategy is to prevent manipulations at the source by disrupting manipulation methods' output [23–25]. The key idea is to disrupt generative neural network models by introducing noise patterns at a low level, making it more challenging for malicious actors to create convincing forgeries. In this study, we investigate the problem of

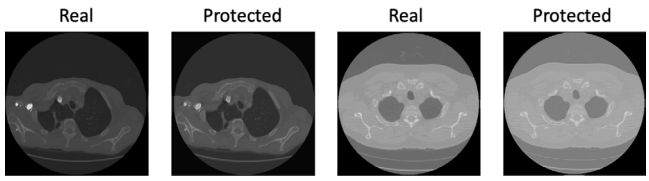

Fig. 3. Comparison between Real unprotected CT scans and protected CT scans generated by the proposed model MITS-GAN. As can be noted, the protected images, which embed the protection noise pattern, are similar to the original one.

image tampering in the medical domain, focusing on the manipulation of CT scans. To address this problem, we propose *Medical Imaging Tamper Safe-GAN* (MITS-GAN) method. In particular, we introduce a framework based on Generative Adversarial Networks with the aim to generate protected images against image manipulation model [18]. Our model generates the protected scans introducing an imperceptible noise with the aim to disrupt the output when the manipulation is performed and minimizing potential artifact that could pose challenges during the review process by medical experts (Fig. 3). MITS-GAN is designed to protect medical images from tampering, addressing risks such as misdiagnosis and medical fraud. Real-world concerns include manipulating CT scans to deceive doctors or commit insurance fraud, as well as using deepfake technology to fabricate medical images. This research is significant for its potential to enhance diagnostic accuracy and bolster healthcare cybersecurity. By ensuring the authenticity of medical images, MITS-GAN supports reliable diagnoses, safeguards patient data, and prevents the misuse of AI technologies in healthcare.

The main contributions of the proposed work are:

- We address the critical issue of medical image tampering by proposing a robust methodology that ensures the integrity and reliability of diagnostic images. This approach is motivated by the urgent need to protect medical imaging from manipulation, which could otherwise compromise diagnostic accuracy and the reliability of Machine Learning methods and other systems based on such datasets;
- We introduce a novel framework called MITS-GAN (Medical Imaging Tamper Safe-GAN) and compare its performance with state-of-the-art methods. MITS-GAN leverages Generative Adversarial Networks (GANs) to safeguard medical images from tampering. Our results demonstrate the superior effectiveness of MITS-GAN in preserving the authenticity and reliability of medical images;
- Our work lays the groundwork for future research aimed at mitigating cyber threats in the field of medical imaging. We emphasize the importance of proactive measures to protect and maintain the integrity of medical scans, highlighting the long-term implications of our approach for the security of medical data.

The document is organized as follows. Section 2 reports the main works in the literature. The proposed approach is described in Section 3. The used for the experiments, the metrics to evaluate the performances, the experimental results and comparison are reported in Section 4. Finally, Sections 5 and 6 conclude the paper with some hints for future works.

2. Related work

2.1. GAN applications in medical imaging

GANs have made significant contributions to the field of medical imaging, addressing various challenges and enhancing the quality and accessibility of medical imagery. GANs' ability to generate realistic

images has been leveraged to alleviate the common issue of data scarcity in medical image analysis by augmenting s through the generation of new images or style translation. For instance, the authors of [26] utilized a conditional GAN (cGAN) to transform 2D slices of CT images into PET images. The authors of [27,28] demonstrated a similar approach employing a fully convolutional network with a cGAN architecture. In [29], domain adaptation was employed to convert MRI images into CT images, while the authors of [6] used CycleGAN to convert MRI images into CT images and vice versa. The authors of [7] use a deep convolutional GAN (DCGAN) to generate 2D brain MRI images. In [30], the authors used a DCGAN to generate 2D liver lesions. In [31], the authors generated 3D blood vessels using a Wasserstien (WGAN). In [5], the authors train two DCGANs for generating 2D chest X-rays (one for malign and the other for benign). Within the medical imaging domain, GANs have also found other interesting applications in segmentation [32], super-resolution [8] and anomaly detection [33].

2.2. Deepfake detection methods

The ability to understand if an image is generated by a generative Neural Network is in some case challenging also for the human eyes representing a complicated problem. To address this problem, numerous methods have been developed over the years to determine the authenticity of an image [34].

Researchers have demonstrated that generative engines leave traces on synthetic content that can be detected in the frequency domain [35,36]. Giudice et al. [37] proposed a method able to identify the specific frequency that characterizes a GAN engine through a deeper analysis of coefficients given from the Discrete Cosine Transform (DCT). These traces are characterized by both the network architecture (number and type of layers) and its specific parameters [38]. Based on this principle, the synthetic images created by various GAN engines are also characterized by different statistics in terms of correlations between pixels. To capture this trace left by the convolutional layers, Guarnera et al. [39,40] proposed a method based on the Expectation-Maximization [41] algorithm, obtaining excellent classification results in distinguishing pristine data from deepfakes. Wang et al. [42] proposed a method to discriminate real images from those generated by ProGAN [43]. The method turns out to be able to generalize with synthetic data created by different GAN architectures.

Recent solutions use Vision Transformer to detect deepfakes [44,45]. For example, [46] combined vision transformers with a convolutional network, achieving excellent results in solving the proposed task.

Researchers are also actively engaged in developing advanced techniques to identify synthetic images generated by Diffusion Models [47]. Corvi et al. [48] investigated the challenges associated with distinguishing synthetic images produced by diffusion models from authentic ones. They assess the suitability of current state-of-the-art detectors for this specific task. Sha et al. [49] proposed DE-FAKE, a machine-learning classifier-based method designed for the detection of diffusion models on four prevalent text-image architectures. Meanwhile, Guarnera et al. [50] introduced a hierarchical approach based on recent architectures. This approach involves three levels of analysis: determines whether the image is real or manipulated by any generative architecture (AI-generated); identifies the specific framework, such as GAN or DM; defines the specific generative architecture among a predefined set.

Experimental results of all these methods have demonstrated that generative models leave unique traces that can be detected to distinguish deepfakes well from real multimedia content.

2.3. Adversarial attacks

Adversarial attack methods are designed to introduce imperceptible changes to images with the aim of disrupting the feature extraction process performed by neural networks. Initially applied in classification tasks [51–53], where their goal was to induce misclassification errors, these methods have been extended to segmentation [54] and detection tasks [55]. However, the optimization process of unique pattern for each individual image can be highly time-consuming. To address this challenge, researchers introduced the concept of generic universal image-agnostic noise patterns [56,57]. Such noise patterns are designed to be versatile and applicable across a wide range of images, eliminating the need for time-consuming, image-specific pattern optimization. While this approach has proven effective in the context of tasks involving misclassification, it has demonstrated limitations when applied to generative models.

2.4. Image manipulation prevention

Prevent image manipulations exploiting adversarial attack techniques has been recently studied as an alternative way to the classification and detection of manipulated images. The authors of [24] propose a baseline methods for disrupting deepfakes by adapting adversarial attack methods to image translation networks. In [23,58] the authors presented an approach to nullify the effect of image-to-image translation models. In [59] authors proposed a novel neural network based approach to generate image-specific patterns for low-resolution images which differs from the previous methods because does not require optimization of a specific pattern for each image separately which is computationally expensive. In [25] an innovative framework called Targeted Adversarial Attacks for Facial Forgery Detection (TAFIM), a innovative framework that accepts a real image X_i and a global perturbation δ_G as inputs to the model. This process generates an image-specific perturbation δ_i. The resulting perturbation is then added to the original image, producing the protected image X_i^p, which is subsequently processed through the manipulation model f_ϕ. The outcome is the manipulated output Y_i^p, utilized for driving the optimization process.

3. Proposed method

Our goal is to prevent image manipulation, specifically the addition or removal of tumors in CT scans, by disrupting the CT-GAN [18] architecture. We designed a proper way to introduce an imperceptible perturbation that disrupts the CT-GAN's output in case of malicious manipulation, making it easier for a human to identify tampered scans, and hence ensuring the integrity of medical imaging process. MITS-GAN operates by applying protection at a slice-by-slice level for 3D CT scans. Rather than implementing a global protection mechanism across the entire 3D volume, our approach applies 2D convolutions to each slice independently. This localized protection ensures that even if only a subset of slices is manipulated, the algorithm remains robust, as each slice is protected individually. This slice-wise protection is particularly advantageous in scenarios where tampering occurs in specific areas of the scan, as it allows the detection of subtle and localized changes. In contrast to recent methods that use 3D-based GANs [18,60] to solve tasks such as creating new datasets or performing attacks on medical images, a slice-wise approach such as the one we propose, can offer greater advantages in terms of computational efficiency and flexibility in both the creation of new synthetic data and the handling of partial manipulations.

The chosen architecture leverages Generative Adversarial Networks (GANs) to generate protected images using a Gaussian perturbation (noise). The primary idea is to ensure that these protected images are indistinguishable from the original ones. By concatenating the noise as an additional channel rather than directly adding it to the CT scan

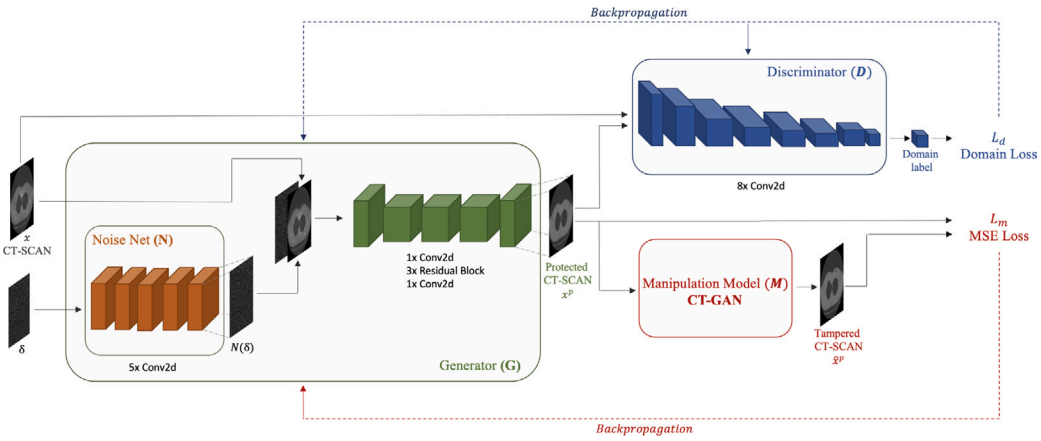

Fig. 4. Model architecture overview: The generator receives the input image x and perturbation noise δ to produce the protected image x^p. Subsequently, x^p is forwarded to the manipulation model and discriminator.

images (x), potential image artifacts are avoided. This approach helps the network treat the perturbation as extra information rather than as part of the image data itself, which could otherwise lead to unwanted artifacts that might be discarded during training. The inclusion of Mean Squared Error (MSE) loss, which is maximized during training, plays a crucial role. This loss function compels the network to generate robust images that resist manipulations from the CT-GAN model, thereby preserving fidelity to the original images.

3.1. Overview

The proposed architecture is illustrated in Fig. 4. Let δ be an image-agnostic perturbation, distributed according to a Gaussian distribution, and $X = \{x_i\}_{i=0}^{N}$ be the of N CT scans. δ and x are fed into the Generator $G(x, \delta; \theta_G)$ of parameters θ_G, which includes a Noise Net N that, for a given input δ, outputs $N(\delta)$. This module contains five 2D convolutional layers, each followed by batch normalization and the ReLU activation function. Subsequently, the CT scan x and $N(\delta)$ are concatenated channel-wise and passed through a sequence of convolutional layers (one 2D convolution, three residual blocks, one 2D convolution). Each layer is followed by batch normalization and the ReLU activation function, except for the last one, which applies the Tanh activation function. Concatenating the noise as a new channel allows the network to consider the perturbation as extra information instead of adding it to x, which could lead to it being considered as image artifacts and therefore discarding them in the training phase to make the generator's output similar to x. The resulting output of G, denoted as x^p, represents the protected scan and is forwarded to the CT-GAN manipulation model M which tamper the x^p producing $\hat{x}^p$ and whose parameters are frozen. Additionally, x and x^p are provided to the discriminator $D(x; \theta_D)$ that outputs the likelihood d that a given image x belong to the real images with the aim to distinguish between a protected image produced by the generator and the original unprotected one. The Discriminator D consists of eight 2D convolutional layers, each followed by batch normalization and the LeakyReLU activation function. The model is trained using a generative adversarial objective, encouraging the generator to produce protected images similar to the original (unprotected) ones.

The goal is to optimize the following min–max objective:

$$\min_{G} \max_{D,M} L_d(D, G) + \alpha L_m(G, M) \tag{1}$$

where L_d represents the domain loss:

$$L_d(D, G) = \mathbb{E}_{x^p}[log\, D(x^p; \theta_D)] + \mathbb{E}_{x,\delta}[log(1 - D(G(x, \delta; \theta_G); \theta_D))] \tag{2}$$

where $\mathbb{E}$ denotes the average value of the enclosed expression over the specified distribution. In detail, $\mathbb{E}_{x^p}[log\, D(x^p; \theta_D)]$ represents the expected log-probability that the discriminator assigns to real data samples. The discriminator aims to maximize this term, meaning it tries to correctly identify real data as real. $\mathbb{E}_{x,\delta}[log(1-D(G(x, \delta; \theta_G); \theta_D))]$ represents the expected log-probability that the discriminator assigns to fake data samples created by the generator. The discriminator aims to maximize this term by correctly identifying fake data as fake (i.e., assigning a low probability to fake data being real).

L_m is the Mean Squared Error (MSE) loss computed between the output of the model M and the generator G:

$$L_m(G, M) = \mathbb{E}_{x,\delta}[(M(G(x, \delta; \theta_G)) - G(x, \delta; \theta_G))^2] \tag{3}$$

where $\mathbb{E}_{x,\delta}[(M(G(x, \delta; \theta_G)) - G(x, \delta; \theta_G))^2]$ denotes that the expectation is taken over the distributions of x and δ, indicating that we are considering the average squared error across all possible input and noise pairs.

α is the weight that controls the interaction of these losses. In particular, the optimization of the loss function L_d concerning both the discriminator D and the generator G constitutes the standard generative adversarial objective. This objective concurrently refines both the generator and the discriminator. Subsequently, the term L_m is introduced to augment the visual dissimilarity between the generated output x^p and its corresponding tampered image $\hat{x}^p$. The inclusion of L_m serves the purpose of increasing noticeable artifacts in the manipulated content $\hat{x}^p$ when attempting to tamper with x^p. Algorithm 1 reports the complete forward procedure of the proposed method.

4. Dataset, metrics and experimental results

In this section, we expound upon the dataset, outline the metrics under consideration, and scrutinize the outcomes derived from the introduced methodology. The manipulation model, denoted as CT-GAN, operates by taking a CT scan as input, identifying a designated square for manipulation, and subsequently producing the manipulated square. This process involves either removing or adding a tumor, resulting in the tampered square, which is then seamlessly integrated into the original scan. It is noteworthy that the tampered scan closely resembles the original, with the sole exception being the generated square. Our model ensures the comprehensive protection of the entire scan, as manipulations can be applied to any part of the scan, necessitating robustness across the entire image.

4.1. Dataset

Our approach is evaluated using the dataset outlined in [61], following the training editing procedure specified in [18]. In this procedure,

Algorithm 1: Forward pass description of the proposed framework

Input: *CT* scan = x, δ = perturbation;
Step 1: Forward δ and x through the Generator G;
Step 2: Feed δ into the Noise Net N within G, obtaining $N(\delta)$, and concatenate it with x;
Step 3: Apply five convolutions in G to generate the protected image x^p;
Step 4: Pass x^p to the Discriminator D and the Manipulation model M (CT-GAN);
Step 5: Compute domain loss for x and x^p through D;
Step 6: Utilize M to extract and manipulate a 32×32 pixel square q. Generate a tampered image $\hat{x}^p$ by pasting q onto x^p;
Step 7: Compute MSE loss between x^p and $\hat{x}^p$;

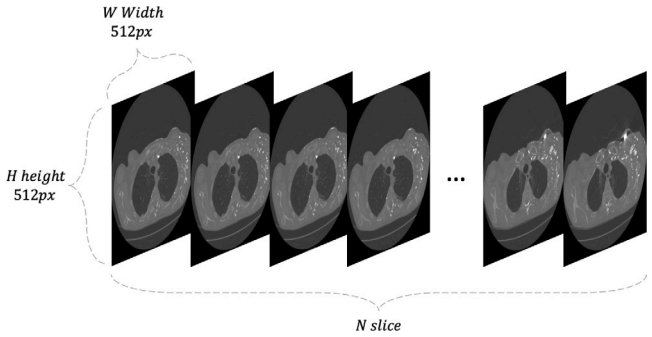

Fig. 5. Example of a CT scan.

the authors' injector model is trained on cancer samples with a minimum diameter of 10 mm, while the remover model is trained on benign lung nodules with a diameter less than 3 mm. The dataset comprises 888 CT scans, and we adhered to the standard split procedure, allocating 80% as the training set and the remaining 20% as the test set. Each CT scan is stored as a DICOM or Raw file, and its dimensions are represented as $N \times H \times W$, where N identify the number of "slices" or thin sections through which the scan was performed, H represents the height, and W represents the width of the scan (see Fig. 5). The considered CT scans have a fixed resolution of 512×512 and a variable number of slices within the range $N \in [95, 764]$.

4.2. Metrics

To evaluate the output quality in a quantitative way, we compute the RMSE, PSNR, LPIPS [62] and SSIM metrics as detailed below:

- RMSE (Root Mean Square Error) measure the deviation between predicted values from a model and the actual observed values. Lower values are better, 0 zero indicates that the predicted values are equals to the observed values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|y_i - x_i\|^2} \tag{4}$$

- PSNR (Peak signal-to-noise ratio) is a metric used to quantify the quality of an image or video by measuring the ratio of the maximum possible signal strength to the noise introduced during compression or transmission. Higher PSNR values generally indicate better image quality.

$$PSNR(I,J) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2 \tag{5}$$

- LPIPS computes the similarity between the feature representations of two image patches extracted by a pre-trained neural network. This metric has demonstrated a strong alignment with human perception. The lower the LPIPS score, the more perceptually similar the image patches are considered to be. For the experiment we used as pretrained network SqueezeNet [63].
- SSIM computes the similarity between two images based on their structural similarity, taking into account factors such as luminance, contrast, and structural patterns. Higher SSIM values indicate greater similarity between the two images according to human visual perception.

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

μ_x and μ_y are the mean intensities of images x and y. σ_x and σ_y are the standard deviations of images x and y. σ_{xy} is the covariance between x and y. C_1 and C_2 are small constants to stabilize the division with weak denominator.

Table 1
Metric results evaluated between the following pairs on the: real-MITS-GAN, real-TAFIM, real-MITS-GAN tampered and real-TAFIM tampered. Lower values are better for RMSE and LPIPS, higher for PSNR and SSIM.

Metric	Real		Tampered	
	MITS-GAN	TAFIM	MITS-GAN T.	TAFIM T.
RMSE	169.481	194.943	198.253	233.780
PSNR	27.949	21.702	21.237	21.469
LPIPS	0.170	0.383	0.226	0.391
SSIM	0.983	0.945	0.970	0.981

Table 2
Metric results evaluated between the following pairs on the tampered square part of the images: real-MITS-GAN, real-TAFIM, real-MITS-GAN tampered and real-TAFIM tampered. Lower values are better for RMSE and LPIPS, higher for PSNR and SSIM.

Metric	Real		Tampered	
	MITS-GAN	TAFIM	MITS-GAN T.	TAFIM T.
RMSE	50.565	66.061	84.349	79.451
PSNR	26.682	18.854	11.289	18.511
LPIPS	0.372	0.3417	0.591	0.346
SSIM	0.992	0.972	0.740	0.866

4.3. Experimental setup

All models were trained for 20 epochs using a NVIDIA V100. The MITS-GAN[2] architecture, implemented using PyTorch,[3] was trained with a batch size of 16, a learning rate set at 0.0002, betas of $[0.5, 0.999]$, and utilizing Adam as the optimizer. For TAFIM, we adopted the configurations suggested by the authors in [25].

4.4. Results

Fig. 6 shows the qualitative results of the proposed MITS-GAN method compared with TAFIM [25]. MITS-GAN exhibits fewer visible artifacts on the reconstructed images and demonstrates a more robust ability to resist manipulation, accentuating the artifacts introduced when the model attempts to manipulate the selected square. Fig. 7 shows the heatmap obtained by performing a pixel-to-pixel difference between the real image and the protected one. In this case, the proposed method generates protected images that are more faithful to the originals than the compared method.

Table 1 reports the results of the considered metrics evaluated between each pair of real-protected and real-protected/tampered on the entire images. MITS-GAN has lower RMSE, LPIPS, and higher PSNR and

[2] https://github.com/GiovanniPasq/MITS-GAN
[3] https://pytorch.org/

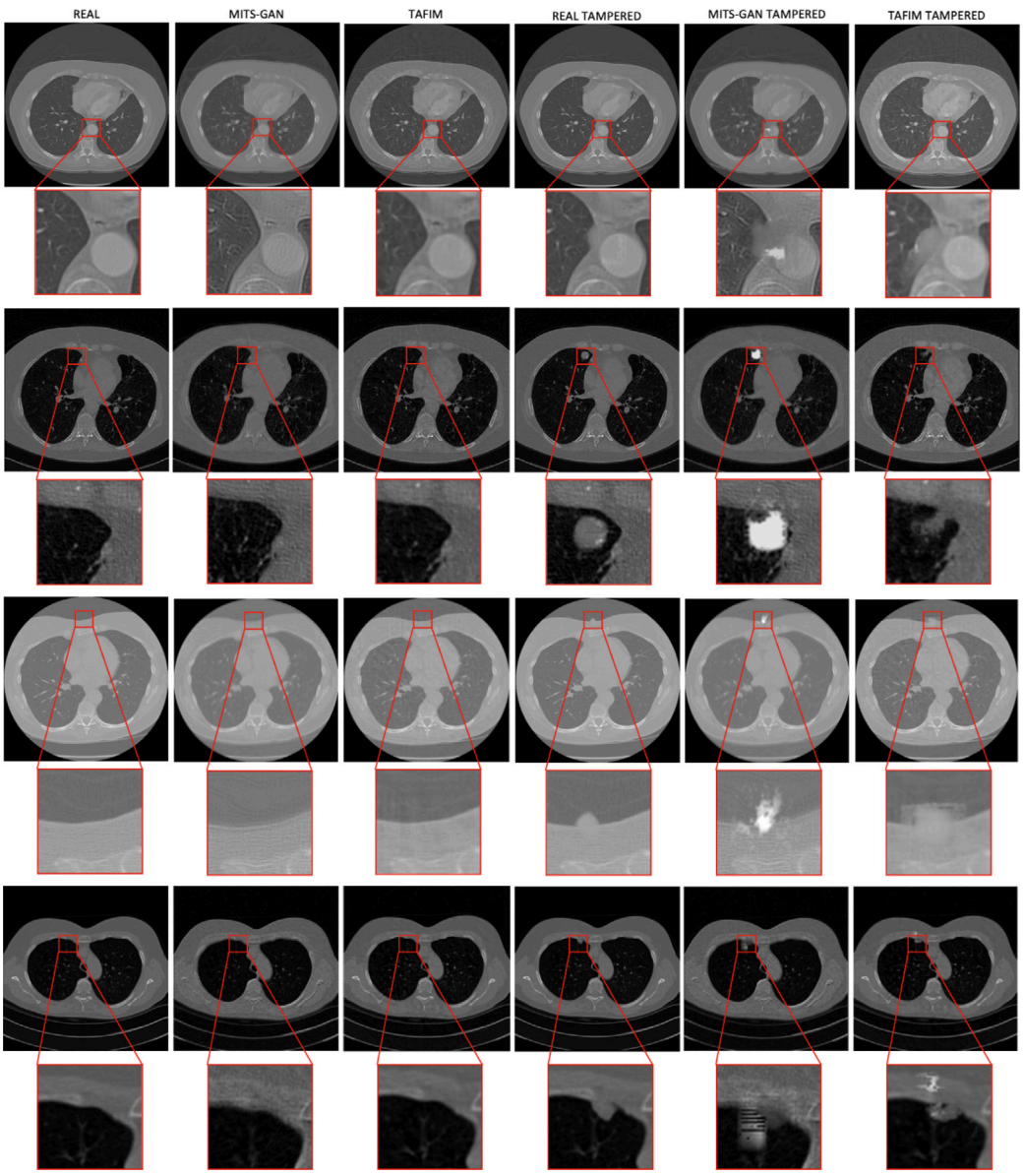

Fig. 6. Qualitative results on the reconstruction task compared with images as manipulation targets.

SSIM values compared to TAFIM, suggesting better reconstruction quality of the images. This advantage is maintained even when considering the images after manipulation. Table 2 shows the results evaluated on the square part subjected to manipulation. In this case, the metrics favor the proposed method. After manipulation, the output produced by the manipulator model appears to be more damaged than the compared method. This suggests that MITS-GAN produces images with less noise but is more robust to manipulation, generating more visible artifacts when attempting to tamper with an image.

4.5. Ablation study

Table 3 presents the results of MITS-GAN varying the hyperparameter α, which regulates the standard GAN losses and the MSE loss used to generate robust images against manipulation by CT-GAN. Since CT-GAN performs manipulation on a square of size 32 × 32 pixels, the evaluation considers which α value provides the best protection. This assessment focuses on maximizing RMSE and LPIPS while minimizing PSNR and SSIM. The goal is to ensure that the output generated by CT-GAN after manipulation is significantly different from the original, introducing artifacts that are clearly visible to the human eye. As shown in the table, the best performance is achieved when $\alpha = 1$.

Table 3
Ablation study about the impact of the MSE loss.

Metric	α				
	0.2	0.4	0.6	0.8	1
RMSE	79.026	80.472	81.920	82.517	84.349
PSNR	18.766	17.145	15.803	13.562	11.289
LPIPS	0.338	0.377	0.425	0.510	0.591
SSIM	0.881	0.854	0.810	0.775	0.740

5. Discussion

The MITS-GAN approach has shown considerable promise in safeguarding medical imaging from tampering, particularly when compared to existing methods such as TAFIM. Experimental results show that MITS-GAN achieves lower RMSE and LPIPS values and higher

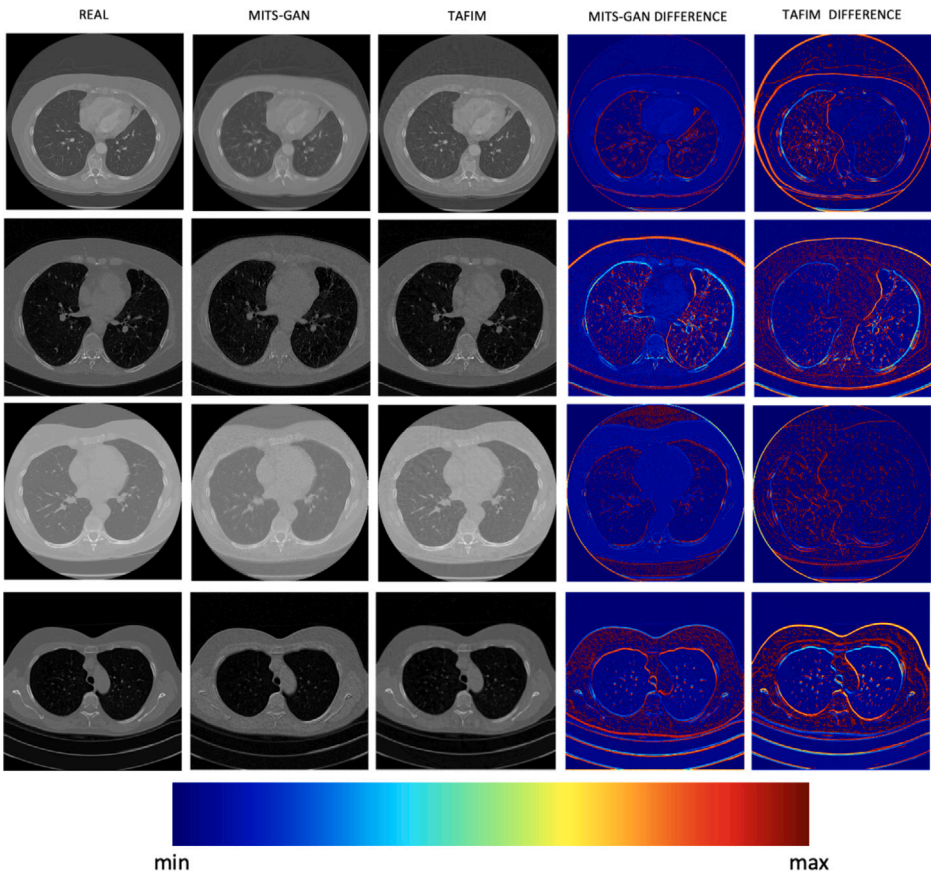

Fig. 7. Heatmap computed between the pairs real-MITS-GAN and real-TAFIM.

PSNR and SSIM values, indicating superior image reconstruction quality and robustness against manipulation. MITS-GAN succeeds in creating high-quality images that are almost completely identical to the originals and with almost no artifacts. This robustness is crucial in medical imaging where clarity and accuracy are fundamental. In addition, the method generates tamper-resistant images, showing more visible artifacts when data protected by MITS-GAN is tampered with by other architectures, making it easier to detect non-authorized alterations. Despite these strengths, some limitations and potential areas for improvement can be identified:

- *Sensitivity of hyper-parameters*: MITS-GAN's performance largely depends on the careful tuning of hyper-parameters, such as the α-value that balances GAN loss and MSE loss. Incorrect tuning can have a significant impact on the effectiveness of the model.
- *Computational complexity*: MITS-GAN training requires some computational resources, including high-performance GPUs and extended training times, which may limit its accessibility and implementation in resource-limited environments.

It is important to note that the computational problems (in terms of time) are mainly related to the model training procedure. The protection of CT scans using the MITS-GAN method does not require heavy computational resources because real-time protection during acquisition is unnecessary. This approach allows for protection to be performed later, in the background, without impacting the primary image acquisition process. Therefore, from the standpoint of scalability and computational efficiency, MITS-GAN proves suitable for practical applications in the medical domain, enabling efficient resource management without compromising service quality.

Future works will focus on improving the MITS-GAN architecture considering:

- *Integration of Diffusion Models* One promising direction involves integrating diffusion models into the MITS-GAN framework. Diffusion models, known for iteratively adding noise to images, could contribute to improving the quality and authenticity of safeguarded medical imagery generated by MITS-GAN.
- *Attention Mechanisms for Robustness* To fortify MITS-GAN against malicious tampering, future work could incorporate attention mechanisms. Attention mechanisms enable the model to focus on relevant regions of the input, potentially making it more resilient to adversarial attacks and ensuring critical details in medical scans are preserved.
- *Exploring Diverse Architectures* The success of MITS-GAN opens the door to exploring diverse generative model architectures. Investigating different GAN variants or hybrid architectures could provide valuable insights into optimizing the trade-off between image quality, computational efficiency, and security.
- *Real-world Deployment and Validation* A crucial step toward practical application involves focusing on real-world deployment and validation of MITS-GAN. Collaborations with healthcare institutions and professionals can provide valuable feedback, ensuring that the proposed method aligns with the practical requirements and standards of the medical imaging community.

6. Conclusion

In this work, we introduced MITS-GAN, an innovative approach to safeguard medical imagery against malicious tampering. The method demonstrated superior performance in disrupting manipulations at the source, resulting in the generation of tamper-resistant images with fewer artifacts when compared to existing techniques. The proactive measures outlined in this study hold significant importance in guaranteeing the responsible and ethical use of generative models, particularly

in critical applications such as healthcare. By addressing the vulnerabilities in medical imaging systems, MITS-GAN contributes to the overall resilience of these systems against potential threats. Looking ahead, future works and potential extensions aim to further refine and enhance the capabilities of MITS-GAN. This ongoing research aligns with our commitment to staying at the forefront of advancements in securing medical imaging technology. By continually pushing the boundaries of innovation, we aim to make meaningful contributions that strengthen the integrity and reliability of healthcare systems, and ensuring the trustworthiness of medical diagnostic tools.

CRediT authorship contribution statement

Giovanni Pasqualino: Writing – original draft, Validation, Software, Data curation. **Luca Guarnera:** Writing – review & editing, Methodology, Investigation. **Alessandro Ortis:** Methodology, Investigation, Formal analysis, Conceptualization. **Sebastiano Battiato:** Writing – review & editing, Project administration.

Declaration of competing interest

None.
All authors affirm that there are no interests to declare.

Acknowledgments

This research is supported by research Program PIAno di inCEntivi per la Ricerca di Ateneo 2020/2022 — Linea di Intervento 3 “Starting Grant ” - SAFE-IA Project, University of Catania, Italy. This research is supported by Azione IV.4 - “Dottorati e contratti di ricerca su tematiche dell’innovazione” del nuovo Asse IV del PON Ricerca e Innovazione 2014–2020 “Istruzione e ricerca per il recupero - REACT-EU”- CUP: E65F21002580005.

References

[1] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[2] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2018, arXiv preprint arXiv:1809.11096.
[3] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188–8197.
[4] S. Shahriar, GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network, Displays 73 (2022) 102237.
[5] A. Madani, M. Moradi, A. Karargyris, T. Syeda-Mahmood, Chest X-ray generation and data augmentation for cardiovascular abnormality classification, in: Medical Imaging 2018: Image Processing, vol. 10574, SPIE, 2018, pp. 415–420.
[6] C.-B. Jin, H. Kim, M. Liu, W. Jung, S. Joo, E. Park, Y.S. Ahn, I.H. Han, J.I. Lee, X. Cui, Deep CT to MR synthesis using paired and unpaired data, Sensors 19 (10) (2019) 2361.
[7] C. Bermudez, A.J. Plassard, L.T. Davis, A.T. Newton, S.M. Resnick, B.A. Landman, Learning implicit brain MRI manifolds with deep learning, in: Medical Imaging 2018: Image Processing, vol. 10574, SPIE, 2018, pp. 408–414.
[8] R. Gupta, A. Sharma, A. Kumar, Super-resolution using GANs for medical imaging, Procedia Comput. Sci. 173 (2020) 28–35.
[9] J. Islam, Y. Zhang, Gan-based synthetic brain pet image generation, Brain Inform. 7 (1) (2020) 3.
[10] F. Li, W. Huang, M. Luo, P. Zhang, Y. Zha, A new vae-gan model to synthesize arterial spin labeling images from structural mri, Displays 70 (2021) 102079.
[11] T. Pang, J.H.D. Wong, W.L. Ng, C.S. Chan, Semi-supervised gan-based radiomics model for data augmentation in breast ultrasound mass classification, Comput. Methods Programs Biomed. 203 (2021) 106018.
[12] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, X. Yang, Ct synthesis from mri using multi-cycle gan for head-and-neck radiation therapy, Comput. Med. Imaging Graph. 91 (2021) 101953.
[13] M. Tian, K. Song, Boosting magnetic resonance image denoising with generative adversarial networks, IEEE Access 9 (2021) 62266–62275.
[14] X. Dong, Y. Lei, T. Wang, M. Thomas, L. Tang, W.J. Curran, T. Liu, X. Yang, Automatic multiorgan segmentation in thorax ct images using u-net-gan, Med. Phys. 46 (5) (2019) 2157–2168.
[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[16] W. Hu, Y. Tan, Generating adversarial malware examples for black-box attacks based on GAN, in: International Conference on Data Mining and Big Data, Springer, 2022, pp. 409–423.
[17] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, Calif. Law Rev. 107 (2019) 1753.
[18] Y. Mirsky, T. Mahler, I. Shelef, Y. Elovici, {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning, in: 28th USENIX Security Symposium, USENIX Security 19, 2019, pp. 461–478.
[19] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15108–15117.
[20] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face X-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.
[21] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 2307–2311.
[22] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of GAN-generated fake images over social networks, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, IEEE, 2018, pp. 384–389.
[23] C.-Y. Yeh, H.-W. Chen, S.-L. Tsai, S.-D. Wang, Disrupting image-translation-based deepfake algorithms with adversarial attacks, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, 2020, pp. 53–62.
[24] N. Ruiz, S.A. Bargal, S. Sclaroff, Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August (2020) 23–28, Proceedings, Part IV, vol. 16, Springer, 2020, pp. 236–251.
[25] S. Aneja, L. Markhasin, M. Nießner, TAFIM: Targeted adversarial attacks against facial image manipulations, in: European Conference on Computer Vision, Springer, 2022, pp. 58–75.
[26] L. Bi, J. Kim, A. Kumar, D. Feng, M. Fulham, Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs), in: Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment: Fifth International Workshop, CMMI 2017, Second International Workshop, RAMBO 2017, and First International Workshop, SWITCH 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14 2017, Proceedings, vol. 5, Springer, 2017, pp. 43–51.
[27] A. Ben-Cohen, E. Klang, S.P. Raskin, M.M. Amitai, H. Greenspan, Virtual PET images from CT data using deep convolutional networks: Initial results, in: Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10 2017, Proceedings, vol. 2, Springer, 2017, pp. 49–57.
[28] A. Ben-Cohen, E. Klang, S.P. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M.M. Amitai, H. Greenspan, Cross-Modality Synthesis from CT to PET Using FCN and GAN Networks for Improved Automated Lesion Detection, Eng. Appl. Artif. Intell. 78 (2019) 186–194.
[29] Q. Dou, C. Ouyang, C. Chen, H. Chen, P.-A. Heng, Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss, 2018, arXiv preprint arXiv:1804.10916.
[30] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.
[31] J.M. Wolterink, T. Leiner, I. Isgum, Blood vessel geometry synthesis using generative adversarial networks, 2018, arXiv preprint arXiv:1804.04381.
[32] Z. Han, B. Wei, A. Mercado, S. Leung, S. Li, Spine-GAN: Semantic segmentation of multiple spinal structures, Med. Image Anal. 50 (2018) 23–35.
[33] T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 146–157.
[34] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Appl. Intell. 53 (4) (2023) 3974–4026.
[35] L. Guarnera, O. Giudice, C. Nastasi, S. Battiato, Preliminary forensics analysis of deepfake images, in: 2020 AEIT International Annual Conference, AEIT, IEEE, 2020, pp. 1–6, http://dx.doi.org/10.23919/AEIT50178.2020.9241108.
[36] R. Durall, M. Keuper, J. Keuper, Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 7887–7896, http://dx.doi.org/10.1109/CVPR42600.2020.00791.
[37] O. Giudice, L. Guarnera, S. Battiato, Fighting deepfakes by detecting GAN DCT anomalies, J. Imaging 7 (8) (2021) 128, http://dx.doi.org/10.3390/jimaging7080128, https://www.mdpi.com/2313-433X/7/8/128.

[38] N. Yu, L. Davis, M. Fritz, Attributing fake images to GANs: Learning and analyzing GAN fingerprints, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 7555–7565.
[39] L. Guarnera, O. Giudice, S. Battiato, DeepFake detection by analyzing convolutional traces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 666–667.
[40] L. Guarnera, O. Giudice, S. Battiato, Fighting deepfake by exposing the convolutional traces on images, IEEE Access 8 (2020) 165085–165098.
[41] T.K. Moon, The expectation-maximization algorithm, IEEE Signal Process. Mag. 13 (6) (1996) 47–60.
[42] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.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, 2020, pp. 8695–8704.
[43] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, Vol. 2018, ICLR, 2018, pp. 1–26.
[44] D.A. Coccomini, N. Messina, C. Gennaro, F. Falchi, Combining EfficientNet and vision transformers for video deepfake detection, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 219–229.
[45] Y.-J. Heo, W.-H. Yeo, B.-G. Kim, Deepfake detection algorithm based on improved vision transformer, Appl. Intell. 53 (7) (2023) 7512–7527.
[46] D. Wodajo, S. Atnafu, Deepfake video detection using convolutional vision transformer, 2021, arXiv preprint arXiv:2102.11126.
[47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning, PMLR, 2015, pp. 2256–2265.
[48] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. 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, IEEE, 2023, pp. 1–5.
[49] Z. Sha, Z. Li, N. Yu, Y. Zhang, DE-FAKE: Detection and attribution of fake images generated by text-to-image diffusion models, 2022, arXiv preprint arXiv:2210.06998.
[50] L. Guarnera, O. Giudice, S. Battiato, Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images, ACM Trans. Multim. Comput. Commun. Appl. 20 (11) (2024) http://dx.doi.org/10.1145/3652027.
[51] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
[52] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.
[53] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.
[54] V. Fischer, M.C. Kumar, J.H. Metzen, T. Brox, Adversarial examples for semantic image segmentation, 2017, arXiv preprint arXiv:1703.01101.
[55] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille, Adversarial examples for semantic segmentation and object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1369–1378.
[56] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, V. Fischer, Universal adversarial perturbations against semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2755–2764.
[57] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1765–1773.
[58] C.-Y. Yeh, H.-W. Chen, H.-H. Shuai, D.-N. Yang, M.-S. Chen, Attack as the best defense: Nullifying image-to-image translation GANs via limit-aware adversarial attack, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16188–16197.
[59] Q. Huang, J. Zhang, W. Zhou, W. Zhang, N. Yu, Initiative defense against facial manipulation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1619–1627.
[60] Y. Shi, H. Tang, M.J. Baine, M.A. Hollingsworth, H. Du, D. Zheng, C. Zhang, H. Yu, 3DGAUnet: 3D generative adversarial networks with a 3D U-net based generator to achieve the accurate and effective synthesis of clinical tumor image data for pancreatic cancer, Cancers 15 (23) (2023) 5496.
[61] S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931.
[62] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[63] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size, 2016, arXiv preprint arXiv:1602.07360.