

Recovering Forensic Traces Through JPEG Artifact Removal for Synthetic Image Detection

Juan Carlos Pérez-Novoa

jcperez@gts.uvigo.es
atlanTTic, Universidade de Vigo
Vigo, Spain

David Vázquez-Padín

dvazquez@gts.uvigo.es
atlanTTic, Universidade de Vigo
Vigo, Spain

Fernando Pérez-González

fperez@gts.uvigo.es
atlanTTic, Universidade de Vigo
Vigo, Spain

Abstract

Synthetic image detection performance degrades significantly when images are subjected to post-processing operations such as JPEG compression. We investigate whether JPEG artifact removal can mitigate this degradation without requiring detector retraining. We evaluate distinct JPEG artifact removal methods across three fundamentally different detector families: fingerprint-based methods, machine-learning-based residual detectors, and semantic-based detectors. Experiments on images generated by multiple state-of-the-art diffusion models and on real camera images show that JPEG artifact removal consistently improves the performance of fingerprint-based and semantic-based detectors, while its impact on residual-based detectors depends on the restoration method employed. Overall, the results demonstrate that JPEG artifact removal can recover useful forensic information and represents a promising detector-agnostic pre-processing strategy for improving synthetic image detection in compressed images.

CCS Concepts

• Computing methodologies → Image and video acquisition.

Keywords

Synthetic Image Detection, JPEG Artifact Removal, Image Forensics, Diffusion Models, Fingerprint-Based Detection, Machine-Learning Residual Analysis, CLIP-Based Detection

ACM Reference Format:

Juan Carlos Pérez-Novoa, David Vázquez-Padín, and Fernando Pérez-González. 2026. Recovering Forensic Traces Through JPEG Artifact Removal for Synthetic Image Detection. In *2nd Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media (DFF '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3840473.3840515>

1 Introduction

In recent years, synthetic image generators have become widely accessible to the general public and increasingly easy to use, requiring only a text prompt to produce photorealistic results. While these technologies enable a wide range of legitimate creative applications, they have also been exploited by malicious actors [6],

contributing to misinformation campaigns, fraud, and challenges to the authenticity of visual media [3].

To counter these threats, numerous forensic approaches for synthetic image detection have been proposed in recent years [4, 9, 10, 17, 24, 26]. These methods exploit a variety of forensic cues, including generator fingerprints, image residuals, and high-level semantic representations extracted from vision-language models such as CLIP (Contrastive Language–Image Pre-training) [29]. Despite their methodological differences, all of these approaches can be affected by post-processing operations applied after image generation.

Among the various post-processing operations that images may undergo after generation, JPEG compression is particularly relevant due to its widespread use in image storage and online sharing. Previous studies have shown that JPEG compression modifies forensic traces [8] and degrades the performance of synthetic image detectors [16]. This motivates the investigation of whether reducing JPEG-induced artifacts can restore forensic traces and enhance the performance of synthetic image detectors.

In this work, we study the effect of JPEG artifact removal on synthetic image detection. Rather than proposing a new detector, our objective is to assess whether JPEG artifact removal can improve the effectiveness of existing detectors. To this end, we generate JPEG artifact-suppressed images using three restoration models: a DnCNN-based model [33], retrained in this work for JPEG artifact removal, and two dedicated JPEG restoration methods, namely FBCNN [19] and SwinIR [22]. To assess the generality of this strategy, we evaluate its impact on representative detectors spanning the three detector families associated with the forensic cues discussed above: fingerprint-based methods, Machine Learning (ML)-based residual detectors, and semantic-based approaches based on CLIP representations.

The main contribution of this work is a comprehensive empirical study of how JPEG artifact removal affects synthetic image detection across detector families that rely on fundamentally different forensic cues. Our results demonstrate that suppressing JPEG compression artifacts can recover useful information for detection and provide a potential pathway to mitigate the degradation caused by JPEG compression. Although our study focuses on JPEG compression, the same principle could be extended to other post-processing operations that degrade forensic traces.

The remainder of this paper is organized as follows. Sect. 2 reviews prior work on synthetic image detection, with particular emphasis on the impact of JPEG compression on detection performance. Sect. 3 introduces the JPEG artifact-removal methods considered in our study. The experimental setup is described in



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

DFF '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2927-0/2026/11

<https://doi.org/10.1145/3840473.3840515>

Sect. 4, including the dataset, compression scenarios, and the representative detectors selected for evaluation. The impact of JPEG artifact removal on the performance of these detectors is then analyzed and discussed in Sect. 5. Finally, Sect. 6 summarizes the main findings and outlines directions for future research.

2 Related Works

The forensic analysis of synthetic images emerged shortly after the introduction of modern AI-based image generators. Early studies revealed that image synthesis processes leave characteristic non-visible traces that can be exploited for forensic purposes. In particular, source-specific artifacts manifesting as periodic peaks and structured patterns in the Fourier spectrum of image residuals were first identified in images generated by Generative Adversarial Networks (GANs) [15, 26] and later observed in images generated by Diffusion Models (DMs) [8]. These traces have enabled a variety of forensic analyses, ranging from the study of generator-specific characteristics in GAN- and DM-generated images [9] to the discrimination between fully synthetic and laundered images [24], as well as synthetic image detection [4].

Building on the observations discussed in Sect. 1, existing synthetic image detectors can be broadly categorized into three detector families according to the information they exploit:

- (1) *fingerprint-based methods*, which rely on source-specific traces left by generative models [26];
- (2) *ML-based residual forensic detector*, which analyze image residuals and low-level artifacts [4, 9, 15, 17, 24]; and
- (3) *semantic-based approaches*, which leverage semantic representations (such as CLIP) extracted from large vision-language models [10].

Despite their differences, all of these methods can be affected by post-processing operations applied after image generation, with JPEG compression being the most common in real-world scenarios.

As a consequence, understanding the impact of JPEG compression on the forensic information exploited by these detectors is essential for assessing their robustness in practice. By modifying both low-level image statistics and frequency-domain characteristics, JPEG compression can partially attenuate or mask the forensic information left by generative models, thereby reducing the effectiveness of synthetic image detectors. Although recent work has questioned the extent to which state-of-the-art deep learning detectors explicitly rely on spectral artifacts [25], the negative impact of JPEG compression on synthetic image detection performance is well documented [4, 9, 10, 15–17].

The first work to analyze this effect was conducted by Grommelt *et al.* [16], who identified a bias in synthetic image detectors caused by the different compression characteristics of synthetic and real images. Since synthetic images are often generated in lossless formats such as PNG, whereas real camera images are commonly stored as JPEG-compressed files, detectors may inadvertently learn a spurious correlation between JPEG compression artifacts and real images. Their results showed that this bias causes compressed synthetic images to be misclassified as real more frequently, particularly at low quality factors. To mitigate this issue, they proposed incorporating JPEG-compressed images during training. Similar strategies based on data augmentation have also been explored in

other works [2, 17, 28]. While effective, these approaches require retraining the detector and are tailored to specific post-processing operations.

An alternative strategy is to mitigate the impact of compression before the detection stage. JPEG artifact removal is a well-established image restoration problem that has been extensively studied over the last decades. Early methods relied on techniques such as DCT-domain filtering [13]. More recently, deep learning approaches have significantly advanced the state of the art. In particular, DnCNN [33] demonstrated the effectiveness of residual-learning CNNs for image denoising, while subsequent methods such as FBCNN [19] and SwinIR [22] further improved JPEG artifact removal through blind quality factor estimation and transformer-based architectures, respectively.

Despite the substantial progress in JPEG artifact removal, its potential benefits for multimedia forensics remain largely unexplored. Existing studies have primarily focused on perceptual restoration quality, typically measured using metrics such as PSNR and SSIM, while the impact of artifact removal on the forensic information exploited by synthetic image detectors has received little attention. We therefore investigate whether JPEG artifact suppression can recover forensic traces degraded by compression and improve detection performance across fingerprint-based, ML-based residual, and semantic-based detector families. Unlike prior approaches that improve robustness through detector retraining or compression-aware data augmentation [4], our methodology operates entirely as a detector-agnostic pre-processing stage.

3 JPEG Artifacts Removal Models

To suppress JPEG-induced artifacts in both natural and synthetic images that have undergone JPEG compression as a post-processing operation, we first considered the DnCNN architecture [33]. Due to its effectiveness in residual extraction, DnCNN has become one of the most widely adopted neural networks in image forensics, forming the basis of Noiseprint [11] and later being adopted in several studies on synthetic image analysis [8, 9, 24]. Since the original model was developed for image denoising rather than JPEG artifact suppression, we retrained it from scratch for the latter task.

The original DnCNN architecture for blind Gaussian denoising consists of 20 convolutional layers with 64 feature channels and 3×3 kernels with padding, using Batch Normalization and ReLU activations throughout the network. The model was trained from scratch using 2,000 natural images from the RAISE dataset [12], using a set of images that does not overlap with the dataset described in Sect. 4.1. Due to memory constraints, a centered crop of size 512×512 was extracted from each image. Training was performed considering a single JPEG compression Quality Factor (QF), namely $QF = 75$. Network weights were initialized orthogonally and optimized using Adam with an initial learning rate of 10^{-3} , reduced by a factor of 10 after epochs 8, 16, 24, and 32. The training process lasted 34 epochs and employed the ℓ_1 -norm as loss function. JPEG compression at $QF = 75$ was applied on-the-fly to image patches of size 64×64 extracted with a stride of 16 pixels. Standard data augmentation techniques, including rotations and flips, were also applied.

QF	JPEG Artifact Removal	PSNR (dB)	SSIM
75	None	36.32	0.9497
	DnCNN	37.32	0.9581
	FBCNN	38.38	0.9641
	SwinIR	37.21	0.9551
60	None	34.76	0.9341
	DnCNN	35.96	0.9468
	FBCNN	37.00	0.9537
	SwinIR	36.83	0.9513
40	None	33.30	0.9144
	DnCNN	34.37	0.9298
	FBCNN	35.69	0.9406
	SwinIR	35.97	0.9428

Table 1: Average PSNR and SSIM with respect to the original uncompressed images for the JPEG-compressed baseline (None) and for JPEG artifact-suppressed images obtained using DnCNN, FBCNN, and SwinIR, evaluated at quality factors $QF \in \{75, 60, 40\}$.

The trained model was evaluated on the test dataset described in Sect. 4.1, comprising 6,000 images compressed at $QF = 75$, the same quality factor used during training. As shown in Tab. 1, the resulting JPEG artifact-suppressed images achieved an average PSNR of 37.32 dB with respect to their original uncompressed counterparts, representing an improvement of approximately 1 dB over the JPEG-compressed baseline (36.32 dB). A corresponding improvement was also observed in terms of SSIM.

To assess the model’s behavior on compression levels not observed during training, the same evaluation was repeated using images compressed at $QF = 60$ and $QF = 40$. The corresponding PSNR and SSIM values are reported in Tab. 1. As expected, the absolute restoration quality decreases as the compression level increases, since lower quality factors introduce more severe distortions. However, the restored images consistently outperform their JPEG-compressed counterparts, yielding PSNR gains of approximately 1 dB and noticeable improvements in SSIM across all evaluated QFs. These results indicate that the model retains its ability to suppress JPEG artifacts beyond the compression level used during training.

In addition to our DnCNN-based model, we evaluated two state-of-the-art approaches specifically designed for JPEG artifact removal. The first is FBCNN [19], a blind restoration model that does not require knowledge of the quality factor and can therefore be applied across a wide range of compression levels. The second is SwinIR [22], for which we employed the model trained for $QF = 40$. Although SwinIR provides separate models for different quality factors, the $QF = 40$ version corresponds to the highest available quality factor and is therefore the most suitable for our experiments, as higher quality factors are generally more representative of real-world JPEG compression scenarios.

The restoration performance of these methods is also included in Tab. 1. Both FBCNN and SwinIR consistently outperform the JPEG-compressed baseline and achieve higher PSNR and SSIM values

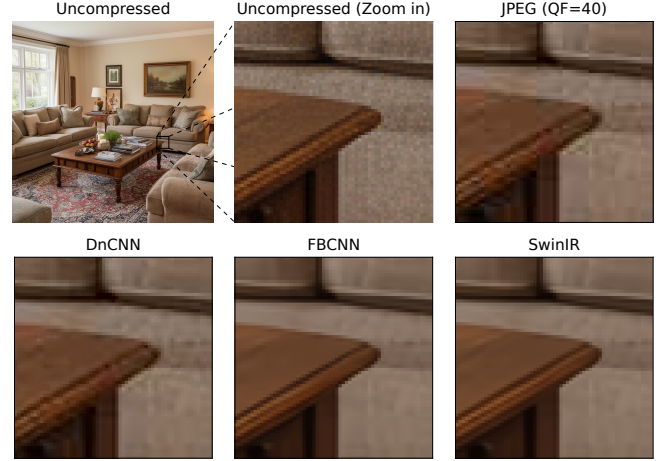


Figure 1: Example of JPEG artifact suppression on a synthetic image generated by FLUX.2 (top-left panel). The figure presents the original uncompressed image and a magnified image region, along with the corresponding region after JPEG compression at $QF = 40$ (top row) and after restoration using DnCNN, FBCNN, and SwinIR (bottom row). The restored images exhibit reduced compression artifacts and a partial recovery of fine image structures relative to the JPEG-compressed version.

than the retrained DnCNN model across nearly all evaluated quality factors, with the sole exception of SwinIR at $QF = 75$. FBCNN provides the best restoration quality at $QF = 75$ and $QF = 60$, whereas SwinIR achieves the highest performance at $QF = 40$. These results highlight the advantages of models specifically designed for JPEG artifact suppression.

A qualitative example is shown in Fig. 1, which illustrates the effect of JPEG compression and subsequent restoration on a synthetic image generated by FLUX.2. The figure presents a magnified image region from the original uncompressed image, its JPEG-compressed counterpart at $QF = 40$, and the corresponding outputs produced by DnCNN, FBCNN, and SwinIR. While JPEG compression introduces visible blocking artifacts and a loss of fine details, all three methods are able to reduce these distortions and partially recover the original image structure, with the improvements being particularly noticeable for FBCNN and SwinIR.

4 Experimental Setup

Having described the JPEG artifact removal models, we now introduce the experimental setup used to assess their impact on synthetic image detection. We first present the dataset and compression scenarios considered in our study, followed by three representative detectors drawn from the fingerprint-based, ML-based residual, and semantic-based detector families.

4.1 Dataset

The dataset used in this study comprises 9,000 images equally distributed across six sources: DALL-E 3 [27], FLUX.1 [20], FLUX.2 [21], two Stable Diffusion (SD) models (SD1.4 [30] and SD3.5 [1]), and

real images. For each of the six sources, the 1,500 available images are split into two non-overlapping subsets. Specifically, 1,000 images per source are allocated to the test set (resulting in a total of 6,000 test images across all sources), while the remaining 500 images per source are reserved for reference-fingerprint estimation (yielding a total of 3,000 images for fingerprint extraction). This partitioning ensures that the evaluation is performed on data that were not used during fingerprint extraction.

The real-image subset consists of 1,500 uncompressed images from the RAISE dataset [12]. We selected this dataset because it contains camera-native images that have not undergone post-processing, thereby providing a clean reference set of real images. The original images were center-cropped to a resolution of 1024×1024 pixels.

The DALL-E 3 subset comprises 1,500 images with a resolution of 1024×1024 , randomly selected from the more than 8,000 evaluation samples released by Betker *et al.* [5], which were generated from captions in the MS-COCO 2014 validation set. Similarly, the FLUX.1, FLUX.2, SD1.4, and SD3.5 subsets each contain 1,500 images generated by the authors and randomly sampled from larger image collections generated from the same caption source.¹ FLUX.1, FLUX.2, and SD3.5 images have a native resolution of 1024×1024 , whereas SD1.4 images have a native resolution of 512×512 .

All synthetic images were originally stored in PNG format, while the real images were stored in TIFF format. Except for the center-cropped RAISE images, all images were used at their native resolution. To evaluate the impact of compression, each image was subsequently JPEG-compressed using Pillow 10.0 at three quality factors, namely $QF \in \{75, 60, 40\}$.

4.2 Synthetic Image Detectors

We evaluate one representative detector from each of the three detector families introduced in Sect. 2:

- (1) a fingerprint-based detector implemented by us using the DnCNN denoiser adopted by Corvi *et al.* [9];
- (2) Corvi2023 [9], an ML-based residual detector; and
- (3) CLIPDet [10], a semantic-based detector.

For Corvi2023 and CLIPDet, we used the official implementations released by their respective authors.

4.2.1 Fingerprint-based Method. Following the principles of Photo-Response Non-Uniformity (PRNU) extraction for camera source identification [23], later extended to GAN-generated [26] and DM-generated images [8], each image source (whether a generative model or a real camera) is represented by a reference fingerprint estimated from a dedicated set of images. Throughout this process, noise residuals are extracted using the DnCNN denoiser configuration adopted by Corvi *et al.* [9] (17 layers, 64 feature maps, $\sigma = 1$, kernel size 3×3 , and no padding) on images normalized to the $[0, 1]$ range. The subsequent fingerprint extraction and matching procedure follows the standard three-step framework described in [14, Sect. 2.1]:

- (i) *Reference Fingerprint Estimation:* A reference fingerprint is estimated from a set of L images belonging to the same source. For each image, the noise residual is extracted via the

aforementioned DnCNN denoiser and resized to 256×256 pixels. An orthonormal FFT is then applied to each residual, the resulting spectra are averaged, and the final fingerprint is obtained through an inverse FFT.

- (ii) *Query Residual Extraction:* Given a query image, its noise residual is extracted using the exact same normalized DnCNN pipeline and resized to 256×256 pixels.
- (iii) *Similarity Matching and Attribution:* The similarity between the query residual and each reference fingerprint is measured using Peak-to-Correlation Energy (PCE) [14]. The query image is then attributed to the source whose PCE value exceeds a predefined threshold.

Given the six image sources considered in our study, a reference fingerprint is estimated for each source from the corresponding set of $L = 500$ reference images (see Sect. 4.1). Fingerprints are computed independently for every image condition analyzed in this work, namely the uncompressed images, the JPEG-compressed images with $QF \in \{75, 60, 40\}$, and the outputs produced by each of the three JPEG artifact removal methods described in Sect. 3.

As an illustrative example, Fig. 2 presents the fingerprints obtained for two representative sources: FLUX.1 and real images. The FLUX.1 fingerprint exhibits the characteristic periodic peaks and structured patterns in the Fourier spectrum that have previously been reported for other generative models [8]. For each source, the figure shows the fingerprints extracted from the uncompressed images, the corresponding JPEG-compressed images at $QF = 75$, and the fingerprints obtained after applying each of the three JPEG artifact removal methods. For brevity, the results for $QF = 60$ and $QF = 40$ are omitted, as they exhibit similar behavior.

As can be observed, JPEG compression significantly alters the reference fingerprints by introducing a characteristic periodic pattern associated with the 8×8 block structure of JPEG. In the frequency domain, this pattern manifests itself as peaks at multiples of $\pm \frac{2\pi}{8}$ radians along both spatial directions, partially masking the source-specific traces and making the fingerprints considerably more similar to one another. After JPEG artifact removal, these compression-induced peaks are noticeably attenuated and the energy concentration along the horizontal and vertical axes is reduced. Consequently, the restored fingerprints more closely resemble those obtained from the original uncompressed images, suggesting that part of the forensic information degraded by JPEG compression can be recovered.

4.2.2 ML-based Residual Method. The Corvi2023 detector considered in this category is the result of a series of developments built upon a residual-learning architecture. It originates from the ResNet50-based detector proposed by Wang *et al.* [32] for the detection of GAN-generated images. Subsequently, Gragnaniello *et al.* [15] modified the architecture by removing the downsampling operation in the first layer and employing more extensive data augmentation during training. Later, Corvi *et al.* [9] adapted the model to the detection of diffusion-generated images by training it on images produced by one of the first DM-based generators (i.e., Latent Diffusion). In our experiments, we use the model weights provided in the repository accompanying the work of Cozzolino *et al.* [10].

¹The dataset is available at <https://github.com/dvazquezpadin/jpeg-artifact-removal>

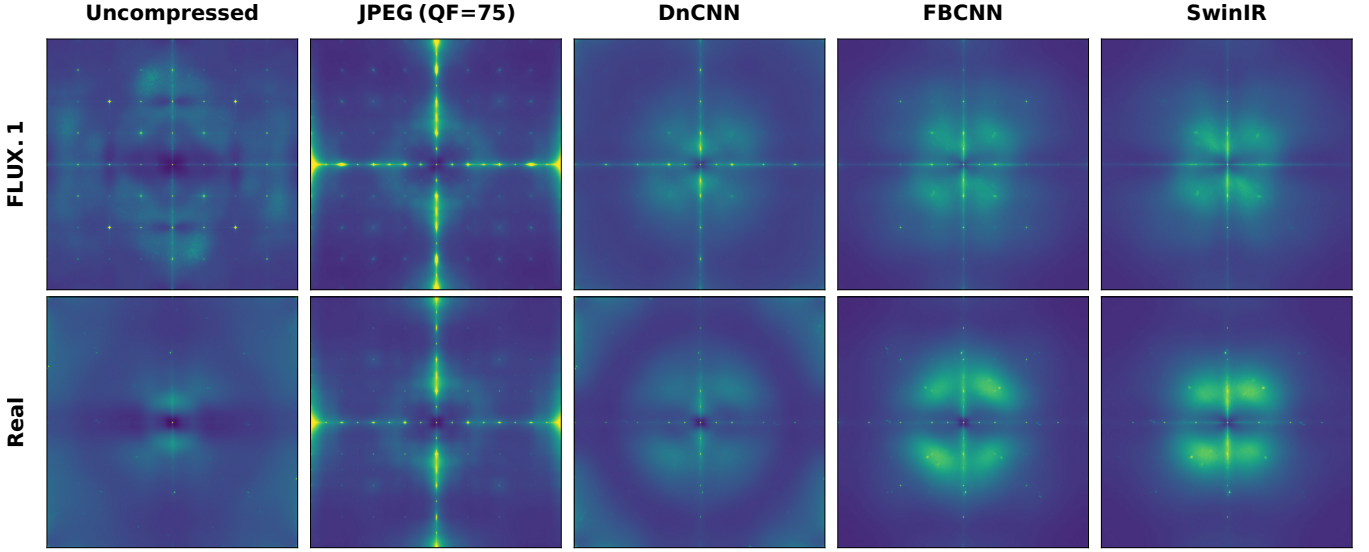


Figure 2: Average power spectral density fingerprints extracted from 500 FLUX.1 images (top) and 500 real images (bottom). The figure compares fingerprints obtained from the original uncompressed images, JPEG-compressed images at QF = 75, and JPEG artifact-suppressed images produced by the restoration models: DnCNN, FBCNN, and SwinIR. JPEG compression introduces a characteristic grid pattern that masks the underlying source-specific traces, whereas artifact removal attenuates this pattern and yields fingerprints closer to those extracted from the uncompressed images.

4.2.3 Semantic-based Method. Recent approaches to synthetic image detection have increasingly explored semantic representations extracted from vision-language models such as CLIP. In particular, CLIPDet [10] has demonstrated promising robustness to several post-processing operations, including JPEG compression, making it a suitable representative of this detector family for our study.

Despite this robustness, recent findings by Uhlenbrock *et al.* [31] showed that post-processing operations can alter the position of images in the latent CLIP space, with JPEG compression inducing increasingly larger shifts as the quality factor decreases. Motivated by these observations, we investigate whether JPEG artifact removal can mitigate these shifts and move compressed images closer to their original uncompressed representations.

To analyze the effect of JPEG artifact removal in the CLIP feature space, we use the OpenCLIP library [18] with the ViT-L-14 model [7] to extract image embeddings. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ denote the CLIP embeddings of two images. Their distance in the latent CLIP space is quantified using the cosine distance:

$$d_{\cos}(\mathbf{e}_1, \mathbf{e}_2) = 1 - \frac{\mathbf{e}_1^T \mathbf{e}_2}{\|\mathbf{e}_1\|_2 \cdot \|\mathbf{e}_2\|_2},$$

where smaller values indicate a higher similarity between the corresponding image representations. We compute this metric for every image in the test set and aggregate the resulting values under two scenarios: (i) between JPEG-compressed images and their corresponding uncompressed counterparts, and (ii) between JPEG artifact-suppressed images, obtained using each of the three considered artifact removal methods, and their corresponding uncompressed counterparts. This analysis enables us to quantify the extent to which JPEG artifact removal reduces the shift caused by JPEG compression in the latent CLIP space.

Fig. 3a reports the average pairwise cosine distance in the CLIP feature space between each processed image and its original uncompressed counterpart for the case of QF = 75. Across all considered image sources, the images obtained after JPEG artifact removal consistently exhibit smaller cosine distances than their corresponding JPEG-compressed versions, regardless of the artifact-removal method employed. This indicates that JPEG artifact removal partially restores the original CLIP representation, bringing the processed images closer to their uncompressed counterparts in the latent feature space. This behavior is expected to translate into improved synthetic image detection performance, as further discussed in Sect. 5.2.

A different trend emerges as the compression quality factor decreases in Figs. 3b and 3c. For $\text{QF} \in \{60, 40\}$, the cosine distance obtained with our trained version of the image denoising DnCNN model increases, which is consistent with the fact that the network was trained exclusively on images compressed at QF = 75. In contrast, both FBCNN and SwinIR, which were specifically designed for JPEG artifact removal and exhibit greater robustness across compression levels, continue to reduce the cosine distance relative to the corresponding JPEG-compressed images. This suggests that these models are better able to recover representations that remain closer to those of the original uncompressed images, even at lower quality factors, an observation that is later reflected in the detection performance reported in Sect. 5.2.

5 Experimental Results

To empirically assess how JPEG artifact removal impacts synthetic image detection, we first establish baseline performance metrics without applying any restoration. The three detectors described in

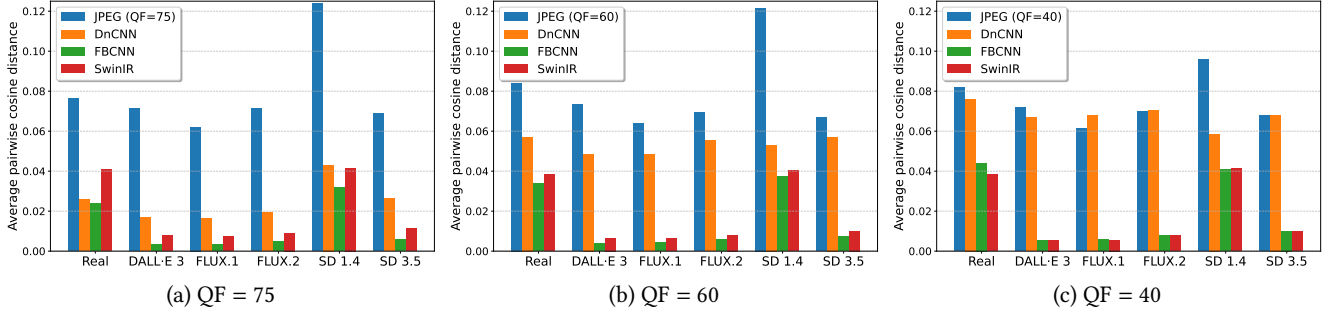


Figure 3: Average pairwise cosine distance in the CLIP feature space between each processed image and its corresponding uncompressed version. Results are reported separately for $QF \in \{75, 60, 40\}$. For each image source, the four bars correspond to the JPEG-compressed images and the JPEG artifact-suppressed images produced by DnCNN, FBCNN, and SwinIR, respectively. Lower cosine distances indicate representations that remain closer to the original uncompressed image in the CLIP feature space.

Sect. 4.2 are evaluated on the test dataset introduced in Sect. 4.1, first using the original uncompressed images (both real and synthetic) and subsequently using versions compressed with JPEG at quality factors $QF \in \{75, 60, 40\}$.

Using these baseline results as a reference, we repeat the same evaluation protocol after applying each of the three JPEG artifact removal methods under study (see Sect. 3). The restored images are then fed into the three synthetic image detectors. Comparing these results against the baseline allows us to quantify the extent to which JPEG artifact suppression restores or influences detection performance across different detector architectures.

5.1 Fingerprint-based Method Performance

Since a distinct fingerprint is estimated for each of the six image sources considered in this study (real images, DALL-E 3, FLUX.1, FLUX.2, SD1.4, and SD3.5), we evaluate the fingerprint-based attribution method under two complementary scenarios:

- *One-vs-All Source Discrimination:* In this setting, we evaluate the detector’s ability to isolate a specific source. Images originating from the source corresponding to the target reference fingerprint serve as the positive class, while images from the remaining five sources constitute the negative class.
- *Real-versus-Synthetic Detection:* In this binary classification setting, we rely solely on the fingerprint extracted from real images. Real images represent the positive class, while all synthetic images (from the five generative models) form the negative class.

In both scenarios, performance is evaluated using the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC).

For these evaluations, reference fingerprints are estimated separately for each source using the $L = 500$ reference images. In addition to extracting fingerprints from the original uncompressed images, we compute dedicated fingerprints for each JPEG artifact removal method using the restored images obtained at each compression level. Consequently, a distinct reference fingerprint is estimated for every unique combination of QF and restoration method. These reference fingerprints are then matched against test images under the various compression and restoration conditions examined in this work. Tab. 2 summarizes the resulting AUC values for

all evaluated test-image conditions: uncompressed images, JPEG-compressed images with $QF \in \{75, 60, 40\}$, and restored images produced by the three artifact removal methods: DnCNN, FBCNN, and SwinIR. For the one-vs-all source discrimination scenario, the reported AUC represents the average performance across the six individual source-specific discrimination tasks.

The results show that JPEG compression severely degrades the effectiveness of fingerprint-based analysis across all quality factors. Using fingerprints extracted from uncompressed images, the average AUC in the source discrimination scenario drops from 0.912 to values between 0.483 and 0.509 (after applying JPEG compression), while in the real-versus-synthetic setting it decreases from 0.887 to values between 0.401 and 0.468 (after JPEG compression). This confirms that JPEG compression substantially alters the generator-specific traces exploited by fingerprint-based detectors.

The removal of JPEG artifacts consistently improves performance with respect to the corresponding JPEG-compressed baseline, regardless of the QF or restoration method considered. The improvement is particularly pronounced in the real-versus-synthetic scenario, where the AUC gains frequently exceed 0.25 and, in some cases, approach 0.40. These results indicate that a significant portion of the forensic information attenuated by JPEG compression can be recovered through artifact suppression.

A consistent observation across our experiments is that the highest performance is generally achieved when the reference fingerprints and test images are processed using the same restoration pipeline. This trend is particularly evident for FBCNN and SwinIR, suggesting that JPEG artifact removal alters the fingerprint domain and that domain alignment between reference and query images is crucial for reliable detection. Among the evaluated methods, FBCNN yields the most robust and consistent performance, achieving AUC values between 0.854 and 0.864 in the real-versus-synthetic scenario, which are remarkably close to the baseline performance obtained with uncompressed images (0.887). In contrast, DnCNN-based restoration exhibits a much stronger dependence on the compression quality factor, likely because the model was trained exclusively on images compressed at $QF = 75$. Furthermore, even under matched domain conditions (where the same restoration pipeline is used for both reference and test images), DnCNN

Scenario	Fingerprint	Test Images					
		Uncompressed	QF	JPEG	DnCNN	FBCNN	SwinIR
Source Discrimination (Avg.)	Uncompressed	0.912	75	0.483	0.575	0.543	0.540
			60	0.501	0.573	0.530	0.541
			40	0.509	0.577	0.530	0.538
	DnCNN	–	75	0.502	0.638	–	–
			60	0.490	0.618	–	–
			40	0.477	0.578	–	–
	FBCNN	–	75	0.522	–	0.615	–
			60	0.516	–	0.600	–
			40	0.505	–	0.582	–
	SwinIR	–	75	0.524	–	–	0.617
			60	0.505	–	–	0.591
			40	0.495	–	–	0.554
Real versus Synthetic	Uncompressed	0.887	75	0.468	0.730	0.774	0.744
			60	0.401	0.686	0.749	0.769
			40	0.415	0.642	0.740	0.724
	DnCNN	–	75	0.592	0.582	–	–
			60	0.613	0.422	–	–
			40	0.654	0.328	–	–
	FBCNN	–	75	0.484	–	0.858	–
			60	0.484	–	0.864	–
			40	0.540	–	0.854	–
	SwinIR	–	75	0.638	–	–	0.813
			60	0.647	–	–	0.808
			40	0.642	–	–	0.746

Table 2: AUC values obtained with the fingerprint-based detector for different reference fingerprints and test image conditions, including uncompressed images, JPEG-compressed images ($QF \in \{75, 60, 40\}$), and JPEG artifact-suppressed images. Source Discrimination (Avg.) reports the average AUC across all one-vs-all source discrimination experiments, while Real vs. Synthetic corresponds to the binary detection task using the fingerprint extracted from real images. Bold values indicate improvements over the corresponding JPEG-compressed baseline at the same quality factor.

fails to improve real-versus-synthetic detection performance. In fact, at lower quality factors, it substantially underperforms the corresponding raw JPEG-compressed baseline. The underlying causes of this behavior are not yet fully understood and need further investigation.

Overall, these findings demonstrate that JPEG artifact removal can substantially mitigate the adverse effects of compression on fingerprint-based synthetic image detection. Although the original fingerprint information is not fully restored, the recovered traces are sufficient to yield significant improvements in both source discrimination and real-versus-synthetic detection tasks.

5.2 ML-based Residual and Semantic-based Methods Performance

The representative detectors chosen from the ML-based residual and semantic-based families (specifically Corvi2023 and CLIPDet, respectively), are both implemented as end-to-end classifiers. For this reason, their results are presented and discussed jointly in this section. Following the same evaluation protocol used for the fingerprint-based method, we evaluate these detectors under all experimental conditions considered in this work: uncompressed

images, JPEG-compressed images with $QF \in \{75, 60, 40\}$, and JPEG-restored images obtained via the three artifact removal methods. Detection performance is quantified using both AUC and classification Accuracy (ACC).

The AUC and ACC values obtained for Corvi2023 and CLIPDet are collected in Tab. 3, showing that the removal of JPEG artifacts has a markedly different impact depending on the detector family. For CLIPDet, all three restoration methods consistently improve both AUC and ACC across all compression levels, with DnCNN yielding the largest gains and nearly recovering the results achieved on uncompressed images at $QF = 75$. These results are consistent with the CLIP-space analysis presented previously, suggesting that artifact removal reduces the shift induced by JPEG compression in the latent semantic space. In contrast, the residual-based detector Corvi2023 exhibits a stronger dependence on the restoration method. While FBCNN and SwinIR generally maintain or slightly improve performance relative to the JPEG-compressed baseline, DnCNN leads to a substantial degradation at lower quality factors, likely because it was trained exclusively on images compressed at $QF = 75$. Overall, the results indicate that JPEG artifact removal can

		CLIPDet		Corvi2023	
JPEG Removal	QF	AUC	ACC	AUC	ACC
None	75	0.749	0.605	0.834	0.707
DnCNN		0.805	0.693	0.831	0.605
FBCNN		0.772	0.662	0.869	0.665
SwinIR		0.764	0.665	0.840	0.747
None	60	0.764	0.610	0.825	0.698
DnCNN		0.767	0.678	0.758	0.588
FBCNN		0.768	0.664	0.852	0.696
SwinIR		0.770	0.666	0.832	0.746
None	40	0.764	0.610	0.810	0.683
DnCNN		0.780	0.667	0.664	0.568
FBCNN		0.756	0.663	0.823	0.729
SwinIR		0.765	0.662	0.813	0.718
Uncompressed Case		0.830	0.658	0.906	0.805

Table 3: AUC and ACC results obtained for Corvi2023 (ML-based residual detector) and CLIPDet (semantic-based detector) on uncompressed images, JPEG-compressed images, and JPEG artifact-suppressed images generated using the three restoration methods. Bold values indicate an improvement over the corresponding JPEG-compressed baseline at the same quality factor.

effectively mitigate compression-induced degradation, although its benefits depend on the forensic cues exploited by the detector.

6 Conclusions

In this paper, we investigated the impact of JPEG artifact removal on synthetic image detection. Rather than proposing a new detector, our objective was to assess whether suppressing JPEG compression artifacts can recover forensic information and improve the performance of existing detectors. To this end, we evaluated three JPEG artifact removal methods (DnCNN, FBCNN, and SwinIR) across three fundamentally different detector families: fingerprint-based, ML-based residual, and semantic-based detectors.

The results show that JPEG compression degrades the performance of all considered detector families, although to different extents. Applying JPEG artifact removal consistently improved fingerprint-based detection and source discrimination, indicating that part of the generator-specific traces masked by compression can be recovered. For the ML-based residual detector, the impact depend on the restoration method, with FBCNN and SwinIR generally improving performance while DnCNN show limited generalization to compression levels not observed during training. In contrast, the CLIP-based detector consistently benefits from JPEG artifact removal across all evaluated quality factors, a finding that was further supported by the observed reduction of the distance between restored and original images in the latent CLIP space.

Overall, our findings suggest that JPEG artifact removal can serve as an effective pre-processing step for mitigating the negative impact of compression on synthetic image detection without requiring detector retraining. Preliminary experiments using the

fusion strategy proposed by Cozzolino *et al.* [10] further indicate that combining detectors based on complementary forensic cues (e.g., Corvi2023 and CLIPDet) may yield additional performance gains after restoration. These observations motivate future work on detector fusion and on extending restoration-based approaches to other post-processing operations, such as image resizing, that similarly interfere with forensic traces.

Finally, our results raise an interesting question regarding the forensic information that is actually recovered by JPEG artifact removal. In particular, it would be valuable to reconcile our findings with those of Mandelli *et al.* [25], who argued that most state-of-the-art deep learning detectors do not seem to rely heavily on the characteristic spectral peaks associated with synthetic image generation. While the improvements observed in this work suggest that artifact removal restores information useful for detection, this does not necessarily imply that the recovered spectral peaks are the primary factor responsible for the observed gains. A plausible explanation is that modern detectors, often trained to be robust against JPEG compression, exploit additional forensic cues that are also restored by the artifact-removal process. Identifying which traces contribute most to the observed improvements remains an interesting direction for future research.

Acknowledgments

This work is supported in part by Xunta de Galicia and the European Regional Development Fund under Project ED431C 2025/41.

References

- [1] Stability AI. 2024. Stable Diffusion 3.5-Large. <https://huggingface.co/stabilityai/stable-diffusion-3.5-large>.
- [2] Efthymia Amarantidou, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2025. Composite Data Augmentations for Synthetic Image Detection Against Real-World Perturbations. In *2025 33rd European Signal Processing Conference (EUSIPCO)*. 1213–1217. doi:10.23919/EUSIPCO63237.2025.11226549
- [3] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, Sara Mandelli, Taiba Majid, Gian Luca Marcialis, Marco Micheletto, Andrea Montibeller, Giulia Orrù, Alessandro Ortis, Pericle Perazzo, Giovanni Puglisi, Nischay Purnekar, Davide Salvi, Stefano Tubaro, Massimo Villari, and Domenico Vitulano. 2025. Deepfake Media Forensics: Status and Future Challenges. *Journal of Imaging* 11, 3 (2025). doi:10.3390/jimaging11030073
- [4] Quentin Bamme. 2024. Synthbuster: Towards Detection of Diffusion Model Generated Images. *IEEE Open Journal of Signal Processing* 5 (2024), 1–9. doi:10.1109/OJSP.2023.3337714
- [5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. *Improving Image Generation with Better Captions*. Technical Report. OpenAI. <https://cdn.openai.com/papers/dall-e-3.pdf> Evaluation samples available at <https://github.com/openai/dalle3-eval-samples>.
- [6] Zhiyi Chen, Jinyi Ye, Beverly Tsai, Emilio Ferrara, and Luca Luceri. 2025. Synthetic Politics: Prevalence, Spreaders, and Emotional Reception of AI-Generated Political Images on X. In *Proceedings of the 36th ACM Conference on Hypertext and Social Media (HT '25)*. Association for Computing Machinery, New York, NY, USA, 11–21. doi:10.1145/3720553.3746675
- [7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2818–2829.
- [8] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 973–982.
- [9] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On The Detection of Synthetic Images Generated by Diffusion Models. In *ICASSP 2023 - 2023 IEEE International Conference on*

- Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10095167
- [10] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. Raising the Bar of AI-generated Image Detection with CLIP. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 4356–4366. doi:10.1109/CVPRW63382.2024.00439
 - [11] Davide Cozzolino and Luisa Verdoliva. 2020. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2020), 144–159. doi:10.1109/TIFS.2019.2916364
 - [12] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. 2015. RAISE: a raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference (Portland, Oregon) (MMSys '15)*. Association for Computing Machinery, New York, NY, USA, 219–224. doi:10.1145/2713168.2713194
 - [13] Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. 2007. Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images. *IEEE Transactions on Image Processing* 16, 5 (2007), 1395–1411. doi:10.1109/TIP.2007.891788
 - [14] Miroslav Goljan, Jessica Fridrich, and Tomáš Filler. 2009. Large scale test of sensor fingerprint camera identification. In *Media Forensics and Security*, Edward J. Delp III, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong (Eds.), Vol. 7254. International Society for Optics and Photonics, SPIE, 72540L. doi:10.1117/12.805701
 - [15] D. Gagnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. 2021. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. doi:10.1109/ICME51207.2021.9428429
 - [16] Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. 2024. Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets. In *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 80–95. doi:10.1007/978-3-031-92089-9_6
 - [17] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. 2025. A Bias-Free Training Paradigm for More General AI-generated Image Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 18685–18694.
 - [18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773
 - [19] Jiaxi Jiang, Kai Zhang, and Radu Timofte. 2021. Towards Flexible Blind JPEG Artifacts Removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4977–4986. doi:10.1109/ICCV48922.2021.00495
 - [20] Black Forest Labs. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
 - [21] Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. <https://bfl.ai/blog/flux-2>.
 - [22] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image Restoration Using Swin Transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 1833–1844. doi:10.1109/ICCVW54120.2021.00210
 - [23] J. Lukas, J. Fridrich, and M. Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214. doi:10.1109/TIFS.2006.873602
 - [24] Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. 2024. When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–6. doi:10.1109/WIFS61860.2024.10810694
 - [25] Sara Mandelli, Diego Vila-Portela, David Vázquez-Padín, Paolo Bestagini, and Fernando Pérez-González. 2026. Beyond Spectral Peaks: Interpreting the Cues Behind Synthetic Image Detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 14307–14311. doi:10.1109/ICASSP55912.2026.11462389
 - [26] Francesco Marra, Diego Gagnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do GANs Leave Artificial Fingerprints?. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 506–511. doi:10.1109/MIPR.2019.00103
 - [27] OpenAI. 2024. Improving Image Generation with Better Captions. <https://cdn.openai.com/papers/dall-e-3.pdf>.
 - [28] Nischay Purnekar, Lydia Abady, Benedetta Tondi, and Mauro Barni. 2024. Improving the Robustness of Synthetic Images Detection by Means of Print and Scan Augmentation. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (Baiona, Spain) (IH&MMSec '24)*. Association for Computing Machinery, New York, NY, USA, 65–73. doi:10.1145/3658664.3659635
 - [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
 - [30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. doi:10.1109/CVPR52688.2022.01042
 - [31] Lea Uhlenbrock, Sandra Bergmann, and Christian Riess. 2025. Latent Landscapes: Topology of the Clip Feature Space for Synthetic Image Detection. In *2025 33rd European Signal Processing Conference (EUSIPCO)*. 51–55. doi:10.23919/EUSIPCO63237.2025.11226506
 - [32] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8692–8701. doi:10.1109/CVPR42600.2020.00872
 - [33] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing* 26, 7 (2017), 3142–3155. doi:10.1109/TIP.2017.2662206