Diffusion models meet image counter-forensics

Matías Tailanián[†,1,2], Marina Gardella[†,3], Alvaro Pardo[2,4], Pablo Musé[1]
[†]Equally contributing authors.
[1]IIE, Facultad de Ingeniería,Universidad de la República
[2]Digital Sense
[3]Centre Borelli, École Normale Supérieure Paris-Saclay, Université Paris-Saclay
[4]Universidad Católica del Uruguay

Abstract

From its acquisition in the camera sensors to its storage, different operations are performed to generate the final image. This pipeline imprints specific traces into the image to form a natural watermark. Tampering with an image disturbs these traces; these disruptions are clues that are used by most methods to detect and locate forgeries. In this article, we assess the capabilities of diffusion models to erase the traces left by forgers and, therefore, deceive forensics methods. Such an approach has been recently introduced for adversarial purification, achieving significant performance. We show that diffusion purification methods are well suited for counter-forensics tasks. Such approaches outperform already existing counter-forensics techniques both in deceiving forensics methods and in preserving the natural look of the purified images. The source code is publicly available at https://github.com/mtailanian/diff-cf.

1. Introduction

Image forgeries are present everywhere [22], from fake news on social media [45] to scientific misconduct. Indeed, many image processing tools are available to create visually realistic image alterations. Yet, these modifications leave traces on the image that are tampering cues. Image forensics aims at detecting these alterations by finding local inconsistencies [22]. Image counter-forensics emerged as the research field that challenges forensics methods and explores their limitations [7].

Adversarial attacks share some common properties with image forgeries in the sense that both techniques introduce subtle modifications to the images that, though imperceptible to the naked eye, disrupt the image's traces. The goal of adversarial attacks is to deceive a model into making incorrect predictions. Adversarial purification can be, therefore, linked to counter-forensics since it aims at preprocessing the input data to remove these adversarial perturbations. Generally, these purification methods are based on generative models [46].

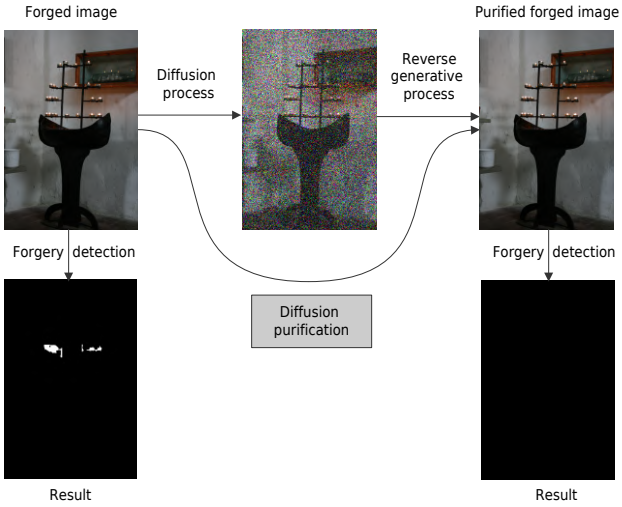

Figure 1. Illustration of using diffusion models as a counter-forensic technique. A forged image from FAU dataset [12], correctly detected by ZERO [44], produces no detection after diffusion purification.

In recent years, diffusion models have emerged as highly effective generative models [26, 50]. These models have showcased impressive capabilities in generating high-quality samples, outperforming traditional Generative Adversarial Networks (GANs) in the realm of image generation. The advancements in diffusion models have led to significant improvements in the fidelity and realism of synthesized images, highlighting their potential as state-of-the-art models in the field.

This work has received funding by the Paris Region Ph.D. grant from Région Île-de-France, the ANR project APATE (ANR-22-CE39-0016), the European Union under the Horizon Europe VERA.AI project, Grant Agreement number 101070093 and by a graduate scholarship from Agencia Nacional de Investigación e Innovación, Uruguay.

In this work, we evaluate, for the first time, the efficiency of diffusion purification methods, currently used for adversarial purification [43, 52], as counter-forensics methods. The rationale behind the use of diffusion models for adversarial purification is that these models learn the distribution of clean data. Hence, by diffusing an adversarial example and then applying the reverse generative process, the diffusion model gradually removes the adversarial perturbations and reconstructs the underlying clean sample.

The same rationale can be applied to hide the forensic traces caused by tampering. Indeed, since diffusion models are trained on pristine images, diffusion purification methods applied to forged images should recover purified images without any inconsistency in the camera traces. Once such disruptions on the camera processing chain are erased, purified images should be able to deceive any forgery detection method relying on them. Fig. 1 shows an example of the aforementioned approach: while ZERO [44] correctly detects the original forgery, once diffusion purification is applied, the method is no longer able to detect it.

2. Related work

2.1. Image counter-forensics

Counter-forensics attacks can be classified into two categories: the first one corresponds to those that focus on a specific trace or method, while the second category corresponds to generic attacks that aim at erasing all the forgery traces and should, therefore, be able to deceive any forensic method. Among methods of the first category, Fan et al. [21] and Comesana et al. [13] propose attacks against histogram-based methods, mainly used to detect JPEG compression traces. Kirchner et al. [28] propose hiding resampling traces by removing the periodic variations in the residual signal in the spatial domain. Do et al. [20] design SIFT-specific attacks that are able to deceive copy-move forgery detectors based on such local descriptors.

With the advent of learning-based forgery detectors, counter-forensic attacks specifically designed for such methods have also been proposed. Marra et al. [39] design a counter-forensic scheme on the feature space. Their goal is to restore the features of the pristine image and, by doing so, to cross the decision boundary of the target detector. In the case of perfect knowledge of the target method, this counter-forensic method delivers great results. However, when the target detector is unknown, the results degrade tremendously. Other methods countering specific learning-based detectors with an optimum attack which relies on gradient descent solutions have also been explored [10, 24].

With limited knowledge of forensic models, counter-forensics attackers focus more on erasing the traces by generic tools [5]. The median filter is a technique commonly used as an anti-forensics attack [59], in deep convolutional neural network versions [27] or even variational formulations [48]. Though this method can be effective on several traces, it leaves a distinctive streaking artifact that can be retrieved [29,61]. To compensate for this, techniques to remove such artifacts have been proposed [23,48].

More recently, Chen et al. [9] proposed erasing camera traces, trying not to damage the signal content by adopting a Siamese-based neural network. Cozzolino et al. [16] and Wu et al. [55] use generative adversarial approaches. Baracchi et al. [4] exploit a real camera firmware to perform the manipulation while reproducing the image statistics. This approach can be most efficient at creating real camera traces, and can easily fool camera identification methods into thinking the image was taken with this camera. However, this method is difficult to use, since it requires disassembling a camera to hack its input field.

2.2. Diffusion-based adversarial purification

Nie et al. [43] were the first ones to propose the use of the forward and reverse processes of a pre-trained diffusion model for image adversarial purification. Their method –DiffPure– first diffuses adversarial examples with a small amount of noise. Then, the clean image is recovered through the reverse generative process. A very similar idea was developed at the same time by Blau et al. [6]. The theoretical fundamentals justifying the performance of such diffusion-based adversarial purification methods are derived in [60].

Wang et al. [52] face the difficult trade-off between choosing a long diffusion time, which guarantees the removal of the adversarial perturbation, and choosing a small one, which guarantees the similarity between the input image and the purified one. They propose to guide the reverse process by the adversarial image. By doing so, the purified image is forced to stay close to the input image.

Wu et al. [56] also guide the reverse process by the adversarial image. However, they propose to sample the initial input from pure Gaussian noise and gradually denoise it. The rationale of their approach is that the diffused image still carries corrupted structures, and the reverse process is likely to get stuck in those corrupted structures.

As the field evolves, several applications of these approaches have been developed. In [1], the authors analyze the performance of DiffPure [43] to purify adversarial attacks on the classification of metastatic tissue. In [51], the authors apply the same principle as in [43] but using an extension of diffusion models to the 3D space [37]. Similarly, [57] also shares the grounds of DiffPure [43] but using a waveform-based diffusion model [30] for adversarial audio purification.

Diffusion purification methods have rapidly gained attention in the field. This interest has even led to questioning the evaluation practices of such techniques [36].

3. Background

In this section, we provide a brief overview of Denoising Diffusion Models [26, 49, 50] that will be used as a basis for the next section. Recently, denoising diffusion models, alternatively called score-based generative models, have emerged as a powerful approach amongst generative methods. Denoising diffusion models consist of two processes: a forward diffusion process that progressively adds noise to the input and a reverse generative process that learns to generate data by denoising.

Forward diffusion process. The diffusion process is a Markov process that gradually adds noise to the clean input data. Let T be the number of steps of the diffusion process, $\mathbf{x}_0$ an input image, and $\mathbf{x}_t$ the forward image until step t ($0 \leq t \leq T$). The diffusion process from clean data $\mathbf{x}_0$ to $\mathbf{x}_T$ is defined as

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

$$\text{with } q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where the variances $\beta_1, \ldots, \beta_T$ are predefined small values.

A notable characteristic of the forward process is that there is a closed-form to generate $\mathbf{x_t}$ at any given time step t directly from $\mathbf{x_0}$ [26]. Indeed, let $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, then we can directly sample $\mathbf{x}_t$ as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{(1-\bar{\alpha})}\,\epsilon, \text{ where } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

Reverse denoising process. The reverse generative process is a Markov process that gradually eliminates the noise added in the forward process. The reverse process from $\mathbf{x}_T$ to $\mathbf{x}_0$ is given by

$$p_\theta(\mathbf{x}_{0:T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{4}$$

$$\text{with } p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 I), \tag{5}$$

where the mean $\mu_\theta(\mathbf{x}_t, t)$ is a trainable network and the variances $\sigma_0^2, \ldots, \sigma_T^2$ can either be fixed or learned using a neural network.

4. Proposed method

Our goal is to introduce subtle modifications to a forged image to erase the traces left by the tampering process while, at the same time, preserving the semantic content. Our proposed approach is based on diffusion purification methods [43, 52, 56]. It consists of two steps: first, we add noise up to a certain time-step $t = t^*$ in the forward diffusion process, and then we gradually remove it following the reverse diffusion process, up to $t = 0$. We refer to this method as *Diffusion Counter-Forensics*, or shortly *Diff-CF*.

The intuition behind this idea is that the probability distributions of the forged and its corresponding clean image are separated in $t = 0$, but by adding noise in the forward process, the boundaries between the distributions get fuzzier, and they begin to overlap, more so the higher the value of t^*. Then, starting from a noisy sample that can belong to either probability distribution, the reverse diffusion process, which was trained on pristine images only, generates a purified version of the image with no forgery traces. See, for instance, Fig. 2 in [40].

The value t^* plays a fundamental role. Intuitively, t^* has to be large enough so that the noise added hides the forgery traces, but small enough so that we can preserve the image semantics and structure. If we set the value of t^* too high, the resulting image would deviate too much from the original one. On the other hand, if the value of t^* is too small, we might be unable to erase the forgery traces correctly. This trade-off is studied more in-depth in Sec. 5.3.

With the purpose of being able to use larger values of t^* without deviating too much from the input image, we also analyze the introduction of guidance in the reverse diffusion process. We refer to this variant as *Counter-Forensics Guided Diffusion*, or *Diff-CFG*. More precisely, we propose to guide the reverse process using the forged image itself, as in [52]. In this way, we encourage the network to produce a clean image as close as possible to the forged one, under the assumption that the forgery traces are subtle enough that they are not reconstructed. In the normal reverse diffusion process, at each time step, a new image is sampled following Eq. 5. Instead, for this variant, we propose to sample from

$$\begin{aligned} &p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \\ &\quad = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t) - s_t \Sigma \nabla_{x_t} \mathcal{D}(x_t, x_{in}), \sigma_t^2 I), \end{aligned} \tag{6}$$

where Σ is the variance of x_t, $\mathcal{D}(x_t, x_{in})$ is some similarity measure between x_t and the input image (forged image) x_{in}, and s_t is a scale factor that depends on the time step t. For high values of t, the forgery traces are completely hidden by the added noise, so we can afford to use large values for s_t, without the risk of guiding the process to reconstruct the forged traces. On the other hand, for small values of t, the forgery traces are more retained, and therefore we should use smaller values for s_t. Similar to what is proposed in [52, 56], we define s_t to be proportional to the added noise, as

$$s_t = s\,\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}, \tag{7}$$

where s is a hyper-parameter.

		CatNet	Choi	Comprint	MantraNet	Noiseprint	Shin	SpliceBuster	TruFor	ZERO	Avg$_w$
Korus	Original	0.0790	0.1971	0.0534	0.1261	0.0988	0.0221	0.1405	0.3428	0.0050	-
		0.0433	0.1261	0.0461	0.0982	0.0792	0.0568	0.1012	0.2575	0.0028	-
	CamTE	**0.0468** (-0.0322)	0.0597 (-0.1374)	0.0356 (-0.0179)	0.0646 (-0.0614)	0.0420 (-0.0569)	0.0305 (0.0084)	0.0817 (-0.0588)	**0.1961** (-0.1467)	0.0000 (-0.0050)	-0.1024
		0.0278 (-0.0155)	0.0400 (-0.0860)	0.0389 (-0.0072)	0.0644 (-0.0338)	0.0545 (-0.0247)	0.0578 (0.0011)	0.0729 (-0.0284)	**0.1489** (-0.1086)	0.0000 (-0.0028)	-0.0479
	BM3D	0.0997 (0.0207)	0.0352 (-0.1619)	0.0278 (-0.0256)	0.0652 (-0.0609)	0.0420 (-0.0569)	0.0155 (-0.0066)	0.0860 (-0.0545)	0.2579 (-0.0850)	0.0000 (-0.0050)	-0.0819
		0.0646 (0.0213)	**0.0227** (-0.1033)	0.0346 (-0.0115)	0.0746 (-0.0237)	0.0514 (-0.0278)	0.0540 (-0.0028)	0.0744 (-0.0268)	0.1964 (-0.0611)	0.0000 (-0.0028)	-0.0358
	Diff-CF	**0.0418** (-0.0371)	**0.0147** (-0.1825)	**0.0024** (-0.0510)	**0.0255** (-0.1005)	**0.0190** (-0.0798)	0.0027 (-0.0194)	**0.0350** (-0.1055)	**0.1454** (-0.1974)	0.0045 (-0.0005)	**-0.1451**
		0.0204 (-0.0229)	0.0246 (-0.1014)	**0.0215** (-0.0246)	**0.0416** (-0.0566)	**0.0360** (-0.0432)	0.0487 (-0.0081)	**0.0401** (-0.0611)	**0.1131** (-0.1445)	0.0027 (-0.0001)	**-0.0677**
	Diff-CFG	0.0852 (0.0063)	**0.0044** (-0.1927)	**0.0125** (-0.0409)	**0.0442** (-0.0818)	**0.0267** (-0.0722)	0.0040 (-0.0181)	**0.0456** (-0.0950)	0.2064 (-0.1364)	0.0005 (-0.0045)	**-0.1177**
		0.0527 (0.0095)	**0.0043** (-0.1217)	**0.0281** (-0.0180)	**0.0552** (-0.0430)	**0.0446** (-0.0346)	0.0491 (-0.0076)	**0.0488** (-0.0525)	0.1601 (-0.0975)	0.0011 (-0.0017)	**-0.0536**
FAU	Original	0.3228	0.3045	0.0393	0.0203	0.0358	0.1134	0.0074	0.4039	0.5855	-
		0.2329	0.2670	0.0305	0.0336	0.0482	0.1289	0.0251	0.3373	0.5003	-
	CamTE	**0.0141** (-0.3087)	0.1426 (-0.1620)	0.0092 (-0.0302)	0.0154 (-0.0049)	0.0242 (-0.0116)	0.0826 (-0.0308)	0.0045 (-0.0029)	0.0553 (-0.3486)	0.0441 (-0.5414)	-0.6120
		0.0085 (-0.2244)	0.1173 (-0.1497)	0.0206 (-0.0099)	0.0427 (0.0091)	0.0434 (-0.0049)	0.1046 (-0.0243)	0.0277 (0.0026)	**0.0572** (-0.2801)	0.0288 (-0.4715)	-0.4259
	BM3D	0.0757 (-0.2471)	0.0679 (-0.2367)	**-0.0017** (-0.0410)	-0.0268 (-0.0470)	**-0.0014** (-0.0372)	0.0411 (-0.0723)	0.0011 (-0.0064)	0.0802 (-0.3237)	**0.0393** (-0.5462)	-0.6145
		0.0517 (-0.1812)	0.0559 (-0.2111)	**0.0126** (-0.0179)	0.0377 (0.0041)	0.0331 (-0.0152)	0.0803 (-0.0486)	0.0243 (-0.0008)	0.0799 (-0.2574)	0.0266 (-0.4737)	-0.4298
	Diff-CF	**0.0070** (-0.3157)	**0.0242** (-0.2803)	0.0001 (-0.0392)	0.0057 (-0.0146)	**-0.0018** (-0.0376)	**0.0128** (-0.1006)	-0.0050 (-0.0124)	**0.0399** (-0.3640)	-0.0007 (-0.5862)	**-0.6922**
		0.0056 (-0.2273)	**0.0458** (-0.2213)	0.0159 (-0.0146)	0.0355 (0.0019)	**0.0199** (-0.0283)	**0.0602** (-0.0687)	0.0123 (-0.0128)	**0.0520** (-0.2853)	**0.0015** (-0.4988)	**-0.4687**
	Diff-CFG	0.0241 (-0.2986)	**0.0137** (-0.2908)	**-0.0059** (-0.0452)	0.0128 (-0.0075)	0.0002 (-0.0355)	**0.0202** (-0.0933)	0.0127 (0.0053)	**0.0470** (-0.3569)	**-0.0043** (-0.5898)	**-0.6882**
		0.0184 (-0.2145)	**0.0220** (-0.2451)	**0.0143** (-0.0162)	0.0339 (0.0003)	**0.0287** (-0.0196)	**0.0646** (-0.0643)	0.0246 (-0.0005)	0.0592 (-0.2781)	**0.0008** (-0.4995)	**-0.4688**
COVERAGE	Original	0.2747	0.0075	0.0230	0.2617	0.0062	0.0615	-0.0571	0.4442	0.0082	-
		0.2199	0.0109	0.0856	0.1856	0.0858	0.1106	0.0423	0.3752	0.0070	-
	CamTE	**0.1480** (-0.1267)	0.0056 (-0.0020)	-0.0015 (-0.0245)	0.0790 (-0.1827)	-0.0230 (-0.0292)	**0.0489** (-0.0127)	-0.0722 (-0.0151)	**0.2614** (-0.1828)	0.0000 (-0.0082)	**-0.1646**
		0.1162 (-0.1038)	0.0079 (-0.0030)	0.0711 (-0.0145)	0.0719 (-0.1137)	0.0770 (-0.0089)	**0.1043** (-0.0063)	0.0361 (-0.0062)	**0.2212** (-0.1541)	0.0000 (-0.0070)	**-0.1048**
	BM3D	0.2666 (-0.0081)	0.0051 (-0.0024)	-0.0281 (-0.0511)	**0.0371** (-0.2246)	-0.0145 (-0.0207)	0.0515 (-0.0100)	-0.0771 (-0.0200)	0.3267 (-0.1175)	0.0000 (-0.0082)	-0.1141
		0.2151 (-0.0049)	0.0036 (-0.0072)	0.0617 (-0.0240)	0.0841 (-0.1015)	0.0773 (-0.0085)	0.1055 (-0.0051)	0.0336 (-0.0087)	0.2863 (-0.0889)	0.0000 (-0.0070)	-0.0571
	Diff-CF	**0.1598** (-0.1149)	0.0011 (-0.0064)	-0.0065 (-0.0295)	**0.0483** (-0.2133)	-0.0115 (-0.0176)	0.0514 (-0.0101)	-0.0602 (-0.0031)	0.2849 (-0.1594)	0.0000 (-0.0082)	**-0.1595**
		0.1278 (-0.0922)	0.0059 (-0.0050)	0.0687 (-0.0169)	**0.0537** (-0.1320)	0.0790 (-0.0068)	0.1055 (-0.0051)	0.0383 (-0.0040)	0.2427 (-0.1325)	0.0000 (-0.0070)	**-0.0974**
	Diff-CFG	0.2003 (-0.0745)	-0.0004 (-0.0079)	-0.0124 (-0.0354)	0.0680 (-0.1937)	0.0024 (-0.0038)	**0.0475** (-0.0140)	-0.0717 (-0.0146)	**0.2738** (-0.1704)	0.0000 (-0.0082)	-0.1478
		0.1607 (-0.0592)	0.0010 (-0.0099)	0.0630 (-0.0226)	**0.0693** (-0.1163)	0.0858 (0.0000)	**0.1051** (-0.0055)	0.0334 (-0.0090)	**0.2386** (-0.1367)	0.0000 (-0.0070)	-0.0890

Table 1. IoU and MCC results for Korus [31, 32], FAU [12] and COVERAGE [54] datasets and all methods, except for Bammey *et al*. [2]. For each dataset, we present in the first row the performance of the forgery detectors over the original images. Then, in the following rows, we show the performance of the same detectors over the considered counter-forensic versions of the images, and the difference to the original performance (metric $_{CF}$ − metric $_{orig}$). The lower this difference is, the better the counter-forensic method erased the forgery traces. The best two scores are shown in bold and underlined for each database. For the sake of readability, methods that are not able to obtain a reasonable performance over the original dataset (MCC < 0.03) are grayed out. Bammey *et al*. [2] is excluded from this table, as it was not able to obtain an acceptable performance over any of the considered datasets. The last column (Avg$_w$), is the average of the differences metric $_{CF}$ − metric $_{orig}$, weighted by the performance in the original dataset.

For all experiments, we used the following values: $t^* = 40$, $s = 10^6$, and $\mathcal{D} =$ – SSIM [53] as the guidance metric. A detailed discussion on the influence of the hyper-parameters is presented in Sec. 5.3. In all cases, the images are divided into patches of 256×256 pixels before running the diffusion process. As for the diffusion model, we used a pre-trained class unconditional checkpoint[1].

[1] https://github.com/openai/guided-diffusion

5. Experiments

To assess the performance of the proposed approaches, we compared both the non-guided (*Diff-CF*) and the guided (*Diff-CFG*) variants with the Camera Trace Erasing technique (CamTE) [9] and with BM3D [18, 35]. While comparison with a plain denoiser is not a common practice in the field, we believe it should be included. Indeed, camera traces are a sort of noise in the sense they produce varia-

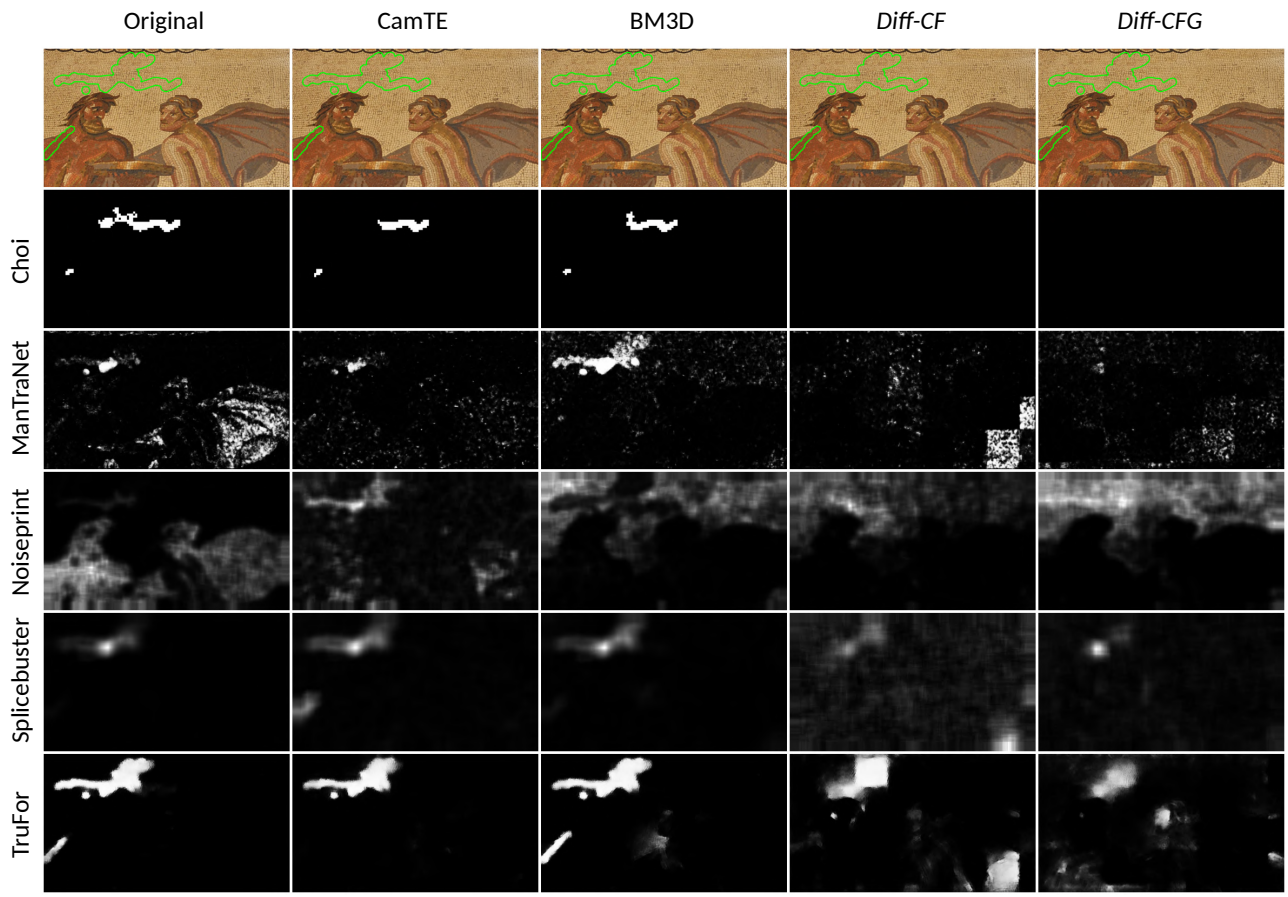

Figure 2. Results obtained by different forensics methods on the different versions of image `r7710a7fat` from the Korus dataset [31, 32]. We observe that Choi [11], ManTraNet [58], and Noiseprint [17] feature no detection when *Diff-CF* or *Diff-CFG* are applied. For Splicebuster [15] and TruFor [25], even if counter-forensics techniques are not completely able to deceive them, the proposed approaches degrade their detections the most. More examples are included in the supplementary materials.

tions in the pixel's values that are not related to the captured scene. On the other hand, we excluded from the comparison the median filtering, which is a widespread technique in counter-forensics, since it was shown to be outperformed by CamTE [9].

We ran our comparisons in four image forgery detection benchmark datasets: Korus [31, 32], FAU [12], COVERAGE [54] and DSO-1 [19]. Since most methods except for Bammey *et al*. [2] deliver poor detection results on the DSO dataset, we decided to exclude them from the main article and report them in the supplementary material.

The goal of counter-forensics methods is to erase all the traces left by the tampering process while preserving the image structures and their semantic content. Therefore, we evaluate two aspects of the counter-forensics techniques under analysis. First, how effectively they hide the forgeries (Sec. 5.1) and second, the quality of the purified images (Sec. 5.2).

5.1. Forgery traces removal

The first point to evaluate is how well the proposed approaches remove the forgery traces. To do so, we ran several state-of-the-art forgery detection methods on the original datasets as well as in their counter-forensics versions (images purified using different techniques). To evaluate their capability of deceiving the forensics methods, we look at the difference between the detection performance before and after purification. The forensics methods that were used are: ZERO [44], Noiseprint [17], Splicebuster [15], ManTraNet [58], Choi [3, 11], Bammey [2], Shin [47], Comprint [38], CAT-Net [33, 34] and TruFor [25]. A brief description of each method can be found in the supplementary material.

To measure detections, we provide scores with the Intersection over Union (IoU) and the Matthews Correlation Coefficient (MCC). F1 scores are not included in the main article but are available in the supplementary material. In terms

of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the IoU is the ratio between the number of pixels in the intersection of detected samples and of ground-truth-positive samples and the number of pixels in the union of these sets. On the other hand, the MCC represents the correlation between the ground truth and detections. The definitions of both the IoU and MCC scores are given in the supplementary material.

The scores were computed for each image and then averaged over each dataset. As most surveyed methods do not provide a binary output but a heat map, to adapt the metrics to the continuous setting, we used their weighted version. We regard the value of a heat map H at each pixel u as the probability of forgery of the pixel. Therefore, given the ground truth mask M, we define the *weighted* TP, *weighted* FP, *weighted* TN and *weighted* FN as:

$$TP_w = \sum_u H(u) \cdot M(u), \tag{8}$$

$$FP_w = \sum_x (1 - H(u)) \cdot M(u), \tag{9}$$

$$TN_w = \sum_x (1 - H(u)) \cdot (1 - M(u)), \tag{10}$$

$$FN_w = \sum_x H(u) \cdot (1 - M(u)). \tag{11}$$

IoU and MCC results for all datasets and all methods are presented in Tab. 1. For each dataset, we present in the first row the performance of the forgery detectors over the original images. Then, in the following rows, we show the performance of the same detectors over the considered counter-forensic versions of the images and the difference to the original performance (metric $_{purified}$ − metric $_{orig}$). The lower this difference is, the better the counter-forensic method erased the forgery traces.

The results in Tab. 1 show that the proposed counter-forensic methods based on diffusion models outperform other counter-forensic techniques in most cases. Indeed, except for the COVERAGE dataset [54] where our methods rank second and third (after CamTE), in all the rest of the datasets *Diff-CF* and *Diff-CFG* achieve the best score reductions. When comparing *Diff-CF* to *Diff-CFG*, we observe that the non-guided version delivers, in most cases, the best results as a counter-forensic method. This can be explained by the fact that when we do not condition the method, the reverse generative process is able to get closer to the distribution of the clean training data.

Regarding the forensic methods individually, we observe that TruFor [25] outperforms the rest of the methods in most cases. Furthermore, it is the only method that still performs acceptably after applying counter-forensics attacks, except on the FAU dataset [12]. Indeed, in this case, once counter-forensic methods are applied, the method delivers highly deteriorated results.

		NIQE (▼)	BRISQUE (▼)	LPIPS (▼)	PSNR (▲)	SSIM (▲)
Korus	Original	5.7271	13.7602	0.0000	80.0000	1.0000
	CamTE	5.5442	34.5632	0.1684	**38.2833**	**0.9433**
	BM3D	5.1004	38.0418	0.0835	**43.1409**	**0.9802**
	Diff-CF	**3.8693**	**23.1161**	**0.0733**	32.9680	0.8769
	Diff-CFG	**4.1070**	**28.3290**	**0.0771**	34.3391	0.9126
FAU	Original	4.7392	20.5726	0.0000	80.0000	1.0000
	CamTE	5.8360	40.1577	0.2098	**37.8765**	**0.9460**
	BM3D	5.4875	42.7470	0.1045	**41.2625**	**0.9797**
	Diff-CF	**3.8896**	**19.8268**	**0.0985**	33.0308	0.8792
	Diff-CFG	**4.2440**	**29.9920**	**0.0952**	34.4725	0.9159
COVERAGE	Original	4.5529	19.0256	0.0000	80.0000	1.0000
	CamTE	5.4513	30.3558	0.0631	**35.7974**	**0.9648**
	BM3D	5.8792	35.9560	**0.0237**	**44.1417**	**0.9888**
	Diff-CF	**4.3343**	**17.1298**	0.0281	33.4959	0.9275
	Diff-CFG	**5.0359**	**27.8903**	**0.0276**	34.6969	0.9487

Table 2. Image quality assessment results of the evaluated counter-forensics techniques. The ▼ indicates that the lower the score the better while the ▲ indicates that the higher the score the better. The best two scores are shown in bold for each database. For the no-reference metrics NIQE and BRISQE, the proposed diffusion-based counter-forensics methods achieve the best performance.

Fig. 2 shows an example of the results obtained by different forensics methods on the different versions of the same forged image. We observe that Choi delivers nearly the same result as in the original forgery when CamTE or BM3D are applied. However, it features no detection when *Diff-CF* or *Diff-CFG* are used as counter-forensics techniques. Noiseprint and ManTraNet provide better detections when CamTE or BM3D are applied, respectively. However, no detection is made when using the proposed approaches. On the other hand, none of the counter-forensics methods is able to deceive Splicebuster and Trufor completely. However, we can observe that their results degrade the most when *Diff-CF* and *Diff-CFG* are applied [2].

5.2. Image Quality Assessment

Another important point to evaluate the pertinence of counter-forensic methods is their resulting image quality. We evaluate this quality in two senses. Firstly, we are interested in how natural the purified images are. To evaluate this, we use the reference-free image quality assessment techniques NIQE [42] and BRISQE [41]. Secondly, it is also important to measure the similarity between the input image and the one obtained after the counter-forensic attack. We, of course, want these two images to be perceptually similar. To evaluate this aspect, we use the full reference image quality assessment methods LPIPS [62],

[2] An analysis of the robustness of these forensic methods is out of the scope of this work and will be addressed in the future.

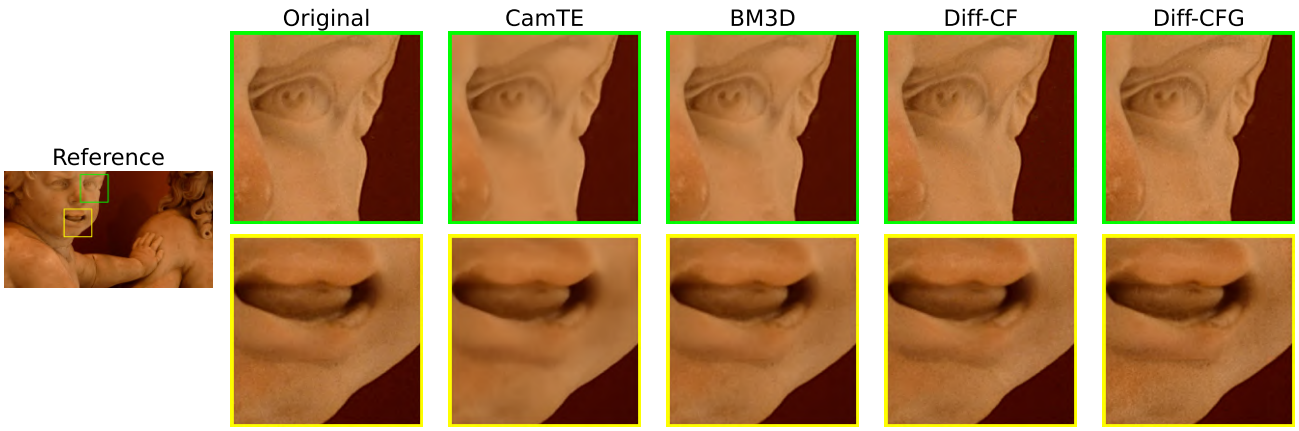

Figure 3. Image quality comparison for all considered counter-forensics methods. We observe that both *Diff-CF* and *Diff-CFG* are good at preserving the fine textures and edges of the image while CamTE and BM3D blur all these fine structures.

SSIM [53], and PSNR. For all the metrics, we use the implementations provided by the PyIQA library [8].

Results are presented in Tab. 2. For all reference-free metrics, the proposed diffusion-based counter-forensics methods achieve the best performance. For the full-reference metrics, we also obtained the best performance for LPIPS, but BM3D and CamTE get better performance in terms of PSNR and SSIM.

Among the proposed methods, the guided variant always achieves better performance in terms of PSNR and SSIM, as expected. Indeed, the guidance explicitly encourages the purified image to be close to the input image. Still, the results are not so conclusive when evaluating the LPIPS score, where the non-guided version shows a slightly better performance on the Korus dataset.

It is important to mention that even if *Diff-CFG* uses SSIM as the guidance distance, this does not imply that the obtained scores on that metric should be perfect. In Eq. 6, the guidance can be interpreted as a sort of gradient descent towards the minimum of $\mathcal{D}(\cdot, x_{in})$. To achieve this minimum, the guidance scale s_t plays a crucial role. Using a non-optimum (in terms of the optimization problem), guidance scale causes the final SSIM score not to be optimal. But this "optimum" guidance scale could not be the best to effectively erase the forgery traces. Sec. 5.3 studied this trade-off more in-depth.

Regarding the reference-free image quality assessment metrics, *Diff-CF* consistently achieves better results than *Diff-CFG*. This can be explained by the fact that the unconstrained generative process gets closer to the distribution of the images with which it was trained. Therefore, these images look more natural.

Fig. 3 shows a qualitative example of the different purified images. We observe that both *Diff-CF* and *Diff-CFG* are good at preserving the fine textures and edges of the image, while CamTE and BM3D blur all these fine structures. For instance, the details highlighted in the green patch show that the granularity in the cherubs' cheeks is blurred out by BM3D and CamTE, while it is preserved by the diffusion-based models. This is also visible in the cherubs' chin, highlighted in the yellow patch. As for the edges, the sharpness of the nose (green patch) and the lips (yellow patch) are also better preserved by the proposed approaches.

5.3. Influence of the parameters

The goal of this work is to provide a first study on the use of diffusion models as counter-forensics techniques. As such, it is important to evaluate how the results vary along with the parameters. The non-guided approach *Diff-CF* has only one parameter: the time step t^*, while *Diff-CFG* has two: the time step t^* and the guidance scale s. In this experiment, we focus mainly on *Diff-CFG* since we think the interaction of both parameters is way more complex than analyzing a single one. The experiments in this section are carried out on Korus dataset [31, 32]. We evaluate both the forgery traces removal capabilities and the image quality of the purified images. For the first, we compute the performance drop for the best-performing methods over the original dataset: Choi, MantraNet, Noiseprint, Splicebuster, and TruFor. For the second, we use all the image quality assessment metrics presented in Sec. 5.2.

Diffusion time-step. The results of the impact of the time-step t^* are presented on the left-hand side of Fig. 4. The analysis is pretty straightforward: the larger the value of t^*, the forgery traces removal performance improves (gets lower). On the other hand, the image quality metrics improve the smaller the value of t^*. There is a clear trade-off in the selection of this parameter, that is simple to understand: with higher values of t^*, we add more noise to the original image in the forward diffusion process, which makes it easier to hide the forgery traces. On the other hand,

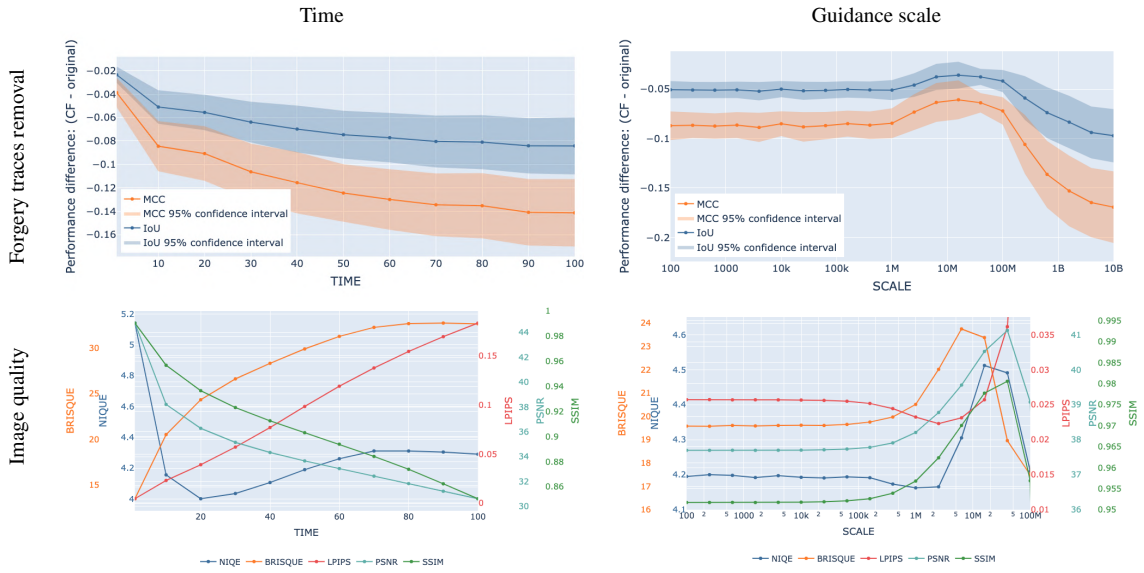

Figure 4. Study of the impact of the time-step t^* (left-hand side), and guidance scale s (right-hand side). For each parameter, we evaluate its influence on the forgery traces removal task (top) and on the purified image quality (bottom). For the forgery traces removal task, we plot the average difference between the performance before and after purification for the best-performing methods in the original dataset as a function of the parameters' value. The colored background area represents the 95% confidence interval. For the Image quality assessment, all five metrics presented in Sec. 5.2 are plotted as a function of the parameters' value, each with a different axis, for better visualization. This figure is best viewed in color. An interactive version of these plots is included in the supplementary material.

starting the reverse process too far away from the original image leads to larger deviations between the original image and the purified one.

In addition, it is interesting to note that all the full-reference metrics keep strongly degrading as we increase t^*, but the reference-free metrics seem to follow a more asymptotic behavior. This evidence can be explained due to the fact that, even if the generated images are more apart from the original one, the diffusion process, following the learned distribution, is still generating natural images.

Guidance scale The guidance scale ensures that the purified image remains close to the manipulated image, thus not modifying its semantic content. However, it is crucial that the chosen guidance scale is not excessively large since it would cause the purified image to match the adulterated image, potentially retaining the manipulation traces [56].

We conducted a series of experiments to study the scale influence, varying the scale value (s in Eq. 7), while keeping a fixed time-step $t^* = 10$. As can be seen in the right-hand side of Fig. 4, the performance difference has small variations for about the first half of the scale range studied, then shows a slight increase, and finally, a great drop. The best point we could choose would be with the lowest value, so at first, one could be tempted to use the highest value for the scale. But if we add the image quality assessment to the analysis, we observe that for those scale values, the quality of the images is highly degraded. Therefore, an intermediate point should be chosen. Note that the optimal point for the removal of forgery traces is not the optimal point in terms of image quality. As mentioned in Sec. 5.2, this could explain why, in our experiments, we do not obtain the best performance in terms of SSIM, even though we are guiding the diffusion process with this metric.

6. Conclusions and Future Work

In this article, we presented a first study on the use of diffusion models for counter-forensics tasks. We showed that such an approach can deliver better results than the existing techniques for both forgery trace removal and image quality. Of course, there is a risk that the shown approaches would be used by people wanting to create forgeries and make them look authentic. The simplicity of this method increases this risk. However, it is also because of its simplicity that the method should be made public: It is important to expose the shortcomings of current methods so that one can know how much trust can be put into an image and so that alternative ways of authentication are developed.

In this direction, future work includes analyzing the traces left by the diffusion purification process [14] to check whether the use of such a counter-forensic approach can be detected or not. Also, it would be interesting to analyze the robustness of the different methods to such kind of counter-forensic methods.

References

[1] Lars Lien Ankile, Anna Midgley, and Sebastian Weisshaar. Denoising diffusion probabilistic models as a defense against adversarial attacks. *ArXiv*, abs/2301.06871:null, 2023. 2

[2] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4, 5

[3] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Image Forgeries Detection through Mosaic Analysis: the Intermediate Values Algorithm. *Image Processing On Line*, 11:317–343, 2021. 5

[4] Daniele Baracchi, Dasara Shullani, Massimo Iuliani, Damiano Giani, and Alessandro Piva. Camera obscura: Exploiting in-camera processing for image counter forensics. *Forensic Science International: Digital Investigation*, 38:301213, 2021. 2

[5] Mauro Barni, Matthew C Stamm, and Benedetta Tondi. Adversarial multimedia forensics: Overview and challenges ahead. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 962–966. IEEE, 2018. 2

[6] Tsachi Blau, Roy Ganz, Bahjat Kawar, Alex Bronstein, and Michael Elad. Threat model-agnostic adversarial defense using diffusion models, 2022. 2

[7] Rainer Böhme and Matthias Kirchner. Counter-forensics: Attacking image forensics. In *Digital image forensics*, pages 327–366. Springer, 2013. 1

[8] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 7

[9] Chang Chen, Zhiwei Xiong, Xiaoming Liu, and Feng Wu. Camera trace erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4, 5

[10] Zhipeng Chen, Benedetta Tondi, Xiaolong Li, Rongrong Ni, Yao Zhao, and Mauro Barni. A gradient-based pixel-domain attack against svm detection of global image manipulations. In *2017 IEEE workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2017. 2

[11] Chang-Hee Choi, Jung-Ho Choi, and Heung-Kyu Lee. Cfa pattern identification of digital cameras using intermediate value counting. In *Proceedings of the Thirteenth ACM Multimedia Workshop on Multimedia and Security*, MM&Sec '11, page 21–26, New York, NY, USA, 2011. Association for Computing Machinery. 5

[12] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security*, 7(6):1841–1854, 2012. 1, 4, 5, 6

[13] Pedro Comesana and Fernando Perez-Gonzalez. The optimal attack to histogram-based forensic detectors is simple (x). In *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 137–142. IEEE, 2014. 2

[14] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022. 8

[15] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2015. 5

[16] Davide Cozzolino, Justus Thies, Andreas Rossler, Matthias Niessner, and Luisa Verdoliva. Spoc: Spoofing camera fingerprints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 990–1000, June 2021. 2

[17] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. 5

[18] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 4

[19] Tiago José de Carvalho, Christian Riess, Elli Angelopoulou, Hélio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013. 5

[20] Thanh-Toan Do, Ewa Kijak, Teddy Furon, and Laurent Amsaleg. Deluding image recognition in sift-based cbir systems. In *Proceedings of the 2nd ACM Workshop on Multimedia in forensics, Security and Intelligence*, pages 7–12, 2010. 2

[21] Wei Fan, Kai Wang, François Cayre, and Zhang Xiong. Jpeg anti-forensics using non-parametric dct quantization noise estimation and natural image statistics. In *Proceedings of the first ACM workshop on Information hiding and multimedia security*, pages 117–122, 2013. 2

[22] Hany Farid. *Photo Forensics*. The MIT Press, 2016. 1

[23] Marco Fontani and Mauro Barni. Hiding traces of median filtering in digital images. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1239–1243. IEEE, 2012. 2

[24] Diego Gragnaniello, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Analysis of adversarial attacks against cnn-based image forgery detectors. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 967–971. IEEE, 2018. 2

[25] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615, June 2023. 5, 6

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1, 3

[27] Dongkyu Kim, Han-Ul Jang, Seung-Min Mun, Sunghee Choi, and Heung-Kyu Lee. Median filtered image restoration

and anti-forensics using adversarial networks. *IEEE Signal Processing Letters*, 25(2):278–282, 2017. 2

[28] Matthias Kirchner and Rainer Bohme. Hiding traces of resampling in digital images. *IEEE Transactions on Information Forensics and Security*, 3(4):582–592, 2008. 2

[29] Matthias Kirchner and Jessica Fridrich. On detection of median filtering in digital images. In Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp III, editors, *Media Forensics and Security II*, volume 7541, page 754110. International Society for Optics and Photonics, SPIE, 2010. 2

[30] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. 2

[31] P. Korus and J. Huang. Evaluation of random field models in multi-modal unsupervised tampering localization. In *Proc. of IEEE Int. Workshop on Inf. Forensics and Security*, 2016. 4, 5, 7

[32] P. Korus and J. Huang. Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Trans. on Information Forensics & Security*, 2017. 4, 5, 7

[33] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, Aug. 2022. 5

[34] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 375–384, 2021. 5

[35] Marc Lebrun. An Analysis and Implementation of the BM3D Image Denoising Method. *Image Processing On Line*, 2:175–213, 2012. https://doi.org/10.5201/ipol.2012.l-bm3d. 4

[36] M. Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. *ArXiv*, abs/2303.09051:null, 2023. 2

[37] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2837–2845, June 2021. 2

[38] Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. Comprint: Image forgery detection and localization using compression fingerprints. *arXiv preprint arXiv:2210.02227*, 2022. 5

[39] Francesco Marra, Giovanni Poggi, Fabio Roli, Carlo Sansone, and Luisa Verdoliva. Counter-forensics in machine learning based forgery detection. In *Media Watermarking, Security, and Forensics 2015*, volume 9409, page 94090L. International Society for Optics and Photonics, 2015. 2

[40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[41] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 6

[42] Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013. 6

[43] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022. 2, 3

[44] Tina Nikoukhah, Jérémy Anger, Thibaud Ehret, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. Jpeg grid detection based on the number of dct zeros and its application to automatic and localized forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 110–118, 2019. 1, 2, 5

[45] M Ali Qureshi and M Deriche. A review on copy move image forgery detection techniques. In *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, pages 1–5. IEEE, 2014. 1

[46] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 1

[47] Hyun Jun Shin, Jong Ju Jeon, and Il Kyu Eom. Color filter array pattern identification using variance of color difference image. *Journal of Electronic Imaging*, 26(4):043015, 2017. 5

[48] Kulbir Singh, Ankush Kansal, and Gurinder Singh. An improved median filtering anti-forensics with better image quality and forensic undetectability. *Multidimensional Systems and Signal Processing*, 30(4):1951–1974, 2019. 2

[49] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 3

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 3

[51] Jiachen Sun, Weili Nie, Zhiding Yu, Z. Mao, and Chaowei Xiao. Pointdp: Diffusion-driven purification against adversarial attacks on 3d point cloud recognition. *ArXiv*, abs/2208.09801:null, 2022. 2

[52] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *ArXiv*, abs/2205.14969:null, 2022. 2, 3

[53] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4, 7

[54] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage – a novel database for copy-move forgery detection. In *IEEE International Conference on Image processing (ICIP)*, pages 161–165, 2016. 4, 5, 6

[55] Jianyuan Wu, Zheng Wang, Hui Zeng, and Xiangui Kang. Multiple-operation image anti-forensics with wgan-gp framework. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1303–1307. IEEE, 2019. 2

[56] Quanlin Wu, Hang Ye, and Yuntian Gu. Guided diffusion model for adversarial purification from random noise. *ArXiv*, abs/2206.10875:null, 2022. 2, 3, 8

[57] Shutong Wu, Jiong Wang, Wei Ping, Weili Nie, and Chaowei Xiao. Defending against adversarial audio via diffusion model. *ArXiv*, abs/2303.01507:null, 2023. 2

[58] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[59] Zhung-Han Wu, Matthew C Stamm, and KJ Ray Liu. Anti-forensics of median filtering. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3043–3047. IEEE, 2013. 2

[60] Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models towards adversarial robustness. *arXiv preprint arXiv:2211.00322*, 2022. 2

[61] Hai-Dong Yuan. Blind forensics of median filtering in digital images. *IEEE Transactions on Information Forensics and Security*, 6:1335–1345, 12 2011. 2

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6