IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 September 2025, accepted 1 October 2025, date of publication 3 October 2025, date of current version 9 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3617766

RESEARCH ARTICLE

A Novel Adversarial Gray-Box Attack on DCT-Based Face Deepfake Detectors

FRANCESCO GUARNERA[1], (Member, IEEE), LUCA GUARNERA[1], ALESSANDRO ORTIS[1], (Senior Member, IEEE), SEBASTIANO BATTIATO[1], (Senior Member, IEEE), AND GIOVANNI PUGLISI[2]

[1]Department of Mathematics and Computer Science, University of Catania, 95124 Catania, Italy
[2]Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy

Corresponding author: Luca Guarnera (luca.guarnera@unict.it)

The work of Francesco Guarnera was supported by MUR in the framework of Piano Nazionale di Ripresa e Resilienza (PNRR), through the Project "Future Artificial Intelligence Research—FAIR" under Grant PE0000013. The work of Luca Guarnera and Sebastiano Battiato was supported in part by the SEcurity and RIghts in the Cyberspace (SERICS) through the Ministero dell'Università e della Ricerca (MUR) National Recovery and Resilience Plan funded by European Union—NextGenerationEU under Grant PE00000014. The work of Alessandro Ortis was supported by the Research PIAno di inCEntivi per la Ricerca di Ateneo 2024/2026—Linea di Intervento i "Progetti di ricerca collaborative," Sistemi di IA Multimodale - Metodologie e Applicazioni (SIAM) Project, University of Catania, Italy.

ABSTRACT In recent years, several techniques have been developed to detect deepfake images, with particular success of approaches that exploit analytical traces (e.g. frequency domain features), such as those derived from the Discrete Cosine Transform (DCT). Despite their effectiveness, these detectors remain vulnerable to adversarial attacks. In this paper, we introduce a novel gray-box adversarial attack specifically designed to evade DCT-based deepfake detectors. Our method accurately tunes the AC coefficient statistics of synthetic images to closely match those of real ones, while preserving high visual quality. The attack assumes full knowledge of the DCT feature extraction process, but not access to the internal parameters of the classifiers. We evaluate the proposed method against a set of DCT-based detectors using deepfakes generated from both Generative Adversarial Networks (GANs) and Diffusion Models (DMs). Experimental results show significant degradation in detection performance, exposing critical weaknesses in systems traditionally considered interpretable and robust. This work raises important concerns about the reliability of frequency domain detectors in forensic and cybersecurity applications.

INDEX TERMS Adversarial imaging, DCT analysis, deepfake images, gray-box attack.

I. INTRODUCTION

Over the past decade, advances in Generative AI (GenAI), particularly Generative Adversarial Networks (GANs) [1] and, more recently, Diffusion Models (DMs) [2], [3], have enabled the automatic synthesis of images and videos—called deepfakes—whose visual fidelity is often indistinguishable from authentic content [4], [5]. While deepfakes have led to various creative applications, they have also blurred the line between what is real and what is fake, giving rise to what Casu et al. [4] have termed Impostor Bias: a generalized distrust that undermines the reliability of digital evidence. The malicious exploitation of deepfakes spans various fields, such as online fraud or pornography, representing a serious threat to both individuals and organizations. One of the most worrisome scenarios involves the dissemination of misinformation through faked videos, in which public figures make false statements. These forgeries can influence public opinion, disrupt democratic processes, destabilize governments, or instigate social disorder. Cybercriminals have used synthetic voices to impersonate corporate managers, obtaining transfers of money or confidential data [6]. Another serious danger is the use of deepfakes to violate people's privacy. Pornographic deepfake videos have been made using the faces of celebrities or ordinary people, often without their knowledge. This type of abuse has devastating consequences for the victim, damaging their reputation and causing psychological trauma. Thus, the

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

necessity to create advanced detection tools to fight the malicious use of deepfakes has arisen [7], [8], especially in the context of cybersecurity and personal data protection. The multimedia forensics community has proposed several detection algorithms, many of which are based on deep convolutional architectures that learn spatial or temporal artifacts [9], [10], while a parallel line of research focuses on more interpretable frequency domain descriptors. Among these, statistics extracted from the discrete cosine transform (DCT) of 8×8 blocks have been shown to be effective and light enough to be matched with simple classifiers [11], [12], [13], [14]. Despite their high classification accuracy, many of these approaches remain vulnerable to adversarial attacks, which subtly manipulate input data while preserving its semantic content, leading to significant drops in classifier performance. In this work, we focus on deepfake detectors based on the Discrete Cosine Transform (DCT), which are commonly regarded as explainable due to the interpretable nature of their features. We propose a novel adversarial method specifically designed to attack these classifiers, showing how easily their decisions can be manipulated through minimal perturbations. Our results reveal critical vulnerabilities even in detectors built on transparent and interpretable features, raising concerns about their robustness and reliability in forensic scenarios. In this paper, we present a threat model in which the adversary has full knowledge of the DCT feature extraction process but lacks access to the internal parameters of the downstream classifiers (i.e. SVM, Random Forest, CNN). This strategy qualifies as a gray-box attack since it assumes knowledge of the input feature space (i.e., DCT statistical profiles), but no access to the internal structure or parameters of the deepfake detector. Our approach involves refining deepfake images by precisely manipulating their AC coefficients statistics. Drawing inspiration from established adversarial attack methodologies, our technique introduces carefully crafted perturbations to the DCT domain coefficients of synthetic images. Specifically, the proposed solution, considering a mathematical model of AC coefficients (i.e., Laplacian distribution) aligns the statistics of the manipulated images to closely resemble those of genuine images. Finally, the Inverse Discrete Cosine Transform (IDCT) is applied to generate a synthetic image that is resilient enough to deceive state-of-the-art deepfake detection mechanisms. We evaluated the effectiveness of our attack against several DCT-based detectors using deepfakes generated by Generative Adversarial Networks (GANs) and Diffusion Models (DMs) [15]. Our findings reveal significant vulnerabilities in current detection methods, underscoring the need for continued innovation in developing robust deepfake detection strategies. The main contributions of this work are as follows:

- We propose a novel gray-box adversarial attack specifically designed to deceive DCT-based deepfake detectors by aligning the frequency statistics of fake images with those of real ones.
- We introduce a DCT-domain perturbation pipeline that achieves high fooling rates while preserving visual quality, leveraging both local rescaling and reference feature selection.

The remainder of this paper is structured as follows: state-of-the-art in the field of deepfakes detection is reviewed in Section II whereas the proposed approach is described in Section III. The experimental results are analysed in Section IV. Conclusions regarding the achieved results and potential hints for future investigation in this area are reported in Section V.

II. RELATED WORKS

Adversarial Machine Learning (AML) focuses on identifying and addressing weaknesses in machine learning models. Therefore, AML techniques play a crucial role in uncovering and mitigating the vulnerabilities inherent in predictive methods [16]. Insights provided by research in AML deepen our understanding and help strengthen the robustness and reliability of AI models against emerging threats such as deepfakes. In the context of identifying deepfakes, approaches based on Discrete Cosine Transform (DCT) [13], [14] leverage traces (i.e., footprints) created by generative models within the DCT space. The study presented in [17] examines how GAN fail to accurately reproduce the spectral distributions in real images. This shortcoming arises from the upsampling techniques employed in GANs, especially transposed convolutions, which introduce spectral distortions that can be exploited as indicators to detect fake or artificially generated images. The work in [18] examines the discrepancies in the Fourier spectrum between real images and those generated by GANs. The study focuses on the final layer of generators, which lead to varying degrees of high-frequency distortions in generated images. Such spectral discrepancies are used for detecting synthetic images. The paper in [19] investigates the vulnerability of neural networks trained to classify synthetic images to adversarial examples. It demonstrates that these networks can be easily fooled by introducing subtle, often imperceptible, modifications to the input images. The authors present both white-box and black-box attack strategies capable of causing these networks to misclassify synthetic images as real.

The research by Husasin et al. [20] highlights the vulnerabilities of deepfake detection systems. The study emphasizes the robustness of the introduced adversarial perturbations, showing that they can withstand image and video compression codecs commonly used in real-world scenarios. Attacks for both white-box and black-box scenarios have been presented. The work in [21] proposes a frequency adversarial attack method against face forgery detectors. Specifically, the authors apply DCT to the input images and introduce a fusion module to capture the salient regions of adversarial perturbations in the frequency domain. Compared to existing adversarial attacks in the spatial domain (e.g., FGSM, PGD), this method is more imperceptible to human observers.

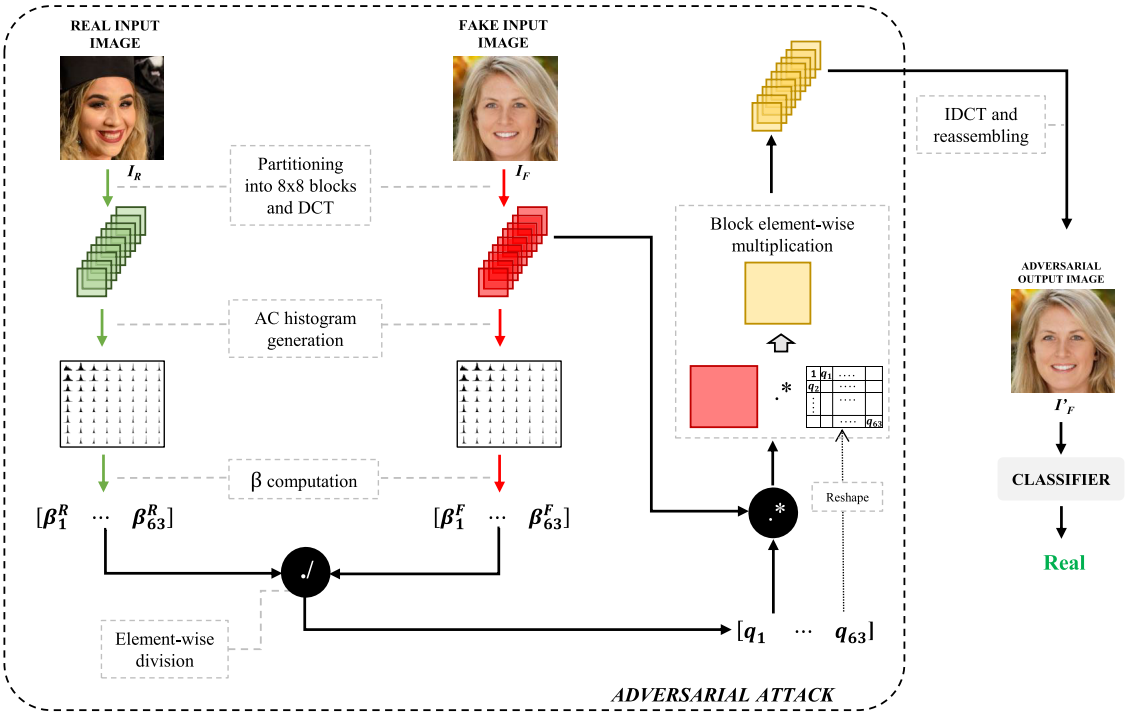


FIGURE 1. Basic solution pipeline. The vectors of features (i.e., the β values related to AC modes) are computed from both real (I_R) and fake image (I_F) respectively. Element-wise ratio between feature vectors is computed (q_i values) and employed to rescale the DCT values computed from I_F. The final adversarial fake image I'_F is then obtained applying the IDCT to the rescaled DCT values.

Furthermore, the authors proposed a hybrid adversarial attack achieving favorable attack performance on both spatial-based and frequency-based fake face detectors. The research in [22] introduced StatAttack, a method designed to evade Deepfake detectors that rely on statistical inconsistencies in images. StatAttack focuses on minimizing the statistical differences between real and fake images by introducing three statistical-sensitive degradations: exposure adjustments, blurring, and noise addition. These degradations are applied to fake images in a way that mimics natural image variations. Unlike existing frequency-based attacks, our method performs the attack with no access to the internal parameters of the classifiers and dataset used for training ensuring high imperceptibility.

III. PROPOSED APPROACH

The proposed method aims at deceiving classifiers based on DCT features. Note that, in almost all approaches [11], [12], [13], [14], the input image I is partitioned and DCT is applied to each 8×8 block. Moreover, the AC coefficients related to each mode are usually modeled as a zero-centered Laplacian distribution [23]:

$$P(x|\beta) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta}\right) \tag{1}$$

where β is the scale parameter obtained by maximum likelihood estimation close form solution. The image content is then summarized by a set of these scale parameters

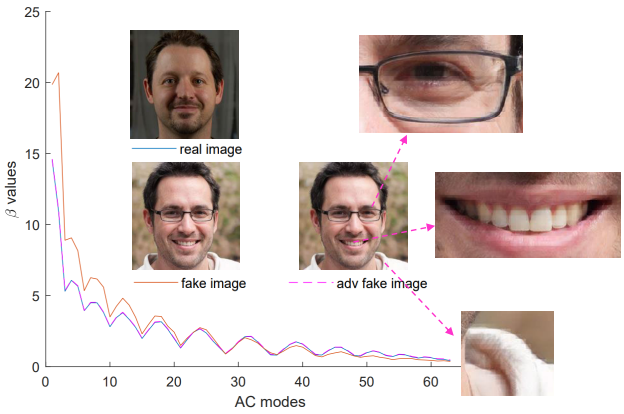


FIGURE 2. Example of adversarial image I'_F generated applying the basic solution depicted in Figure 1. Although the vector of features (i.e., β values related to AC modes) of real image I_R and I'_F are similar, due to the block based computation of the described solution, visible blocking artefacts are generated (see enlarged details).

$[\beta_1, \ldots, \beta_{63}]$ that can be exploited to perform the various classification tasks ([11], [13], [24], [25], [26]).

The proposed solution tries to modify the feature values, i.e., the vector of β extracted from a generic fake image I_F, to become similar to the ones computed from a reference real picture I_R (see Figure 1). In particular, the following property of Laplace distributions has been exploited: if a random variable $X \sim Laplace(0, \beta)$ then $qX \sim Laplace(0, q\beta)$.

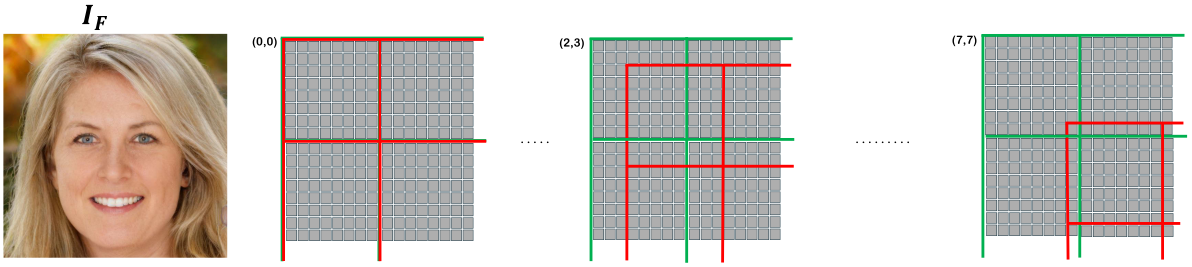


FIGURE 3. Schematic illustration of the 64 cropped images generation process.

A proper rescaling of the 8×8 DCT blocks can be then employed to obtain the target feature vector. Taking into account the $i-th$ AC mode ($i = 1, \ldots, 63$), the ratio $q_i = \beta_i^R / \beta_i^F$ with β_i^F and β_i^R values of β related to fake and real images respectively, is computed. These q_i values are then multiplied by the corresponding AC coefficients of the fake picture I_F. Finally, IDCT (Inverse Cosine Discrete Transform) is applied to 8×8 blocks and the final picture I'_F is generated. The overall pipeline is shown in Figure 1.

The described approach can be exploited to generate an adversarial image I'_F with a feature vector similar to the one computed from a real picture I_R. However, due to the 8×8 partitioning and operations applied to each block independently, blocking artifacts are quite evident from a visual inspection in the output picture I'_F (see Figure 2). To sum up, although the solution depicted in Figure 1 is able to effectively deceive a classifier based on DCT features, the overall quality of the output picture has been considerably reduced. To cope with this problem, several strategies have been adopted. A first improvement consisted in the application of the basic algorithm to 64 cropped versions of the $N \times M$ input fake picture I_F generated as follows: $I_F^s = I_F(r+1 : N-8+r, c+1 : M-8+c)$ with s linear shift index computed as $r \times 8 + c$ ($r, c \in \{0, 1, ...7\}$). Later, each $(N-8) \times (M-8)$ adversarial picture generated from I_F^s is zero padded adding r and $8-r$ rows on top and bottom, c and $8-c$ columns on left and right respectively (this step is shown in Figure 3). In this way, the 64 $N \times M$ adversarial padded pictures $I_F^{s\prime}$ are aligned with respect to the original image content. Note that, the misalignment introduced in these cropped versions of I_F, implies a different 8×8 partitioning and hence adversarial output picture $I_F^{s\prime}$. As a consequence, also the blocking artifact locations of the 64 $I_F^{s\prime}$, are not aligned to each other and depend on s (i.e., the linear shift index). The misalignment of these visual perturbations can then be exploited to design a proper filtering strategy to reduce their impact. Specifically, the adversarial image I'_F is computed from this set of images ($I_F^{s\prime}$ with $s \in \{0, 1, ...63\}$) concatenated in a $3D$ tensor applying a median filter along the third dimension followed by a cropping (an 8 pixel border is removed) to avoid the influence of the previously included zero padding.

Algorithm 1 Adversarial Image Generation Pipeline Based on DCT Manipulation

Require: D_R : dataset of real pictures,
I_F : fake input picture,
k : number of clusters
Output: adversarial picture I'_F

$\beta_{real} \leftarrow get\_beta(D_R)$
$Centroids \leftarrow kmeans(\beta_{real}, k)$
$\beta_{fake} \leftarrow get\_beta(I_F)$
$\beta_{target} \leftarrow nearest\_centroid(\beta_{fake}, Centroids)$
for each shift (r, c) in 8×8 grid **do**
  $I_F^s \leftarrow crop(I_F, r, c)$
  $\beta_s \leftarrow get\_beta(I_F^s)$
  $q \leftarrow \beta_{target} / \beta_s$
  $B \leftarrow DCT\_blocks(I_F^s)$
  $B \leftarrow B \cdot q$
  $I_F^{s\prime} \leftarrow IDCT\_reconstruct(B)$
end for
$I \leftarrow median\_combine(I_F^{s1\prime}, \ldots, I_F^{s64\prime})$
$I \leftarrow crop\_borders(I)$
$\beta \leftarrow get\_beta(I)$
$q \leftarrow \beta_{target} / \beta$
$B \leftarrow DCT\_blocks(I)$
$B \leftarrow B \cdot q$
$I'_F \leftarrow IDCT\_reconstruct(B)$

Although artifacts have been considerably reduced (see Figure 4(a)), the feature vector of the adversarial fake image, especially for the low frequency components, is also modified by the additional filtering process. This shortcoming can be reduced by a proper selection of the reference profile (i.e., feature vector representing real images). Starting from a dataset of real images D_R, k different sets of pictures have been obtained through k-means algorithm applied in the feature space and the related centroids are also computed. Later, the centroid closest to the profile extracted from the fake image under analysis is then considered as the reference one. This solution allows us to obtain a feature vector similar to those extracted from real pictures without generating visual artifacts (see Figure 4(b)). Note that, this strategy includes several special cases. For example, if $k = 1$ the reference

TABLE 1. Collected images for LQ and HQ scenarios.

	Resolution	# Images	Real	Deepfake
HQ	1024×1024	10,000	FFHQ	StyleGAN2
LQ	256×256	10,000	StyleGAN2	ProGAN VQGAN DDIM LDM

profile corresponds to the average feature vector computed considering all the pictures in the dataset D_R. On the other hand, if $k = |D_R|$, the closest real image (in the feature space) is selected as the reference one. The impact of this hyperparameter (i.e., the number of clusters) in terms of trade-off between the effectiveness of the adversarial attack and image quality preservation is deeply studied in Section. IV. To sum up (see Algorithm 1), starting from the input picture I_F, 64 cropped versions are generated (I_F^s with $s \in \{1, 2, \ldots, 63\}$). Later, the basic algorithm is applied to these cropped versions together with a proper zero padding ($I_F^{s\prime}$). A 3D tensor is then built from the adversarial pictures $I_F^{s\prime}$, and a median filter followed by a cropping is applied to generate I_F'. Finally, due to the similarity of fake and reference feature vectors, the algorithm depicted in Figure 1 is applied a last time to the previously computed I_F'.

IV. EXPERIMENTAL RESULTS

The method proposed in the previous section has been designed to properly modify some features of a fake image to lead to errors in a classifier. Specifically, the goal of the attack is to bring about a binary classifier, capable of distinguishing between real and fake images, to classify the fake images as real. To test the effectiveness of the proposed attack, we proceeded as follows:

1) Choose a balanced and representative dataset of real and fake images.
2) Train and test some binary DCT-based classifiers that can distinguish between real I_R and fake I_F images with satisfactory accuracy.
3) Attack the fake images I_F of the test set with the method detailed in Section III generating the adversarial fake images I_F'.
4) Compare the results of the classifiers by giving as input fake images I_F (before) and adversarial images I_F' (after) and quantify the performance drop.

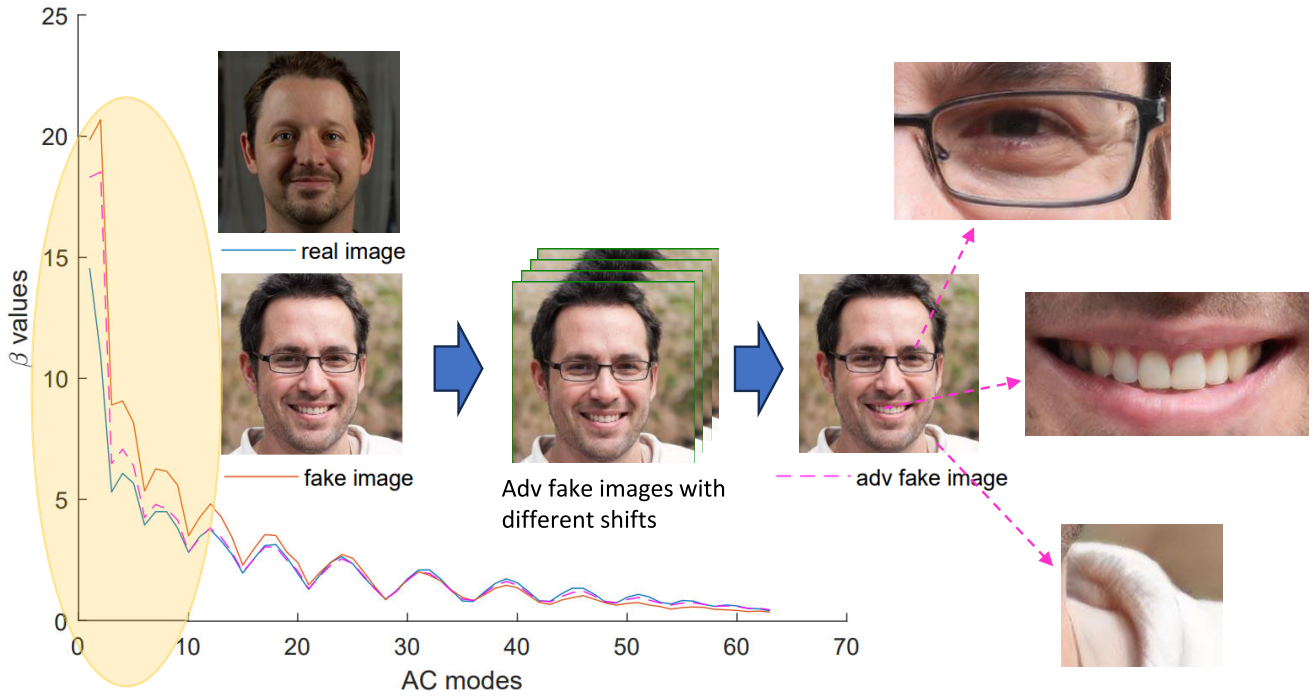
In the following, the aforementioned steps will be illustrated.

A. EMPLOYED DATASET

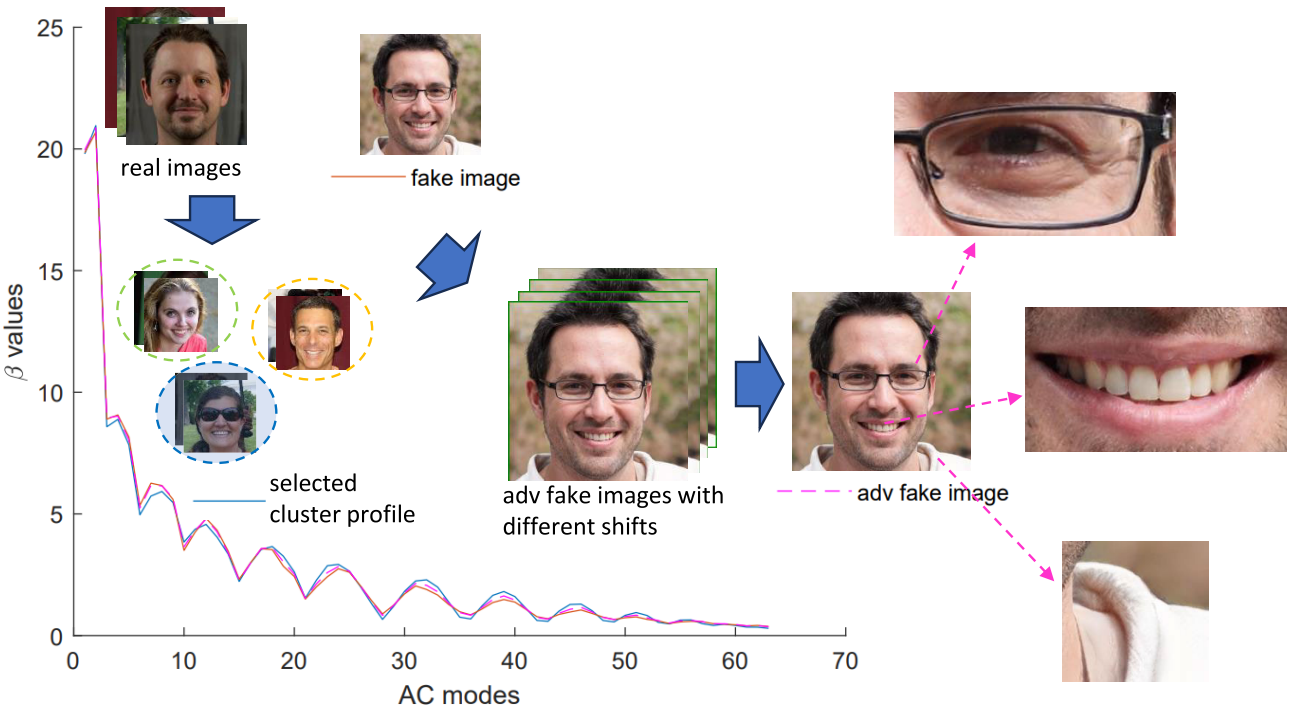
The collected dataset consists of both real and deepfake data. In the frequency analysis of an image, the higher the resolution, the more precise it is, so we decided to test the proposed method in two different scenarios, High Quality (HQ) and Low Quality (LQ). For the HQ scenario we chose FFHQ [27] as real dataset and StyleGAN2 [28] as fake; both datasets are composed of 50001024 $\times$ 1024 images. For the LQ scenario CelebA [29] was selected as real while 4 different deepfake datasets were adopted to test the method: the analysis carried out by [30] demonstrated the deep difference between DMs and GANs generation, so ProGAN [31], VQGAN [32], DDIM [33] and LDM [3] were selected. The division of the DeepFake LQ dataset into GAN and DM samples has been introduced to highlight the generator-agnostic nature of the attack. This distinction will not be mentioned in the following analysis. Also for LQ scenario the datasets real and fake are composed of 5000256 $\times$ 256 images (the 5000 fake images are divided into the 4 classes); the overall datasets employed are described in Table 1. In total, we collected 10, 000 real images and 10, 000 deepfake images. The dataset is balanced in terms of categories (real and deepfake) and the number of images generated by each architecture in LQ scenario (GANs and DM). The dataset was split into 80% training set, 10% validation set, and 10% test set and all the images are related to the same context: FACE (see Figure 2).

B. CLASSIFIERS

It is clear that our attack is linked to those classifiers that in some way exploit features closely related to DCT, our work should therefore be regarded as complementary to adversarial approaches operating in the spatial domain, as it specifically targets frequency-domain traces left by deepfakes. In order to demonstrate the generalization capability of the attack, showing that the proposed method is not tied to a specific classifier, many have been trained. Note that, to force the classifiers to base their decision on DCT features, the input at training time is represented by the 63β values computed from the AC distributions. We chose to train two ML classifiers (Support Vector Machine, SVM and Random Forest) and a Deep Learning one (a custom Convolutional Neural Network, CNN). SVM (with RBF kernel) is a margin-based classifier that uses a Gaussian kernel to handle non-linear class boundaries, offering strong generalization on moderately sized datasets. Random Forest is an ensemble of decision trees that improves robustness and reduces overfitting through bootstrap aggregation and feature randomness, excelling with noisy or tabular data. CNN is a deep learning model specialized for grid-structured data (e.g., images), automatically learning spatial hierarchies of features via convolutional layers. In this work, SVM and Random Forest has been simply implemented using the 63β values as input and their relative binary labels while the custom CNN presents 5 convolutional layers and a final fully connected layer followed by a softmax function to perform the final classification. A visual representation of the CNN is visible in Figure 6 trained for 20 epochs with an initial learning rate of 0.01, a learning rate drop factor of 0.2 and a mini batch size of 64. For the Vision Transformer (ViT) model, we adopted the ViT B/16 [34] configuration with 384 $\times$ 384 input resolution. The network was fine-tuned using a binary classification setup with 2 output classes, a batch size of 15, a learning rate



(a) First step of improved approach with the reduction of the generated artefacts. The basic algorithm (see Figure 1) is applied to 64 shifted version of the input fake picture I_F. Later, the adversarial image is computed from this set of images applying a median filter. Although artefacts have been considerably reduced (see enlarged details), the additional step modifies the feature vector of the adversarial fake image, especially for the low frequency components.



(b) The output of step (a) is used to obtain an adversarial fake image close to real pictures. Taking into account a dataset of real images, k different sets of pictures are obtained through a clustering algorithm (e.g., k-means) in the feature space. The centroid closest to the profile (i.e., feature vector) extracted from the fake image under analysis is then considered as the reference one.

FIGURE 4. Improved approach with the reduction of the generated artefacts (a) and the extraction of profile centroid (b). This solution, allow us achieving a feature vector similar to those extracted from real pictures without generating visual artefacts.

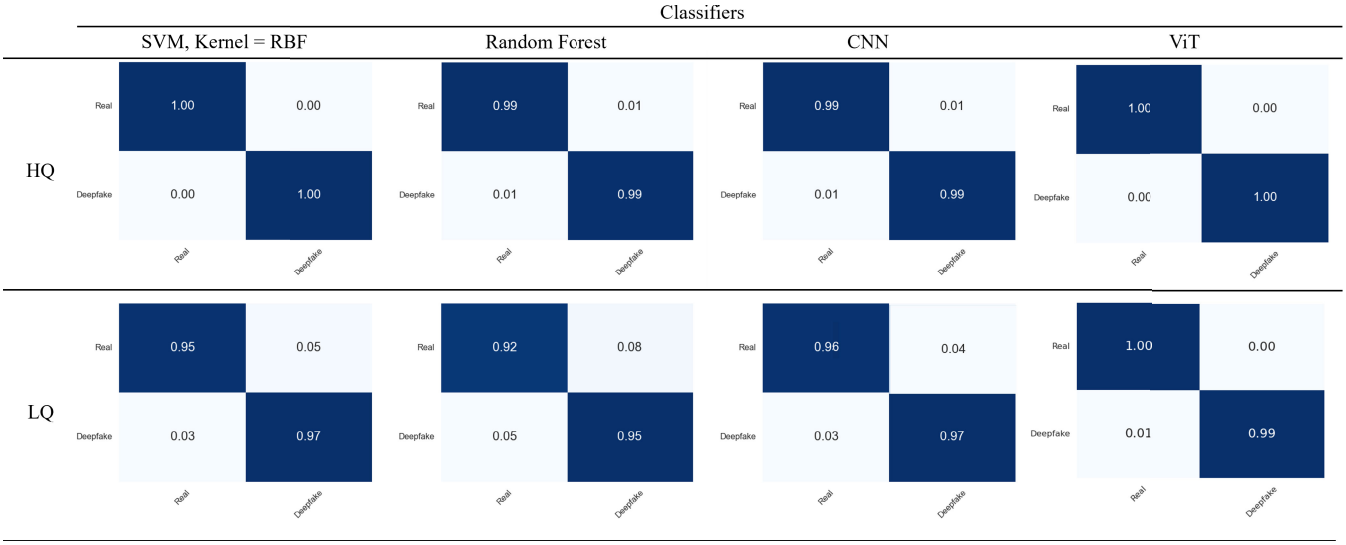


FIGURE 5. Confusion matrix for each couple classifiers/dataset. High quality scenario describes an almost perfect classification while LQ scenario show negligible wrong classifications. Good classifiers are the basis to demonstrate the goodness of the attack described in this work.

FIGURE 6. CNN structure with type of operations and activation values.

of 1e-5, and trained for 100 epochs. Each of the three methods creates different models by exploiting different features of the dataset, so an attack that works on all three means having an adversarial method that is not overfitted on the characteristics of the model itself. The performances of the classifiers are reported in Figure 5, where for each classifier (y axis) the confusion matrix of both scenarios (LQ, HQ) is shown. Note that, in confusion matrices y axis represents the ground truth while x axis the predictions. The results shown in Figure 5 demonstrate how the 63β values computed from AC distributions are discriminatory for the binary classification of real/deepfake images.

C. EXPERIMENTAL ATTACK SCENARIOS

Starting from the test set described in Section IV-A (500 images for both scenarios) the attack of Section III was performed. Following the pipeline in Figure 4 the DCT values of each fake test image I_F were adapted referring to the profile extracted from real images via the proposed method. Given a fake image I_F, the real reference profile must be calculated starting from a reference dataset D_R composed of real images.

The effectiveness of the proposed adversarial strategy has been evaluated under three distinct *feature-aware* scenarios resembling white-box conditions, where the attacker has access to the input features rather than the model itself. Note that this setting is more challenging than a standard white-box scenario, as it requires crafting effective attacks without direct knowledge of the model's architecture or parameters. This setup allows a systematic assessment of how different levels of data access influence the success of adversarial perturbations. The scenarios differ in the type of reference dataset used to guide the perturbation process:

- Training-aware attack: the adversary has access to the real images used to train the classifier, namely the D_{train} data including real images used to train the classifier.
- Test-aware attack: the adversary exploits the D_{test} real data used during the testing phase. While the features are known, this scenario simulates an attacker who can

access evaluation data, a realistic setup for assessing robustness [20].

- Cross-dataset attack: the adversary selects reference real images from an external dataset with different semantic and statistical properties, named D_{ext}. Despite not knowing the images used for training or testing, the attack remains effective due to shared DCT-based features. This scenario is aligned with the notion of cross-domain adversarial attacks discussed in [21], where perturbations are designed to generalize across datasets.

The three scenarios describe different cases in which different knowledge is owned by the attacker. Training-aware and test-aware attacks assume the knowledge of the dataset (real part) in addition to the feature (DCT) used to classify, while in the last scenario the attacker is only aware of the latter. Knowing the training dataset is often realistic since many works disclose it, and frequency information (e.g., via DCT) is commonly exploited by classifiers either directly or indirectly, making both assumptions plausible in practice.

D. RESULTS

The fairness of the experiments is represented by Figure 5 where the results of all classifiers are almost identical in all tests, validating them. The effectiveness of our adversarial attack is demonstrated in Table 2, where is shown how all the classifiers drop a relevant percentage in the detection of deepfake images (false negative rates on Deepfake images). The proposed adversarial attack has been evaluated under varying conditions, considering image quality, classifier type, and reference data scenario. On low-quality images (LQ), CNNs exhibit decreasing performance degradation as the number of clusters increases, with a drop at 64 clusters. Random Forest follows a similar trend with higher results for each cluster, while SVM in this scenario always reaches maximum results. This behavior is consistent across both training-aware and test-aware settings, while using an external dataset (FFHQ), the attack reaches the best performances; this behavior could depend on the FFQH resolution (1024×1024) which results greater than the input image, so the DCT frequencies seem to be more accurate and closer to the real ones. On high-quality images (HQ), the degradation is lower and the performance becomes less evident with increasing clusters. In this case, the attack has a lower effect on Random Forest w.r.t. CNNs, in contrast to the LQ scenario. In this case, the cross-dataset scenario describes a behavior consistent across all the classifiers, as for the LQ cross-dateset scenario.

The analyzed results suggest that the richer frequency content in high-resolution images reduces the impact of adversarial perturbations, probably due to the best representation of classes in classifiers. In cross-dataset scenarios, the attack generally reach better results with all classifiers that got bad results always. In training/test-aware attacks there is a different behavior of ML/DL classifiers w.r.t. resolution: ML classifiers work better in LQ scenario, while the results of CNN are higher in HQ scenario. The results suggest that ML methods are more vulnerable to the proposed attack on low-quality images, as DCT-based perturbations more strongly alter the statistical distribution of the extracted β features, which they rely on for classification. As a result, the same perturbation becomes less effective in high-resolution contexts, emphasizing the role of input complexity in adversarial robustness.

In addition to traditional machine learning and convolutional approaches, we also evaluated the robustness of Vision Transformers (ViT) against the proposed attack. Results reported in Table 2 show that ViT classifiers, although capable of handling frequency-based features, are similarly susceptible to DCT-domain perturbations. Specifically, in the LQ scenarios, ViT classifiers exhibit a performance drop comparable to CNNs, particularly in the training-aware and test-aware settings where the false negative rate increases as the number of clusters grows. Notably, in the cross-dataset scenario, the attack is especially effective: while the SSIM values remain high, the false negative rate for ViT peaks, confirming the generalization ability of our method. In HQ scenarios, ViT maintains slightly higher resilience than other models, but still demonstrates reduced detection capability as the perturbation becomes more sophisticated. Overall, these findings confirm that even advanced transformer-based architectures are vulnerable to our adversarial attack when relying on DCT-derived statistical features.

Across all settings, Structural Similarity Index Measure (SSIM) [36] analysis confirms a trade-off between stealthiness and effectiveness. Indeed, structural similarity increases with cluster count and reaches high values in both settings (HQ and LQ); this can be explained by the fact that with more centroids, the one selected is closer to the fake image, resulting in a less effective attack but higher SSIM. The SSIM/MEAN(FN) curves (Figure 9) further highlight this relationship: SSIM represents visual quality, while MEAN(FN) quantifies attack effectiveness, with different curves and points corresponding to various attack types, cluster configurations and reference datasets. In the HQ scenario, the larger number of 8×8 blocks reduces the effect of modifying a single block's DCT coefficient, making perturbations less impactful on the classifier's decision. This explains why the attack is less effective compared to the low-resolution case. An example of obtained attacked images in different scenarios is provided in Figure 7. In addition to SSIM, we also considered the Learned Perceptual Image Patch Similarity (LPIPS) metric [37], which evaluates perceptual similarity by leveraging deep network features rather than relying solely on structural fidelity. As reported in Table 2, LPIPS values remain consistently low across different attack scenarios, confirming that the introduced perturbations are largely imperceptible to the human eye. While SSIM highlights the preservation of global structural information, LPIPS provides a more fine-grained, perceptual assessment, reinforcing the evidence that our adversarial strategy maintains high visual quality.

TABLE 2. False negative rates for our Deepfake images generated by our methods across different SSIM- LPIPS levels, classifier types, and reference scenarios. Columns correspond to values of number of clusters used during evaluation.

LQ								
Training-aware attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	1	1	1	1	1	1	1	1
Random Forest	1	1	1	0.88	0.82	0.83	0.75	0.79
CNN	0.87	0.93	0.92	0.82	0.72	0.81	0.67	0.74
ViT	0.89	0.91	0.93	0.84	0.75	0.89	0.71	0.72
SSIM - LPIPS	0.93 - 0,065	0.96 - 0,049	0.97 - 0,040	0.98 - 0,034	0.98 - 0,028	0.98 - 0,024	0.99 - 0,021	0.99 - 0,017
Test-aware attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	1	1	1	1	1	1	1	1
Random Forest	1	1	1	0.92	0.8	0.8	0.84	0.78
CNN	0.83	0.90	0.84	0.90	0.76	0.65	0.65	0.67
ViT	0.89	0.91	0.90	0.83	0.78	0.79	0.70	0.73
SSIM- LPIPS	0.93 - 0,065	0.96 - 0,049	0.97 - 0,038	0.98 - 0,033	0.98 - 0,029	0.98 - 0,024	0.99 - 0,020	0.99 - 0,017
Cross-dataset attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	1	1	1	1	1	1	1	1
Random Forest	1	1	1	1	1	1	0.99	0.99
CNN	1	1	1	1	1	1	1	1
ViT	1	0.87	0.86	0.88	0.84	0.78	0.7	0.64
SSIM- LPIPS	0.86 - 0.094	0.93 - 0.057	0.95 - 0.045	0.96 - 0.037	0.96 - 0.033	0.97 - 0.028	0.97 - 0.025	0.97 - 0.022
HQ								
Training-aware attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	0.99	0.96	0.92	0.86	0.83	0.79	0.73	0.7
Random Forest	0.99	0.97	0.88	0.78	0.77	0.72	0.62	0.56
CNN	0.99	0.97	0.97	0.94	0.93	0.91	0.89	0.88
ViT	1	0.97	0.97	0.91	0.89	0.86	0.8	0.8
SSIM- LPIPS	0.98 - 0.030	0.98 - 0.023	0.99 - 0.019	0.99 - 0.015	0.99 - 0.012	0.99 - 0.010	0.99 - 0.009	0.99 - 0.008
Test-aware attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	0.99	0.96	0.91	0.86	0.83	0.78	0.76	0.7
Random Forest	0.99	0.96	0.85	0.81	0.75	0.63	0.62	0.54
CNN	0.99	0.97	0.95	0.94	0.92	0.94	0.91	0.86
ViT	1	0.97	0.97	0.94	0.9	0.88	0.83	0.77
SSIM- LPIPS	0.98 - 0.030	0.98 - 0.023	0.99 - 0.018	0.99 - 0.015	0.99 - 0.012	0.99 - 0.011	0.99 - 0.009	0.99 - 0.009
Cross-dataset attack								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**
SVM (RBF)	1	1	1	1	1	1	1	0.99
Random Forest	1	1	0.94	0.96	0.98	0.98	0.77	0.83
CNN	1	1	1	1	1	1	1	1
ViT	1	1	1	1	1	0.99	0.95	0.97
SSIM- LPIPS	0.95 - 0.079	0.96 - 0.071	0.96 - 0.069	0.97 - 0.051	0.97 - 0.042	0.98 - 0.035	0.98 - 0.032	0.98 - 0.028

Notably, the joint analysis of SSIM and LPIPS indicates a favorable trade-off: as SSIM approaches values close to 1.0, LPIPS decreases below 0.05 in most cases, meaning that the attacks not only deceive classifiers but also yield outputs that are perceptually indistinguishable from genuine images. Regarding generation time, which depends solely on image resolution, the attack was executed on a standard laptop without GPU or hardware acceleration, requiring on average 471 and 10,512 ms per low- and high-resolution input respectively.

E. COMPARISON

As discussed previously, our method targets a specific attack scenario involving classifiers designed to analyze image frequency components, either directly or indirectly. Unlike from us most existing methods rely on different feature domains or prior knowledge of the classifier, therefore, we compare primarily with which shares similar conditions. Recently, the authors of [35] proposed an attack method that exploits the same type of features used in our approach. In order to compare our method, the attack of [35] was tested in all the same scenarios described in Table 2. Similarly to our method, the approach proposed in [35] performs the attack by varying a parameter, specifically the frequency components that are considered. Accordingly, Figure 8 presents, for each scenario and method, a graph in which the blue line indicates the SSIM value, reflecting the preserved image quality after the attack, while the orange line represents the average false negative rate obtained across the classifiers

FIGURE 7. Quantitative and qualitative comparison of original and attacked images.

described in Section IV-B. An analysis of Figure 8 reveals that both methods, despite variations in their respective parameters (number of clusters in our approach and number of frequencies in [35]), aim to preserve a high similarity with the original image, as indicated by the high SSIM values. However, in our method, this similarity is achieved together with a high attack effectiveness, whereas in the method proposed in [35] there is a high probability that using more frequencies the attack quality will improve, but since they are the lowest, this should lead to a lower SSIM. Finally, it is important to highlight the different behavior of our method when the attack is performed using an external profile. Although the SSIM is slightly lower in this case, the attack achieves better performance in certain settings. Conceptually, if we consider the decision boundary in the feature space defined by the classifiers trained for each scenario, using the same profile for training and testing allows the attacked image to cross the fake/real boundary while remaining close to it. In contrast, the use of an external profile displaces the image further in the feature space, potentially enabling more effective attacks at the cost of a slight reduction in perceptual image quality (see Figure 7).

F. ADVERSARIAL ATTACK ROBUSTNESS TO POST-PROCESSING

In order to evaluate the robustness of the proposed attack, we applied a set of common image manipulations to adversarial examples and analyzed their behavior when processed by a Vision Transformer (ViT) model. It is important to emphasize that the purpose of this evaluation is not to measure the robustness of the classifier itself (indeed, we rely on the pretrained ViT used only in inference mode, without retraining) but rather to assess whether the adversarial perturbations introduced in the DCT domain maintain their effectiveness after post-processing. In other words, the classifier is not the focus of our evaluation; rather, we aim to test the persistence of the attack itself, verifying whether adversarial fake images continue to be misclassified as real after undergoing realistic transformations. The considered manipulations include JPEG compression with quality factors QF = {50, 60, 70, 80, 90}; rescaling at 50% (downscaling) and 200% (upscaling); rotations by 45°, 135°, 225°, and 315°; and Gaussian noise with $\sigma = 3$ and $\sigma = 9$. For each setting, we measured the False Negative Rate (FNR).

The results shown in Figure 10 clearly demonstrate that in many cases the adversarial effect remains almost entirely intact.

Under HQ training-aware setting, the attack remains almost fully effective against JPEG compression up to QF = 80, with only slight drops at QF = 90. Upscaling (to 200%) does not reduce the effectiveness of the attack, while rotations slightly reduce the false negative rate (to around 0.85-0.93 depending on the angle). Adding Gaussian noise ($\sigma = 3$) still produces very high FNRs (close to 1), although strong noise ($\sigma = 9$) drops the FNR to around 0.74-0.81. The worst post-processing for

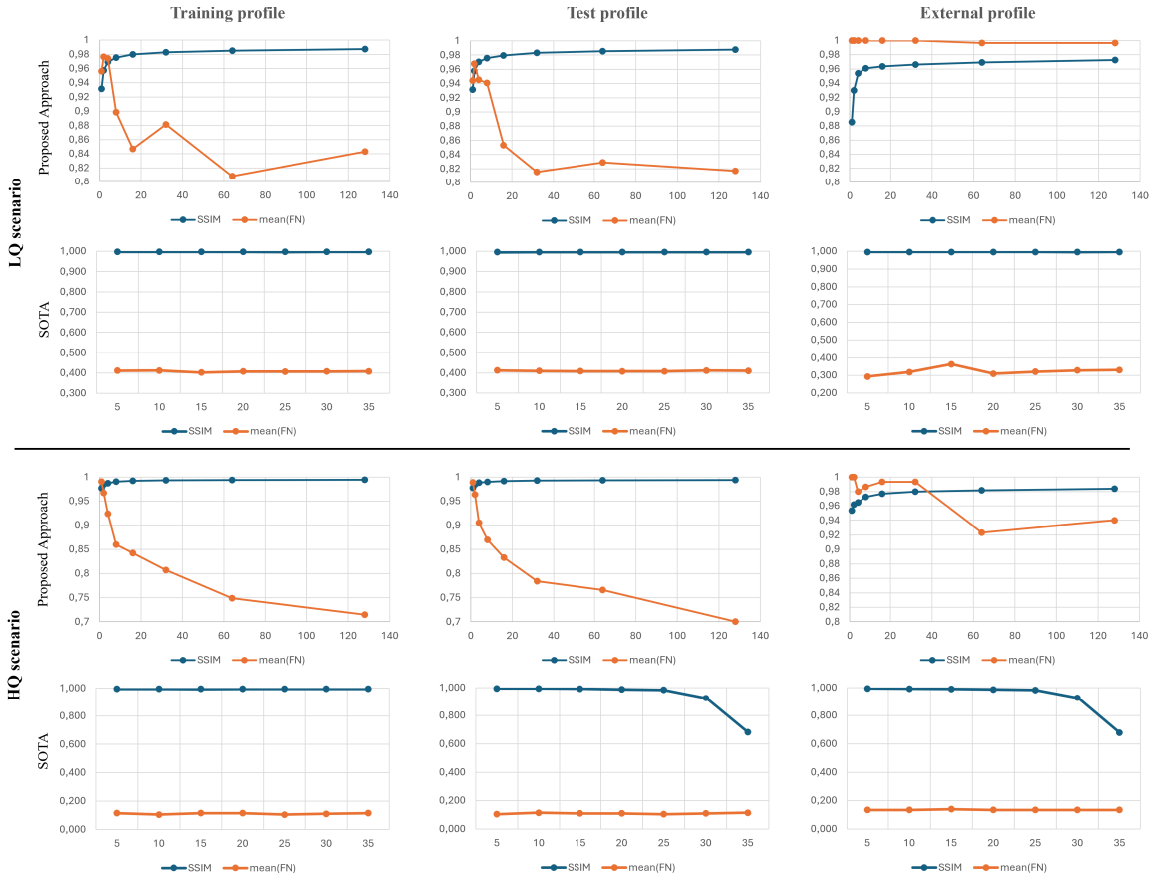
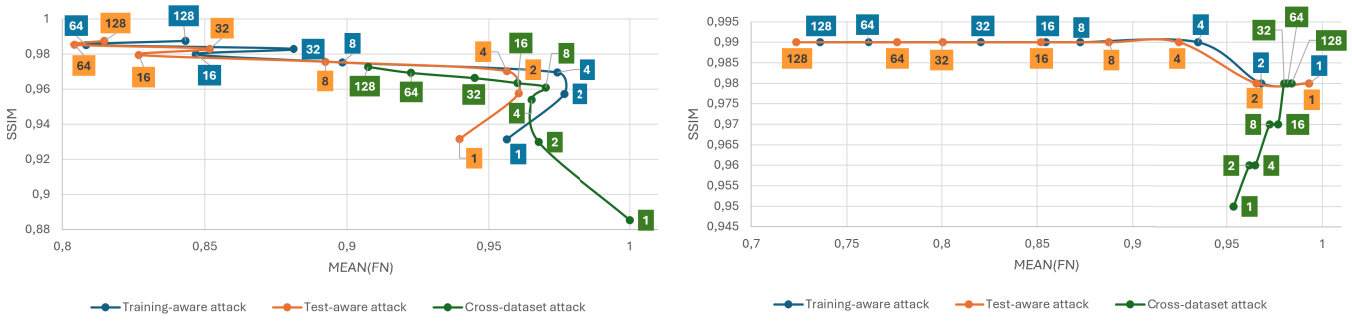


FIGURE 8. Results of our method compared with [35] in terms of SSIM and mean False Negative which in this case represents the effectiveness of the attack.



(a) (b)

FIGURE 9. SSIM vs. MEAN(FN) trade-off curves in different scenarios: LQ images on the left and HQ images on the right, across various attack configurations, (number of clusters) and reference datasets.

HQ train-aware is 50% downscaling, which drastically lowers the FNR (to ∼0.08-0.21). In the HQ test-aware setting, the results are very similar: performance under compression and mild transformations remains robust, while downscaling and strong noise affect the attack. In the external HQ scenario, the attack is again largely preserved under the same mild transformations, and even somewhat resilient under downscaling for larger clusters (i.e., FNR increases compared to train-aware/test-aware under downscaling), although noise of $\sigma = 9$ still causes a drop.

Moving on to the LQ scenario, similar patterns persist, but with overall lower robustness. In the training-aware LQ scenario, JPEG compression up to QF = 70 does not affect the attack (FNRs are still high), but higher compression (QF = 80-90) produces more evident drops; Slight rotations still maintain moderate success for some degrees (particularly 45°, 315°), but other rotations show greater degradation. Gaussian noise with $\sigma = 3$ already causes a drop in FNR (sometimes up to ∼0.5-0.6), while $\sigma = 9$ almost completely eliminates the attack (values close to zero).

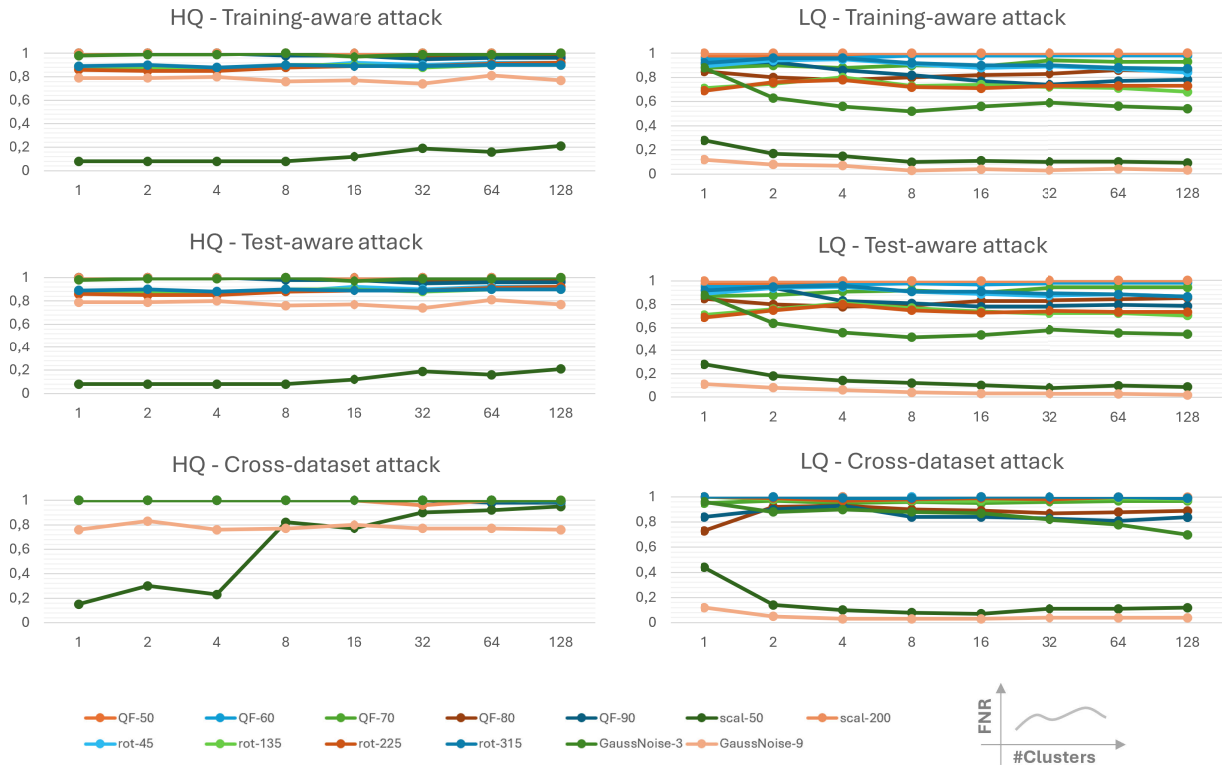


FIGURE 10. False Negative Rate (FNR) of the proposed adversarial attack under different manipulations, evaluated on HQ and LQ images across training-aware, test-aware, and cross-dataset scenarios. The x-axis indicates the number of clusters, while the y-axis reports the corresponding FNR values.

Downscaling also has negative effects on the attack, with very low FNRs. In the test-aware LQ scenario, the results are similar to those in the training-aware scenario, showing the same weakness against strong noise and downscaling. In the external LQ scenario, attack performance improves: JPEG compression even at QF = 90 maintains relatively high FNRs, and although downscaling and strong noise continue to reduce performance, the results are better than the training-aware/test-aware LQ configuration in terms of robustness.

Crucially, these findings do not reflect the resilience of the ViT model itself, but rather confirm that the perturbations crafted in the DCT domain retain their deceptive effect across most practical conditions, and fail only when the underlying frequency structure of the image is severely compromised.

V. CONCLUSION

This study presents a comprehensive analysis of current DCT-based deepfake detection methods introducing a novel adversarial attack designed to exploit these weaknesses. The proposed attack involves precise manipulation of DCT coefficients to make deepfake images closely to real ones, thereby significantly reducing the detection of various classifiers. Specifically, the attack strategy aligns the statistical properties of the manipulated images with those of real images by adjusting the AC coefficients in the DCT domain, thus generating synthetic images that are resilient against detection mechanisms. Results show that ML/DL-based deepfake detectors suffer a significant accuracy drop when using different reference datasets. In this specific scenario, the proposed adversary attack proves to be highly effective. From our perspective, a potential defence strategy could involve the use of an ensamble of classifiers or a model trained specifically to recognise this attack. This could be achieved by incorporating adversary examples, created with knowledge of our method of attack, into the training process or using specially constructed strategies against adversarial attacks on deepfake images. To confirm this we did a simple test considering the HQ scenario with 64 clusters under the Test-aware attack scenario, where the attack exhibits varying effectiveness across different classifiers. In this scenario we simulated a defense against False Negatives by constructing an ensemble of two classifiers (Random Forest and ViT), classifying an image as real only if both classifiers agreed. The attack success decreased from 0.62 (Random Forest, the best-performing single classifier) to 0.59; to note that this strategy is restrictive for adversarial attacks, as it assumes prior knowledge of the defense. While this approach could increase the robustness of the model against similar attacks, it would likely result in a reduction in overall classification performance due to the inherent trade-off between robustness and accuracy. Future work will explore integrating DCT features with additional cues and testing classifiers trained without frequency-based features.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.

[2] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[4] M. Casu, L. Guarnera, P. Caponnetto, and S. Battiato, "GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions," *Forensic Sci. Int., Digit. Invest.*, vol. 50, Sep. 2024, Art. no. 301795.

[5] A. U. Hirte, M. Platscher, T. Joyce, J. J. Heit, E. Tranvinh, and C. Federau, "Realistic generation of diffusion-weighted magnetic resonance brain images with deep generative models," *Magn. Reson. Imag.*, vol. 81, pp. 60–66, Sep. 2021.

[6] *Facing Reality? Law Enforcement and the Challenge of Deepfakes*, Europol Innovation Lab, Luxembourg, 2022.

[7] I. Amerini et al., "Deepfake media forensics: Status and future challenges," *J. Imag.*, vol. 11, no. 3, p. 73, Feb. 2025.

[8] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[9] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.

[10] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Q. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, N. Messina, G. Amato, G. Perelli, S. Concas, C. Cuccu, G. Orrù, G. L. Marcialis, and S. Battiato, "The face deepfake detection challenge," *J. Imag.*, vol. 8, no. 10, p. 263, 2022.

[11] S. Concas, G. Perelli, G. L. Marcialis, and G. Puglisi, "Tensor-based deepfake detection in scaled and compressed images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3121–3125.

[12] A. K. Das, S. Mukhopadhyay, A. Dalui, R. Bhattacharya, and R. Naskar, "Fighting deepfakes by detecting DCT frequency anomalies," in *Proc. Int. Symp. Devices, Circuits Syst. (ISDCS)*, vol. 1, May 2023, pp. 1–5.

[13] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting GAN DCT anomalies," *J. Imag.*, vol. 7, no. 8, p. 128, Jul. 2021.

[14] O. Pontorno, L. Guarnera, and S. Battiato, "On the exploitation of DCT-traces in the generative-AI domain," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2024, pp. 3806–3812.

[15] L. Guarnera, O. Giudice, and S. Battiato, "Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 11, pp. 1–24, Nov. 2024.

[16] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2014.

[17] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7887–7896.

[18] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at Fourier spectrum discrepancies for CNN-generated images detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7196–7205.

[19] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 658–659.

[20] S. Hussain, P. Neekhara, B. Dolhansky, J. Bitton, C. C. Ferrer, J. McAuley, and F. Koushanfar, "Exposing vulnerabilities of deepfake detection systems with robust attacks," *Digit. Threats, Res. Pract.*, vol. 3, no. 3, pp. 1–23, Sep. 2022.

[21] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang, "Exploring frequency adversarial attacks for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4103–4112.

[22] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading DeepFake detectors via adversarial statistical consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12271–12280.

[23] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.

[24] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, and S. Battiato, "Representing scenes for real-time context classification on mobile devices," *Pattern Recognit.*, vol. 48, no. 4, pp. 1086–1100, Apr. 2015.

[25] S. Battiato, F. Guarnera, and G. Puglisi, "Exploiting AC histogram statistics for misalignment estimation in double JPEG compressed images," *IEEE Access*, vol. 12, pp. 64622–64632, 2024.

[26] S. Battiato, O. Giudice, F. Guarnera, and G. Puglisi, "First quantization estimation by a robust data exploitation strategy of DCT coefficients," *IEEE Access*, vol. 9, pp. 73110–73120, 2021.

[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[28] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[30] Y. Lu and T. Ebrahimi, "Towards the detection of AI-synthesized human face images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2024, pp. 3778–3784.

[31] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[32] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[35] L. Guarnera, F. Guarnera, A. Ortis, S. Battiato, and G. Puglisi, "Evasion attack on deepfake detection via DCT trace manipulation," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2025, pp. 157–169.

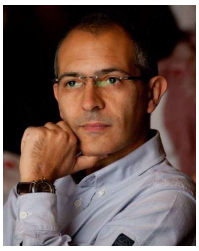
[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

FRANCESCO GUARNERA (Member, IEEE) received the bachelor's and master's degrees (summa cum laude) in computer science from the University of Catania, in 2009 and 2018, respectively, and the Ph.D. degree from the Department of Mathematics and Computer Science, University of Catania, in 2022. His Ph.D. thesis titled "Advanced Methods for Image Forensics: First Quantization Estimation and Document Authentication." From 2009 to 2016, he was a Developer/Analyst/Project Manager of web applications. In April 2022, he was a Research Fellow with the Department of Pharmaceutical Sciences, University of Catania studies with the grant titled "Modeling Using Agent-Based Modeling (ABM) Methodologies by In Silico Trials for Drug Evaluation" joining medical imaging research projects focused on the analysis of magnetic resonances of the brain. He is currently a Researcher with the Department of Mathematics and Computer Science, University of Catania. Since 2017, he has been a member of the IPLab Research Group, University of Catania. He has participated at the International Computer Vision Summer School (ICVSS) in the 2019 edition. Since 2022, he has been part of the committee of the International Forensics Summer School (IFOSS). He has co-authored more than 20 papers in international journals and conference proceedings. His research interests include image processing, multimedia forensics, medical imaging, and deep learning.

LUCA GUARNERA was born in Catania, in October 1992. He received the Ph.D. degree in computer science from the Department of Mathematics and Computer Science, University of Catania, in October 2021. His Ph.D. thesis titled "Discovering Fingerprints for Deepfake Detection and Multimedia-Enhanced Forensic Investigations.". Part of the Ph.D. research was carried out with the University of Hertfordshire, College Lane Campus, Hatfield, U.K., under the supervision of Prof. Salvatore Livatino, working on the creation of a software for forensic ballistics analysis and firearms comparison, using the Oculus Rift S headset. Since January 2022, he has been a Research Fellow in computer science with the University of Catania. Since 2022, he has been part of the committee of the International Forensics Summer School (IFOSS). His main research interests include computer vision, machine learning, and multimedia forensics and its related fields with a focus on the deepfake phenomenon.

ALESSANDRO ORTIS (Senior Member, IEEE) received the Ph.D. degree in mathematics and computer science from the University of Catania, in 2019. He became an Assistant Professor (in 2022) and an Associate Professor (in 2025) with the University of Catania. He is an Active Member of the Image Processing Laboratory (IPLAB), the Director of the Biometric Research Group (BioRG), and a member of the IEEE Biometrics Council. He supervises several Ph.D. students and postdoctoral researchers and coordinates various national and international research projects. He has co-authored over 50 conference papers and more than 25 journal articles. He is also a co-inventor of an international patent. His research interests include image and video understanding, adversarial machine learning, biometrics, and multimedia forensics. He serves as a reviewer and an editorial board member for several international journals and conferences.



SEBASTIANO BATTIATO (Senior Member, IEEE) received the degree in computer science from the University of Catania, in 1995, and the Ph.D. degree in computer science and applied mathematics from the University of Naples, in 1999. He has held various positions with the University of Catania, starting as an Assistant Professor, in 2004, then becoming an Associate Professor, in 2011, and a Full Professor, in 2016. Throughout his career, he has been involved in leading roles within academic programs and research initiatives, including chairing the Undergraduate Program in Computer Science and serving as the Rector's Delegate for Education. Currently, he serves as the Deputy Rector for Strategic Planning and Information Systems. He is highly active in research, directing the IPLab Research Laboratory, and participating in several national and international projects. He has supervised several Ph.D. students and postdoctoral researchers, edited several books, and authored over 100 papers in various publications. His research interests include computer vision, imaging technology, and multimedia forensics. He is the Director and the Co-Founder of the International Computer Vision Summer School (ICVSS) and the International Forensics Summer School (IFOSS). He is recognized for his contributions and received accolades, such as the 2017 PAMI Mark Everingham Prize and the 2011 Best Associate Editor Award from IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He also plays key roles in organizing international events and serves as an Associate Editor for *SPIE Journal of Electronic Imaging*.

GIOVANNI PUGLISI received the M.S. degree (summa cum laude) in computer science engineering and the Ph.D. degree in computer science from the University of Catania, Catania, Italy, in 2005 and 2009, respectively. From 2009 to 2014, he was with the University of Catania, as a Postdoctoral Researcher. In 2014, he joined the Department of Mathematics and Computer Science, University of Cagliari, as an Associate Professor. He has edited one book; and co-authored more than 50 papers in international journals, conference proceedings, and book chapters. He is a co-inventor of several patents and serves as a reviewer for different international journals and conferences. His research interests include image/video enhancement and processing, camera imaging technology, and multimedia forensics.