DeepEvader: An Evasion Tool for Exposing the Vulnerability of Deepfake Detectors Using Transferable Facial Distraction Blackbox Attack

Qurat Ul Ain[a], Ali Javed[b], Aun Irtaza[a]
[a] Computer Science Department, University of Engineering and Technology, Taxila, 47050, Pakistan
[b] Software Engineering Department, University of Engineering and Technology, Taxila, 47050, Pakistan

*Correspondence: ali.javed@uettaxila.edu.pk
Contributors: quratul.ain2@students.uettaxila.edu.pk, aun.irtaza@uettaxila.edu.pk

Abstract
Deepfakes are highly realistic fabricated media created using sophisticated generative techniques that are massively used to spread disinformation in cyberspace. Implementing a deepfake detector is crucial to identify and counter these threats, thereby ensuring the integrity of digital media. However, these detection systems are susceptible to adversarial attacks that exploit vulnerabilities to circumvent identification. The present study investigates the vulnerability of deepfake detectors to adversarial attacks. To develop an adversarial attack that is visually realistic, resilient, and demonstrates formidable attacking capabilities, we proposed a new Facial Distraction Black-Box Attack (FDB attack) framework based on biological vision. The proposed black-box attack is capable of successfully evading deepfake detectors without the need for access to the target detector's parameters or architectural specifications. It exhibits high transferability across a variety of deepfake detectors, including end-to-end deep learning, fused, and unified models, utilizing five standard datasets. The proposed attack beats the performance of the best detector and reduces the accuracy of the detectors significantly from 99.9% to 49.7%. Rigorous experimentation was performed to show the effectiveness of the proposed attack in comparison with the state-of-the-art attacks. In addition, we have developed a specialized penetration testing tool named DeepEvader to uncover and analyze the vulnerabilities of existing deepfake detectors systematically. By exposing weaknesses in current detection methodologies, our work highlights the urgent need for robust detection mechanisms to combat deceptive digital content effectively. Our research reveals flaws in existing detection methods, emphasizing the immediate requirement for strong and durable detection techniques to combat deepfakes successfully.

Keywords: *Adversarial attack, Black-box attack, Deepfakes detection, DeepEvader, Facial distraction black-box attack.*

1. Introduction

The rapid advancement in artificial intelligence, particularly in generative models like Generative Adversarial Networks (GANs) and autoencoders [1, 2], has led to the proliferation of deepfake technologies. These technologies can generate hyper-realistic audio and visual content posing significant threats to individuals and society by spreading disinformation [3]. Deepfake technologies have already been used to create convincing fake videos of public figures saying and doing things they never did [4]. As these technologies continue to improve, it will become increasingly difficult to discern between real and deepfake. The increasing sophistication of deepfakes has necessitated the development of advanced detection methods to mitigate their harmful impacts. Deepfake technologies can spread disinformation, but progress is being made in developing detection methods to counter their negative impact. It has become essential to focus on improving detection techniques rather than solely addressing the negative implications of these technologies.

Due to the growing complexity of deepfakes, there is a need for more advanced detection methods [5-9] to reduce their negative effects. Researchers have developed deepfake detectors to identify deepfakes, these technologies can be categorized into traditional machine learning (ML) models and deep learning (DL) approaches, each with its unique strengths and areas of application. Traditional ML models rely on hand-crafted features [10, 11] and statistical analysis [12, 13] to identify inconsistencies in videos and images that may indicate manipulation. However, DL approaches, such as Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) leverage the power of neural networks to learn and identify subtle cues that distinguish genuine content from deepfakes [5-9, 14]. These methods can automatically learn complex patterns and anomalies in data without the need for manual feature selection, making them particularly effective against the continuously improving quality of deepfakes.

Advancements in adversarial machine learning have made deepfake detectors vulnerable to attacks, which manipulate models to make inaccurate predictions, resulting in their ineffective performance. The adversarial attacks impact the reliability of deepfake detectors, potentially enabling harmful deepfakes to spread without control. There are two common types of attacks, white-box and, black-box, which are utilized to evaluate model performance and vulnerability. White-box attacks involve full knowledge of the model architecture and parameters [15-17], while black-box attacks only have access to the model's input-output behavior [16, 18-20]. In

real-life scenarios, black-box attacks are applicable when the attacker has limited information about the model they are trying to deceive. This makes black-box attacks more challenging to defend against, as they mimic real-world conditions where attackers may not have complete access to the target system. However, existing approaches are not as effective due to knowledge gaps and limitations in transferring adversarial examples among different models [18]. Improving defenses against black-box attacks requires innovative strategies that can adapt to varying levels of information access.

Black-box attacks are crucial in deepfake detection due to their lack of transferability [18, 21] across different models. These attacks simulate real-world scenarios without requiring access to the inner workings of the model they aim to deceive. However, existing black-box attacks often struggle to generalize across different detection systems, resulting in attacks that are effective against one model but ineffective against another [15]. To expose the vulnerability of deepfake detectors, there is a need for an adversarial attack that is robust, transferable, looks natural, and cannot be distinguished from real content. In this research, we proposed a black-box attack methodology that is both robust and transferable, while ensuring the generated content remains natural-looking, and represents a significant advancement in adversarial machine learning. The proposed technique creates adversarial examples that are effective at evading detection and can be applied across a range of models without losing effectiveness.

Current detection methods rely on detecting real and deepfake-generated images by analyzing features such as inconsistencies in facial expressions, lighting, and shadows, algorithms can identify patterns that are indicative of manipulation [22]. To expose the vulnerability of existing deepfake detectors, we proposed imperceptible adversarial perturbations to the test data, causing the deepfake detectors to misclassify it. By carefully crafting these perturbations, we deceive the algorithm into incorrectly identifying a deepfake image as real. In addition, we propose a penetration testing tool that can simulate various adversarial attacks to evaluate the effectiveness and expose the vulnerability of deepfake detection algorithms. This highlights the ongoing challenge of developing robust detection methods that can withstand sophisticated adversarial attacks. The following are major contributions of this research:

- This work presents a novel Facial Distraction Black-Box Attack framework inspired by biological vision to develop an adversarial attack that is visually natural, resilient, transferable across different deepfakes detectors, and demonstrates formidable attacking capabilities.
- Our proposed black-box attack can successfully evade deepfake detectors without requiring access to the target detector's parameters or architectural specifications.
- We employ attention distraction and lateral inhibition mechanisms, inspired by biological observations, to present an optimized attack strategy, which redirects the model's focus toward unintended areas.
- We employ degradation methods to reduce statistical disparities, including illumination adjustment, brightness values in localized pixels, noise addition, and uniform Gaussian blurring.
- We also propose a novel penetration testing tool to test the vulnerability of deepfake detectors on proposed and state-of-the-art adversarial black-box attacks.
- Rigorous experiments were performed including quantitative, qualitative, explainability, transferability, and comparative analysis with existing adversarial attacks on different deepfake detectors i.e., end-to-end, fused, and unified models using five benchmark datasets.

2. Literature review

This section examines the current state of deepfake detection technologies and the adversarial attacks that have been attempted to compromise the security of deepfake detectors. The following are some current techniques that have mainly focused on the detection of deepfakes, neglecting to account for vulnerability to adversarial attacks. For the detection of deepfakes, various handcrafted and local feature extractors have been utilized.

2.1 Deepfake detectors

For the detection of deepfakes, a hierarchical feature selection (HFS) method was introduced in [23]. From the inputs, both handcrafted and deep learning features were extracted, and HFS was utilized to select the nearest set of features. While the method exhibited strong performance on the same data, its accuracy was significantly diminished when applied to cross-dataset scenarios. Kolagati et al. [24] utilized a convolutional neural network and a multilayer perceptron to learn and predict features for deepfake video classification by transmitting facial landmark features to them. This model cannot detect fake videos reliably in low-light conditions or with multiple faces. Current algorithms designed to detect deepfakes using a single image rely heavily on a variety of deep-learning techniques. CNN is compared to naive approaches because it learns to distinguish between genuine and fake data. In this category, Xception [25] and MesoNet [26] are two of the most widely used. Previous methodologies have utilized a range of deep learning-based strategies to identify deepfakes. Cao et al. introduce RECCE [27], an autoencoder that acquires compact latent representations of actual faces. As a result, deepfakes

are categorized as out-of-distribution samples characterized by increased reconstruction error. Liang et al. [28] utilized U-net to forecast the depth map of an image, which was subsequently employed to guide a transformer-based triplet feature extraction network. Instances involving extreme illumination and face occlusion render the method inapplicable. Ilyas et al. [5] introduced a new and efficient capsule network, referred to as E-Cap Net, designed specifically for detecting forgeries. The utilization of capsule network design enhances the identification of deepfake images and videos, however, this approach is computationally more intricate. Khalid et al. [7] introduced a combined truncated DenseNet121 model for identifying deepfakes via transfer learning, truncation, and feature fusion. The model effectively identifies deepfakes in a wide range of datasets. A cross-modality attention-based deepfake detector was introduced in [29]. Nawaz et al. [9] proposed ResNet-Swish-Dense54 for reliable deepfake detection. The model extracts and analyses human faces from video frames, distinguishing between real and manipulated content.

To validate that the CNN has accurately identified distinguishing characteristics, explicit spatial modeling of particular deepfake artifacts was employed in [30], [31]. Nguyen et al. [31] proposed a method called the capsule network that takes advantage of the spatial and hierarchical relationships among image components. A feature extraction module employing learnable high-pass filters and Gabor convolutions; a shallow texture module enhancing texture and high-frequency features; and a cross-modality attention module enhancing feature learning and fusion comprise the three sections of the architecture. The computational expense of this approach is substantial because of its intricate architecture. A 5-layered 3DCNN for detecting facial manipulations such as FaceSwap, Face2Face, and Deepfakes was introduced in [32]. However, the generalization capabilities of this model were not assessed, which is a complex issue in deepfake detection. A vision transformer and distillation method were employed in [33] to distinguish between forged and authentic videos. Despite the method's promising outcomes, its efficacy and practicality are constrained by an absence of reliability. In our prior study [14], we introduced DFGNN, a comprehensible and versatile GNN designed for the identification of deepfake content. The method utilizes facial landmarks to generate a graph, enhancing the capacity to analyze and apply the results to various scenarios. However, sophisticated deepfakes could potentially impact the performance of DFGNN. Further investigation is required to assess its efficacy in detecting intricate deepfakes. The performance of DFGNN may be impacted by structurally regular advanced deepfakes.

There is a dire need for a more resilient and widely applicable approach to identifying intricate artificial content. Yuyang et al. [34] redirected their attention toward frequency space image analysis, acknowledging that authentic and fabricated data generally possess indistinguishable frequency spectrums. By utilizing frequency-aware image decomposition and local frequency statistics, they develop F3-Net. Ilyas et al. [6] suggested using a combination of Swish and ReLU activation functions to enhance the ability of the Efficient-Net architecture to detect deepfakes by improving its representation capabilities. The model demonstrated satisfactory performance in detecting deepfakes. However, it does not include a thorough analysis of how well the model can adapt to detecting deepfakes that it has not been trained on. Liu et al. [35] proposed that the phase spectrum of natural images retains a multitude of frequency components in contrast to manipulated images. By integrating spatial indicators with this information, their Spatial-phase shallow learning (SPSL) method enhances the effectiveness of deepfake detection. Luo et al. [36] suggested an additional frequency-based method, which used residual-guided spatial attention module-extracted low-level RGB features and high-frequency image disturbances.

2.2 Adversarial Attacks on Deepfake Detectors

The identification of deepfakes is not just a significant concern, but also a serious security issue that demands immediate attention. Adversaries are constantly modifying their approaches to evade current deepfake detectors, emphasizing the need to remain one step ahead of their strategies. Unfortunately, the efficacy of existing modern deepfake detection techniques in opposing scenarios has not been investigated, resulting in a deficiency in security standards. A method called a double-masked guided attack [37] was proposed to deliberately deceive the deepfake detector into identifying GAN-generated fake faces. This method involves introducing perturbations to the crucial facial areas that are often focused on by deepfake detectors during the detection of fake photos. The attack was executed on nine forensic classifiers using both white-box and black-box methodologies. However, there is a need to enhance the transferability and resilience of the adversarial cases. In addition to perturbations in the picture space, Carlini et al. [15] also explore the application of perturbations in the latent space of the generative model, resulting in the creation of adversarial images. In contrast, Liu et al. [38] engage in blind post-processing of pre-generated deepfakes, eliminating identifiable traces that are left behind by the deepfake development process. The resulting deepfakes are thus more real and difficult to identify. Huang et al. [39] have created FakePolisher, a shallow dictionary model that is specifically trained to effectively eliminate common GAN artifacts by accurately reconstructing only actual data.

Several researchers performed white and black-box attacks, Gandhi et al. [17] conducted adversarial attacks, specifically FGSM and C&W, on VGG16 and ResNet18 models. These attacks were carried out using synthetic

images generated by Few-Shot Face Translation GAN [40]. Attacks were executed in both the black-box and white-box scenarios. The success rate of white-box assaults was 100% in all scenarios, except for the FGSM attack on the ResNet18 model. However, for black-box attacks, the success rate decreases dramatically. Shahriyar et al. [41] demonstrated the efficacy of FGSM and C&W assaults on a deepfakes detector that relies on sequence-based analysis. The experiments were conducted in both white-box and black-box scenarios. To achieve this objective, the victim models, Conv-LSTM [42] and FacenetLSTM [43] were utilized using undisturbed photos from the FF++ dataset. The white-box approach designed for one model was employed as a black-box attack for the other model, and vice versa. In the white-box configuration, this adversarial assault diminishes the performance of detectors, however, in the black-box attack setting, the success rate of the attack is greatly reduced.

Neekhara et al. [44] assess several methods of disruption in black-box scenarios. The study showcases the consistent ability of their generated adversarial samples to attack various deepfake detection methods. Lim et al. [45] conducted a black-box attack by putting cosmetic artifacts (eyeliner, blush, lipstick) on specific areas of facial landmarks, which led to distorted photographs. The assault resulted in a 50% reduction in the precision of the victim models, specifically MesoInception-4 and TwoStreamNet. However, this approach is less effective when compared to other conventional attacks such as PGD [46] and FGSM [47]. Lou et al. [48] have presented a new form of black-box attacks that target the reduction of GAN fingerprints. These fingerprints are frequently utilized as indicators in the process of detecting deepfakes. To accomplish this, they employ a training method that involves an autoencoder. This autoencoder is designed to generate images of excellent quality while also applying subtle changes to individual pixels that are difficult to detect. The differential evolution one-pixel assault [21] and the simulated annealing one-pixel attack [18] are two separate forms of black-box attacks commonly employed in image classification tasks to create adversarial instances. Su et al. [21] proposed a technique called differential evolution one-pixel assault, which involves iteratively altering the pixel values of a picture to generate an adversarial example. The CIFAR-10 dataset was subjected to an assault, specifically targeting images with a dimension of 32×32 pixels. This attack has the effect of reducing the performance of images with larger dimensions. Zhou et al. [18] introduced a simulated annealing one-pixel attack, which is an optimization approach that systematically alters the pixel values of an image to identify an adversarial instance. Both techniques [18, 21] generate adversarial samples through repetitive processes, resulting in computational complexity. Additionally, the effectiveness of these attacks diminishes when applied to high-dimensional images.

The existing research contains examples of adversarial instances that were sent to detectors, which made the identification of deepfakes challenging. However, these methods do not possess the transferability property of adversarial instances inside a single setting. This is because an attacker creates an adversarial perturbation in a white-box setting and then transfers it to a black-box setting within the same setting. We offer a simple attack in comparison to existing attacks and are transferable based on a model attention distraction phenomenon. This attack is visually natural and transferable across several deepfake detectors, allowing them to fail their performance even in a black-box situation. Our method demonstrates the vulnerability of deepfake detectors to adversarial attacks, highlighting the need for robust defense in this rapidly evolving field. By exploiting model attention distraction, we show how easily deepfake detectors can be deceived, raising concerns about the reliability of current detection systems.

3. Methodology of Proposed Attack Framework

Several deepfake detection methods [49, 50] highlight statistical discrepancies that exist between real and deepfake visual content. The variations in brightness and statistics observed in GAN-generated images make them different in contrast to natural images. Some GANs might exhibit limitations when it comes to producing images that encompass a wide spectrum of intensity values, leading to an absence of saturated and underexposed areas. This demonstrates that GANs invariably introduce high frequency into manipulated images. These differentiations direct our attention toward integrating corresponding adversarial degradations. Compared to the existing adversarial attacks [18, 21, 51, 52], adding perturbation doesn't look natural, and several attacks lack transferability. In this section, we first present the problem definition and then elaborate on the proposed attack overview.

3.1 Problem definition

Suppose the deepfake detectors, based on deep neural networks D_θ are trained on original instances I_{original} (real and deepfake) with the labels l. The black box attack involves an optimization task, where the objective is to minimize the loss function $L(I_{original}, D_\theta)$, while adhering to the constraint $\|\mathrm{P}\| \leq \varepsilon$. In this context, L represents the deviation between the actual label and the predicted label, and ε signifies a minor positive constant that restricts the magnitude of perturbations. The aim is to identify the perturbation P which reduces detection accuracy to the greatest extent possible while remaining within the perturbation constraint as follows:

$$\text{minimize } L\left(I_{original}, D_\theta\right) \tag{1}$$

$$I_{adv} = I_{original} + P \quad s.t. \ \|\mathrm{P}\| \leq \varepsilon \tag{2}$$

The test set adversarial instances I_{adv} can expose the vulnerabilities of D_θ. In addition, as an adversary, we aim to add the perturbation to target instances which ensures the visual natural ε for the adversarial instances I_{adv}. The adversarial perturbation can lead the model to make a false prediction as follows:

$$D_\theta\left(I_{adv}\right) \neq l \quad s.t. \ \| I_{adv} - I_{original} \| \ < \varepsilon \tag{3}$$

The proposed adversarial attacks emphasize the development of a natural-looking adversarial attack that deceives the authenticity and exposes the vulnerability of deepfake detection systems.

3.2 Method overview

In this work, we present a novel framework inspired by biological vision's mechanisms [51, 53] and principles to develop an adversarial attack that is visually natural, capable of transferring across models with resilience, and demonstrates formidable attacking capabilities. Even if the attacker lacks access to the target detector's parameters and architectural specifications, the proposed method can still evade the detectors by utilizing both human and model attention. Initially, heatmaps were generated using the surrogate model following the attention distraction method [53] and the lateral inhibition mechanism [51] to assess the explainability factor [54] and examine the principal characteristics of the facial frames focused by the trained model. To optimize the applicability of our attack, we incorporate insights from biological observations and deliberately shift the focus of the model from the intended targets to unintended regions, including the background. Concerning visual naturalness, we aim to circumvent the bottom-up attention that is unique to human vision by developing an adversarial attack that appears visually natural. A mask is created to accurately represent the precise region of the facial frame that is focused by the detector. We employ a set of facial degradation methods in the proposed attack to effectively reduce statistical disparities. We initially employ the illumination adjustment and extract the mask of the frame landmark area. After that, we adjust the brightness values in localized pixels of masked areas. The addition of noise demonstrates that the introduction of random noise into fabricated images effectively reduces regular artifacts. Lastly, uniform Gaussian blurring blurs statistical differences in the frequency domain by removing high-frequency components from images. The details are provided in subsequent sections. Figure 1 illustrates the methodology, and the following sections provide further elaboration on the proposed approach.

3.3 Surrogate model creation

A surrogate model S_θ resembles the target model and is intended to exploit or manipulate system vulnerabilities to comprehend the behavior of the target model. Additionally, the concept of lateral inhibition [55] provides an intrinsic neural process identified in biological systems, including inhibition of the activity of adjacent neurons, thereby augmenting the ability to differentiate between contrast and features. These abilities are inherent in all deep learning models. By drawing inspiration from the principles of lateral inhibition, we create surrogate models to make proposed adversarial attacks transferable across different deep-learning models. To simultaneously strengthen the resilience of adversarial perturbations and implement attention distraction, we leverage the concept of lateral inhibition [55] in our attack strategies.

Deep neural networks' perception is crucial for recognition and classification tasks. Surrogate model lateral inhibition methods allow us to attack model attention. Simultaneously, this process restricts the activation of the ground-truth class, resulting in an improved ability to make effective attacks. Thus, the model must focus on the wrong class, which affects its attention to the true label. Based on the above description, we chose the VGG-19 architecture [56], a pre-trained model that is computationally efficient for many detection tasks. The target model is any deepfake detector that verifies video authenticity. Our goal as an adversary is to modify the test set videos to reject the target model's label. We want the target to mistake a deepfake video for a real one, or vice versa. We employ a surrogate model to approximate the behavior of target models. The surrogate model misclassified data after testing it with proposed attacked test set samples. Based on surrogate model findings, we visualize [57] thetraits of facial frames focused by a surrogate model. This provides us with the information to add the proposed adversarial attack to these ROIs.

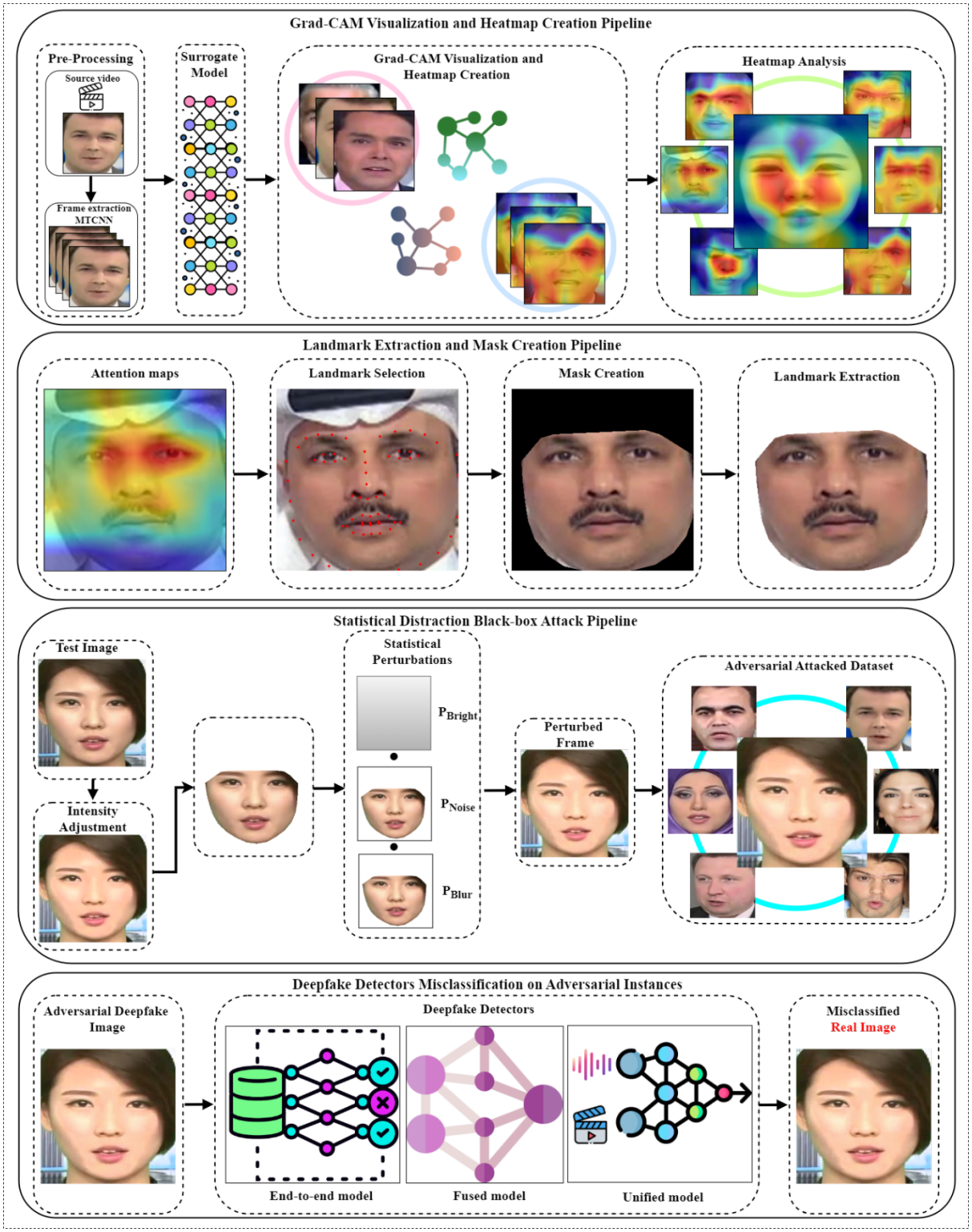

Figure 1: Framework of proposed method. To emphasize the attention mechanism, a surrogate model is created; heatmaps are extracted using Grad-Cam visualization, and the landmark region is eventually extracted to generate the mask. In addition, we generate a visually natural adversarial attack to bypass the human-specific visual attention mechanism. Finally, we assess the transferability of the proposed attack across various deep learning models, such as unified models, fused models, and end-to-end models.

3.4 Grad-cam Visualization

In biology, there are noteworthy observations regarding selected attention features and their impact on cerebral activities. These observations have revealed that these attention features can stimulate similar patterns of cerebral

activity across different individuals [53]. This finding suggests the presence of comparable characteristics in terms of neuron hyperperception among these individuals. In this research, we explored the potential similarities in attention patterns exhibited by deep neural networks when predicting the same objects. The implementation of artificial neural networks is influenced by the functioning of the human central nervous system, which forms the basis for this assumption. Our objective is to capture attention structures that are independent of the deep learning models to improve the practicality of adversarial attacks.

Visual attention techniques, including CAM, Grad-CAM, and Grad-CAM++ [57], have been the subject of extensive research to gain insights and comprehension into the behaviors of deep learning models. In predictive modeling, it is a common practice for models to direct their attention toward specific target objects. To effectively execute an attack on a model, it is imperative to employ a strategy that diverts the model's attention away from significant objects. This can be observed by manipulating the attention map shared by the model on the salient area. Figure 2 shows heat maps generated from our surrogate model. Visual assessment shows the surrogate model's focus on facial landmark areas including foreheads, brows, nose, and eyes. The red-highlighted portion of the facial frame is the area of interest. This investigation shows that perturbing only the highlighted region of facial frames reduces the computational complexity of the attack. Additionally, this perturbation bypasses the deepfake detector's intended performance.

3.5 Proposed Facial Distraction Blackbox Attack

As we have already discussed in the method overview for the proposed Facial Distraction Blackbox (FDB) Attack, we employ different statistical degradation methods that effectively reduce statistical disparities from generated images. In the first step, we adjust image intensity, then extract facial ROI via masking and apply some other techniques to the masked region. We design three distinct stages in our approach to improve the effectiveness and covert nature of proposed adversarial perturbations. In the masked region, we adjust channel-wise brightness and add uniform noise and Gaussian blur. We carefully construct adversarial degradations to covertly manipulate visual signals with the intention of misleading image authenticity. The visual presentation of the proposed attack pipeline is provided in Figure 1 and the details of all the degradations are as follows:

3.5.1 Adversarial illumination adjustment

Specific visual cues identify most of the generated images as computer-generated. Those cues encompass a range of visual characteristics, including the presence of unrealistic lighting and shading effects and a noticeable absence of fine details. To make these generated images look realistic, we adjust the intensity, which refines the brightness of the generated deepfake images. We achieve this by manipulating their intensity within the HSV color space. This technique isolates intensity information while suppressing color details to enhance the realism of deepfake outputs.

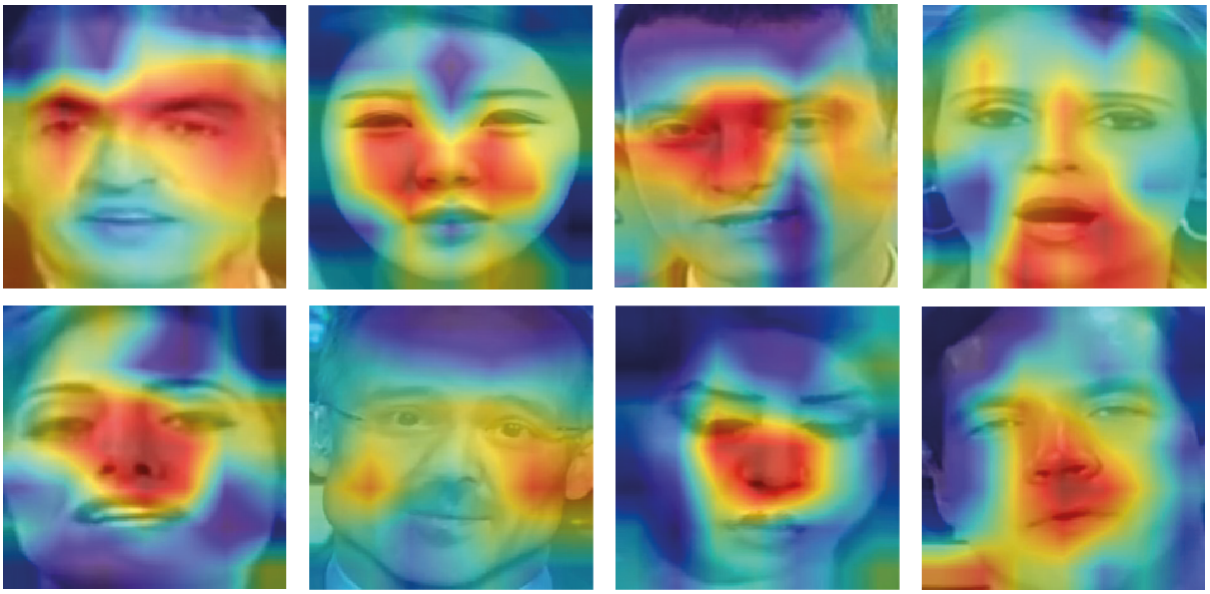

Figure 2: Heatmap visualization generated through the Grad-Cam visualization technique.

To adjust the intensity, we first convert the default BGR color channel to the hue, saturation, and intensity (HSI) color space. This color space allows the separation of intensity data from color tones. Incorporating the given description, the conversion from the original BGR channel I_{BGR} to the HSI channel I_{HSI} can be denoted as follows:

$$I_{HSI} = (I_{BGR(original)}) \tag{4}$$

Following this, an adjustment is made to the intensity channel (I_{HSI_adj}) by increasing image intensity. By multiplying intensity channel pixel values by the intensity factor, the image will appear brighter, improving detail visibility and visual impact. By utilizing the clip function, the adjusted intensity values remain within the allowed range of [0, 255]. We can mathematically represent the adjustment as:

$$I_{HSI\_adj} = clip(I_{HSI} \times intensity\_factor) \tag{5}$$

In the last step, the adjusted intensity channel is transformed back into the BGR color space, resulting in the generation of the modified image denoted as I_{BGR_adj}. This conversion allows for the restoration of the modified image to its original color representation, ensuring that the visual information is preserved and represented as:

$$I_{BGR\_adj} = (I_{HSI_{adj}}) \tag{6}$$

3.5.2 Mask Generation

Extending the concepts of attention distraction and heat map analysis, the creation of a facial landmark area mask M_{xy} introduces a novel approach to perturb specific regions within an image. By creating a mask, we can manipulate the region of interest by disrupting specific facial characteristics. This method determines critical facial landmark areas, enabling precise disruption of the image. Utilizing the mask enables one to manipulate visual attention and assess the influence of particular regions on overall perception in a more precise manner. Facial landmarks are identified for every intensity-adjusted frame through the utilization of a landmark's predictor. This results in the generation of a set of (x, y) coordinates that collectively represent 68 distinct facial landmarks. With these coordinates, we draw a convex hull H_{xy}, a polygon that includes all the landmark features of the face. The convex hull enables precise capture of the facial landmarks and boundaries, ensuring the adjustment process exclusively incorporates the intended landmarks. This process involves iterating through each of the 68 facial landmarks, extracting their (x, y) coordinates, and generating the convex hull. Let $\{(x_i, y_i)\}_{i=1}^{68}$ denote facial landmarks, so the convex hull is:

$$Convex\ Hull = H_{xy}(\{(x_i, y_i)\}_{i=1}^{68}) \tag{7}$$

After deriving the convex hull, we construct a binary mask. The mask effectively highlights the facial landmark area by filling the region within the hull with white pixels (255). This results in the creation of a binary representation of the identified region.

$$M_{xy}(H_{xy}) = \begin{cases} 255, & H_{xy} \\ 0, & otherwise \end{cases} \tag{8}$$

By applying the facial landmark area mask that is obtained, the original image is subsequently modified in a way that permits precise perturbation of the desired region. Mathematically, the process of applying the mask M_{xy} to the intensity-adjusted image I_{BGR_adj} is denoted as follows:

$$I_{masked}(M_{xy}) = I_{BGR\_adj} \,\&\, M_{xy}(H_{xy}) \tag{9}$$

Finally, a masked facial frame $I_{masked}(M_{xy})$ is generated through the "bitwise AND" of the intensity-adjusted image I_{BGR_adj} and its mask $M_{xy}(H_{xy})$. The utilization of the mask allows for the diversion of the detector's focus toward the essential facial characteristics, such as the mouth, eyes, and nostrils. As a result, the attack execution becomes more effective, and the computational load is reduced. The mask generation pipeline is also provided in Figure 1.

3.5.3 Adversarial Brightness

In the next step, we slightly enhance the brightness in this masked area, the facial landmarks, to make them more visible and more apparent. This adjustment helps to highlight the eyes, nose, and mouth, allowing for better recognition and interpretation of the facial expressions. The method accesses specific pixels in the region with the facial landmark, augmenting each color channel with a constant value and verifying that the resultant values fall

within a suitable range of brightness. This is achieved through the addition of a brightness factor to each color channel and the mathematical expression representing this operation is:

$$P_{Bright} = M_{xy}([b_B, b_G, b_R])) \tag{10}$$

Here, M_{xy} and the constant array $[b_B, b_G, b_R]$ represent the original pixel at coordinates (x, y). The brightness factor b is added to each corresponding color channel (red, green, and blue). The degree of brightness enhancement can be precisely adjusted by modifying the constant values, providing further control to the overall appearance of adversarial bright images.

3.5.4 Adversarial Uniform Noise

After adversarial brightness addition, we add uniform noise to each color channel of a masked image that occurs within a specified range. The procedure includes the independent generation of random noise values for the red, green, and blue channels. Adding small adversarial uniform noise to each color channel results in a realistic and natural appearance of the image. By carefully controlling the range of the uniform noise, we can adjust the level of distortion and achieve the desired balance between preserving image details and introducing a realistic level of noise.

$$P_{Noise} = (M_{xy}[c] + noise[c_B, c_G, c_R]) \tag{11}$$

The noise is added to the color channel $[c_B, c_G, c_R]$ of M_{xy} at coordinates (x, y). After adding the generated noise, the resulting intensity is reduced to ensure that it remains within the acceptable range of [0, 255]. The noise added to each channel c introduced slight variations in the mask texture and gave the image a more natural look. Simulating the existence of noise in the region containing facial landmarks enhances the overall visual realism and dynamism of the image.

3.5.5 Adversarial Blur

In the last step of the proposed attack, we add Gaussian blurring. Employing this technique eliminates the high-frequency components present in manufactured images, reducing the statistical difference in the frequency domain. To introduce Gaussian distortion to a masked image M_{xy}, the image is convolved with a Gaussian kernel represented as:

$$P_{Blurred} = \sum_{i=0}^{255} x \sum_{j=0}^{255} y \; M_{xy}(x+i, y+j) * G(i,j) \tag{12}$$

The outcome of convolving the functions $M_{xy}(x+i, y+j)$ represents the pixel intensity of the image at coordinates (x, y), while $G(i,j)$ denotes the Gaussian kernel positioned at the given offset in the context of image processing. The Gaussian kernel $G(i,j)$ is defined as follows:

$$G(i,j) = \frac{1}{2\pi\alpha^2} e^{-\frac{m^2+n^2}{\alpha^2}} \tag{13}$$

Where α represents the base of the natural logarithm and denotes the standard deviation of the Gaussian distribution. The Gaussian kernel convolution operation results in a noticeable opacification of the image as standard deviations increase. The adversarial Gaussian kernel blurs the fabricated image.

We sequentially employ these three perturbations to mask landmark image Equations (8), (9), and (10) and provide the following overview of the perturbations:

$$I_{perturbed} = P_{Blurred}(P_{Noise}(P_{Bright}(M_{xy}))) \tag{14}$$

The perturbed frame I_{adv} is formed through "bitwise AND" of the perturbed frame masked $I_{perturbed}$ with the bright-adjusted frame I_{BGR_adj} mentioned as:

$$I_{adv} = (I_{BGR\_adj}) \,\&\, (I_{perturbed}) \tag{15}$$

Figure 3: The Top row represents deepfake images from the dataset, bottom row represents perturbed frames by the proposed attack.

To assess the performance of the proposed attack, we provided the victim models with a test set of attacked facial frames. The original and attacked samples can be analyzed in Figure 3. As it can be observed the proposed perturbed frames look visually natural. The modified frames maintain the main facial features and expressions of the original pictures, posing difficulty for human observers to detect any abnormalities. The visual similarity highlights the effectiveness of the attack in evading detection by both human observers and deepfake detectors. The flow of the proposed method is described in Algorithm 1.

Algorithm 1: Proposed Facial Distraction Blackbox Attack.

Input: Input Video, Surrogate Model, Target Model.
Output: Attack creation, Target models misclassification (Acc drop), Attack Success rate.

1. **Initializing**
2. Selection of target model $\boldsymbol{D_\theta}$ and creation of a surrogate model $\boldsymbol{S_\theta}$:
3. $\boldsymbol{S_\theta}$
4. Train$(\boldsymbol{I_{original}})$ // train instance with $\boldsymbol{S_\theta}$
5. Heat map generation
6. Analyzing heatmaps on landmark ROI
7. Proposed Facial Distraction Blackbox Attack
8. $\boldsymbol{I_{HSI}} \leftarrow (\boldsymbol{I_{BGR(original)}})$ //each facial frame color space conversion for illumination adjustment
9. $\boldsymbol{I_{HSI\_adj}} \leftarrow (\boldsymbol{I_{HSI}} \times \boldsymbol{intensity\_factor})$
10. $\boldsymbol{I_{BGR\_adj}} \leftarrow (\boldsymbol{I_{HSI_{adj}}})$ //color conversion to BGR
11. **For**
12. $\boldsymbol{M_{xy}}$ for each $\boldsymbol{I_{BGR\_adj}}$, // mask creation
13. $Convex\ Hull \leftarrow \boldsymbol{H_{xy}}$ //convex hull
14. $\boldsymbol{M_{xy}}(\boldsymbol{H_{xy}}) \leftarrow 255$ // $area\ inside\ H_{xy}$
15. $\boldsymbol{I_{masked}}(\boldsymbol{M_{xy}}) \leftarrow \boldsymbol{I_{BGR_{adj}}}$ & $\boldsymbol{M_{xy}}(\boldsymbol{H_{xy}})$
16. $\boldsymbol{P_{Bright}} \leftarrow \boldsymbol{M_{xy}}([\boldsymbol{b_B}, \boldsymbol{b_G}, \boldsymbol{b_R}])$ // adversarial brightness perturbation in M_{xy}
17. $\boldsymbol{b}$ // brightness factor
18. $\boldsymbol{P_{Noise}} \leftarrow (\boldsymbol{M_{xy}}[\boldsymbol{c}] + \boldsymbol{noise}[\boldsymbol{c_B}, \boldsymbol{c_G}, \boldsymbol{c_R}]$ // adversarial noise perturbation in M_{xy}
19. $\boldsymbol{noise}[\boldsymbol{c_B}, \boldsymbol{c_G}, \boldsymbol{c_R}]$ // noise addition to each color channel
20. $\boldsymbol{P_{Blurred}} \leftarrow \boldsymbol{M_{xy}} * \boldsymbol{G}(\boldsymbol{i}, \boldsymbol{j})$ // adversarial blurriness perturbation in M_{xy}
21. $\boldsymbol{G}(\boldsymbol{i}, \boldsymbol{j})$ // Gaussian kernel
22. $\boldsymbol{I_{perturbed}} \leftarrow \boldsymbol{P_{Blurred}}(\boldsymbol{P_{Noise}}(\boldsymbol{P_{Bright}}(\boldsymbol{M_{xy}})))$ // perturbed mask creation
23. $\boldsymbol{I_{adv}} \leftarrow (\boldsymbol{I_{BGR\_adj}})$ & $(\boldsymbol{I_{perturbed}})$ // Attacked frame creation through proposed attack
24. Test set $(\boldsymbol{I_{adv}}) \leftarrow$ target model $\boldsymbol{D_\theta}$ //target model testing using attack instances
25. **Output**←misclassification of $\boldsymbol{D_\theta}$.
26. **End**

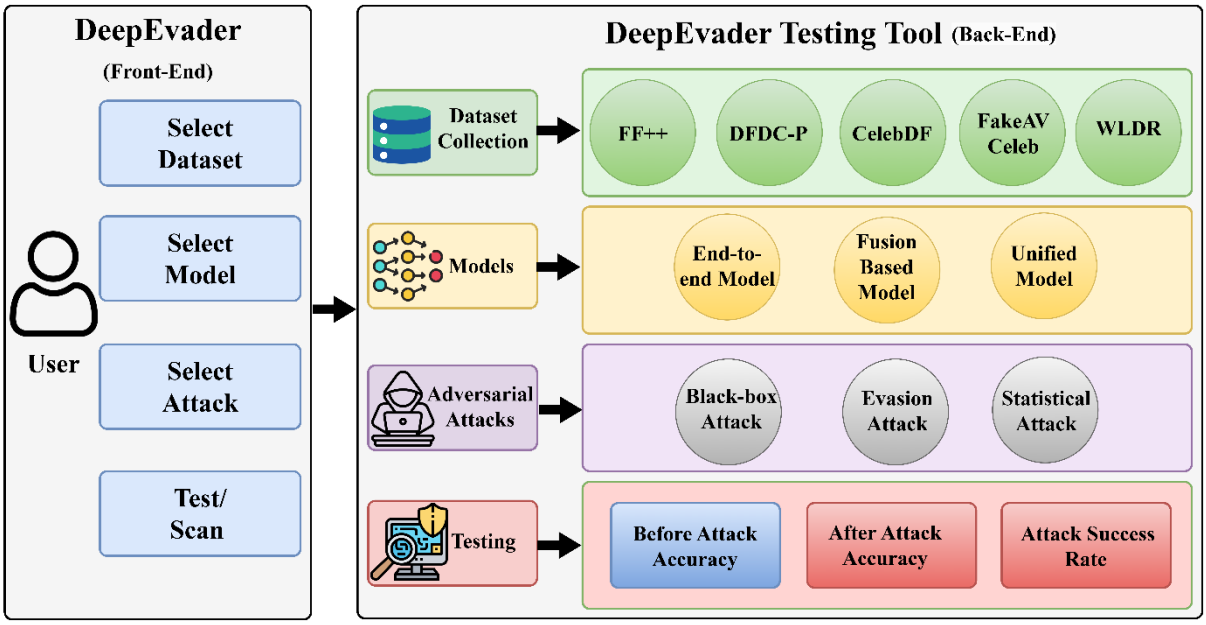

Figure 4: DeepEvader Flow Diagram.

4. DeepEvader Testing Tool

To detect the vulnerabilities of existing deepfake detectors, we proposed a novel penetration testing tool named DeepEvader. It uncovers the vulnerabilities of deepfake detectors on proposed and state-of-the-art adversarial black-box attacks. This emerges as an indispensable tool for pinpointing weaknesses within deepfake detectors, thereby evaluating the robustness and vulnerability of detectors against diverse manipulation strategies. This part explains the underlying concept of DeepEvader (Section 4.1), as well as the components (Section 4.2) and workings (Section 4.3) of the proposed system. Figure 4 displays the testing tool's flow diagram. Additionally, the evaluation results of DeepEvader on various deepfake detection models showcase effectiveness in identifying vulnerabilities. The tool also includes a user-friendly interface for easy navigation and utilization by researchers and practitioners in the field.

4.1 Concept of DeepEvader

A DeepEvader, acting as a black-box adversary, can carry out black-box adversarial attacks on deepfake detectors to undermine their integrity. This tool performs the following tasks: (i) allows the user to select different models and test them with perturbed and unperturbed samples, (ii) the test set data is changed to create adversarial outputs; and (iii) based on the attack, it generates actual outputs. This tool provides a platform to evaluate the resilience of detectors by simulating adversarial scenarios. DeepEvader and various deepfake detection models showcase their effectiveness in identifying vulnerabilities. The tool also includes a user-friendly interface for easy navigation and utilization by researchers and practitioners in the field. The interface of the developed testing tool is provided in Figure 5.

4.2 Components of DeepEvader

The DeepEvader framework is composed of several critical components, each of which is essential for the evaluation of the security and robustness of machine learning models. These components collaborate to simulate and analyze potential adversarial scenarios.

4.2.1 Model Selection

For now, we use three different varieties of deepfake detectors, including deep learning [5] fusion-based [7], [6] and unified deepfake detectors [8]. Section 5.2 provides the details of all models.

4.2.2 Dataset Selection

In the proposed application, we provide several datasets, including FF++, DFDC, Celeb-DF, FakeAVCeleb, and WLDR. We also offer attacked versions of all datasets, including existing and proposed black box attacks.

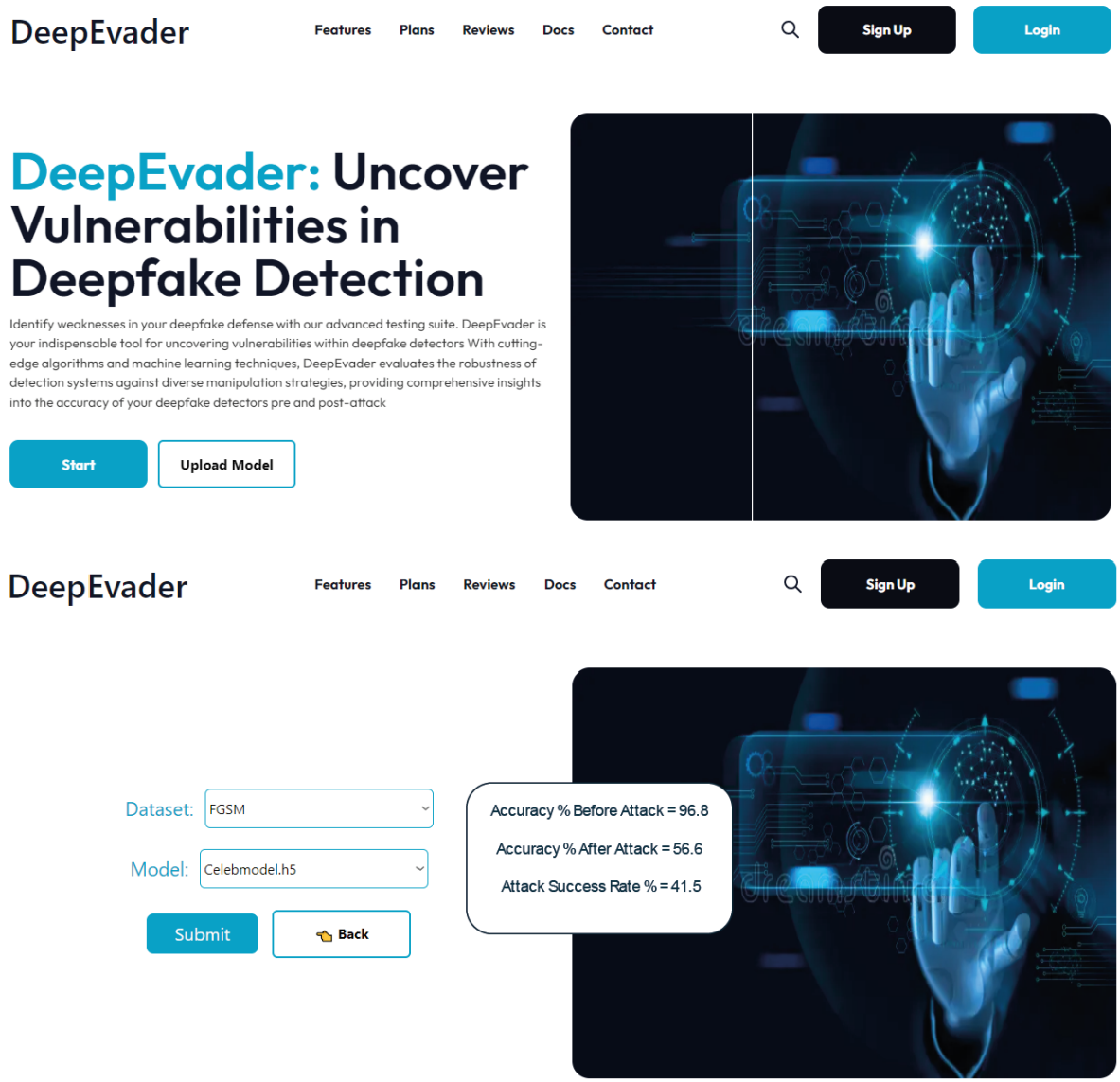

Figure 5: Proposed penetration testing tool "DeepEvader" Interface.

4.2.3 Attack Selection

The application includes various attack methods to test the resilience and expose the vulnerability of deepfake detection models. These attacks include Blur, Noise, Pix-DE, Pix-SA, Mole, FGSM, PGD, and the proposed attack.

4.3 How DeepEvader Works

DeepEvader utilizes a range of attack methods and its own proposed attack to expose the vulnerability of deepfake detection models. After selecting the model, dataset, and attack method, the user can initiate the test or scan process. The application will run the chosen deepfake detection model on the dataset, applying the provided blackbox attacks to assess the model's performance. The application provides a detailed comparison of the model's performance metrics, including accuracy before attack, accuracy after attack, and attack success rate. However, it is important to note that for now, we provide a few models, datasets, and attacks, and we plan to scale this process up in the future by providing access to users to upload their required models with datasets and desired attacks. This expansion will enable users to tailor the evaluation process to their needs and explore various scenarios. By incorporating a variety of models, datasets, and attacks, users can gain a more thorough understanding of the vulnerabilities present in different machine-learning systems.

5. Experiments and Results

In this section, we provide a detailed description of the experiments that were conducted to assess the efficacy of our proposed attack, including the details of datasets and metrics that were employed.

5.1 Datasets

Experiments are performed on several distinct datasets to evaluate the proposed attack's performance. These datasets are FaceForensics++ [2], Celeb-DF [1], World Leaders Dataset [4], DFDC-Preview [58], and FakeAV Celeb [59]. For the fake AV Celeb dataset, we perform the attack on the video subset only. The details of these datasets are discussed in the following sections.

5.1.1 FaceForensics (FF)++

One of the most challenging deepfake datasets, FaceForensics (FF)++ [2], comprises one thousand authentic YouTube videos featuring frontal features and no occlusions. Each video features individuals of different ethnicities ornamented with accessories, including spectacles, and features frames with contrasting lighting. These elements contribute to the challenge of distinguishing between real and counterfeit samples. These authentic videos are manipulated using computer graphics and deep learning techniques to generate the subsets i.e., Deepfakes, FaceSwap, Face2Face, NeuralTextures, and FaceShifter.

5.1.2 DFDC-P

The Deepfake Detection Challenge Preview (DFDC-P) dataset [58] comprises 5,000 manipulated and authentic videos. The authentic videos were obtained from the compensated actors, while several deepfake, GAN-based, and unlearned techniques were used to generate the fake videos. These videos are produced using facial manipulation methods such as DeepFakes and Face2Face, among others. DFDC considers various acquisition scenarios (e.g., outdoor, and indoor environments), illumination conditions (e.g., day and night), subject distance to the camera, pose variations, and more. DFDC is varied in numerous respects, including age, gender, and skin tone.

5.1.3 CelebDF

The Celeb-DF dataset [1] consists of 590 genuine videos in addition to 5639 fake videos. The authentic videos, which feature interviews with personalities of various ages, genders, and ethnic backgrounds, are obtained from YouTube. The recordings captured in the real world demonstrate a wide range of variations, encompassing disparities in facial dimensions (measured in pixels), orientations, illumination conditions, and backgrounds. The methodology employed to generate deepfakes centers on augmenting the luminance and contrast of facial images to mitigate discrepancies between the altered material and its environment. As a result, the videos that have been altered exhibit an elevated degree of deceit and highlight superior visual expertise.

5.1.4 WLDR

The World Leaders Dataset (WLRD) [1] is a selection of YouTube recordings featuring internationally renowned political figures, such as Barack Obama, Hillary Clinton, Joe Biden, Elizabeth Warren, Donald Trump, and Bernie Sanders. This dataset is additionally subdivided into several subsets, which comprise comedic imitations, face exchanges, and the original videos. Additionally, subdivisions for puppet master and lip-sync are included in the Obama section of this dataset. The WLRD dataset contains a face exchange subset in which the authentic appearance of a leader is substituted with that of a comedic impersonator. It is important to acknowledge that this subset lacks balance, as the quantity of fake videos is comparatively small in comparison to the number of authentic recordings about each featured leader.

5.1.5 FakeAVCeleb

The FakeAVCeleb dataset [59] consists of more than 20,000 false celebrity videos and 500 authentic ones. RealAudioRealVideo (RaRv), FakeAudioFakeVideo (FaFv), RealAudioFakeVideo (RaFv), and FakeAudioRealVideo (RaRv) are its four subsets. Videos of men and women from distinct ethnic groups: American, European, African, South Asian, and East Asian comprise the FakeAVCeleb dataset, which is diverse in terms of gender and ethnic distribution. From this dataset, we took Rv and Fv, which are video-only subsets because our proposed attack is performed on visual content.

5.2 Victim Detectors

For experiments, we choose two fusion-based deep learning models: Fused Truncated DenseNet [7] and Fused Swish-ReLU Efficient-Net [6], and two end-to-end deep learning models: Efficient-capsule Net [5] and ResNet-Swish-Dense54 [9]. Initially, this model [9] was trained on the FaceSwap and Face-Reenactment subsets of the FF++ and DFDC datasets, but we trained it on other subsets of the FF++, WLDR, and Celeb-DF datasets. In

addition, we use a unified end-to-end Dense Swin Transformer [8], which combines audio-visual deepfake detection, but we attack the video stream because the proposed attack targets visual content. Selecting these deep learning models as victim detectors ensures a fair evaluation of proposed attack transferability, as these models may still have similar vulnerabilities or weaknesses that can be exploited. To evaluate the robustness of the suggested adversarial method, it is necessary to select a variety of deepfake det ectors. We select these diverse deepfake detectors to ensure a comprehensive assessment of the proposed attack's impact on different deepfake detectors. This allows us to gain insights into the overall effectiveness and generalizability of our attack across various detection frameworks. The accuracy (ACC_{Before}) of these detectors on each dataset is presented in Table 1.

5.3 Evaluation Metrics

The following evaluation metrics were implemented to assess the effectiveness of the proposed attack on the dataset. These metrics were also utilized by the comparative methods [15, 18, 20]. These established metrics aid in evaluating the effect of the proposed attack.

5.3.1 Accuracy

Accuracy quantifies the proportion of correct predictions about the overall number of predictions generated by the model. It gives information regarding the proportion of instances that were accurately classified into the predetermined categories.

$$Acc\% = \frac{No.of\ Correct\ Predictions}{Total\ no.of\ Predictions} \times 100 \tag{16}$$

5.3.2 Attack Success Rate (ASR)

ASR is an essential metric utilized to assess the efficacy of a particular attack. The efficacy of the attack is assessed through the quantification of the percentage of model predictions that were altered successfully. It is worth noting that instances that were initially misclassified by the classifier are excluded from this metric.

$$ASR = \frac{Acc_{Before} - Acc_{After}}{Acc_{Before}} \times 100 \tag{17}$$

5.4 Experiment parameters

For the video preprocessing, we use MTCNN [60] for frame extraction of size 224×224. The proposed attack is performed only on a test set of all datasets in the black box setting. For the best results and visual naturalness of the proposed Facial distraction black box attack, we select the parameter of each adversarial perturbation very carefully. For hyperparameters, we set the intensity factor = 1.2 because intensity factors 0–3 give a visually natural outcome. For the proposed adversarial brightness, noise, and blur, we choose the minimum value of 3 w.r.t. the range of [0, 255], choosing these values slightly increases the illumination of the frame and perturbs the landmark region naturally. In addition to other state-of-the-art attacks, we set the same value = 5 for noise and blur. For PGD and FGSM attacks [61], we set the epsilon = 5. Whereas for one pixel DE [21] and one pixel SA [18], only one pixel is modified per frame, but for mole attack, a maximum of 15 pixels per frame are modified. Selecting these values for all attacks guarantees uniformity and the ability to make significant comparisons in evaluating the various attacks. Based on these parameters, we perform all the experiments.

5.5 Performance evaluation of proposed attack

To comprehensively assess the efficacy of our proposed attack in evading modern deepfake detectors, we carried out this experiment. Each model was evaluated before and after the attack. Before the attack, the models are evaluated through an accuracy assessment on all unperturbed datasets. Then an attack is performed on all datasets to perturb their test instances, and models are individually evaluated with perturbed samples. The results, including before-attack accuracy (ACC_{Before}), after-attack accuracy (ACC_{After}), and attack success rate of all deepfake detectors, are mentioned in Table 1, and details of the experiments are given in subsequent sections.

5.5.1 End-to-end deep learning target model

The objective of this experiment was to evaluate the susceptibility of end-to-end deep learning models utilized in deepfake detection to proposed adversarial attacks. A substantial decline in accuracy is observed after proposed adversarial attacks. The CAP net [5] has the highest accuracy loss on the NT subset of the FF++ dataset, with 50.4%. On WLDR, a loss of 50.1% was reported on the Joe subset. DFDC-P has a 51.2% accuracy drop, while CelebDF reported an accuracy loss of 60%. On the other hand, on RNSD [9], the F2F subset of FF++ had the lowest accuracy of 50.5%, followed by WLDR at 51.0% on the Joe subset and CelebDF at 49.8%. The DFDC-P dataset had the lowest accuracy drop of 45.6%. The overall highest ASR on CAP [5] is reported on the WLDR dataset, which is 51.7%, and on RNSD [9], the highest ASR is 49.3% on the DFDC-P dataset.

Table 1: Models Before and After Attack Accuracy % and ASR.

Models	Target model	Measures	Models Before and After Attack Accuracy % and ASR on different datasets											
			FF++					WLDR					DFDC-P	Celeb DF
			FS	DF	F2F	SH	NT	Obama	Clinton	Joe	Sander	Warren		
End-to-end deep learning	CAP[5]	ACC_{Before}	99.5	98.6	99.6	96.1	95.1	98.8	93	99.9	99.7	99.9	87.9	83.6
		ACC_{After}	**54.2**	**52.5**	**53.9**	**51.4**	**50.4**	**60.9**	**59.2**	**50.2**	**52.2**	**48.3**	**51.2**	**60.8**
		ASR	45.5	46.8	45.9	46.5	47.0	38.4	36.3	49.7	47.6	51.7	41.8	27.3
	RNSD [9]	ACC_{Before}	98.7	98.5	98.1	97.9	96.6	97.1	96.2	97.3	95.8	91.0	96.8	92.3
		ACC_{After}	**64.6**	**56.4**	**50.5**	**52.8**	**60.1**	**59.3**	**54.8**	**51.0**	**57.1**	**52.7**	**45.6**	**49.8**
		ASR	34.9	42.7	48.5	46.1	35.0	40.2	44.8	48.6	40.2	42.7	49.3	44.9
Fused Models	SRE [6]	ACC_{Before}	98.8	97.1	98.5	96	92.11	99.7	99.8	97.5	99.6	90.4	88.4	98.7
		ACC_{After}	**51**	**50.8**	**55.7**	**54**	**45.7**	**48.3**	**47.9**	**48.9**	**49.7**	**50.0**	**59.6**	**57.7**
		ASR	48.4	47.7	43.5	43.8	50.4	51.6	52.0	49.8	50.1	44.7	32.6	41.5
	Fused [7]	ACC_{Before}	95.7	93.9	92.6	83.5	60.9	94.5	84.1	89.6	89.5	93.1	86	96.8
		ACC_{After}	**52.9**	**58.1**	**52.5**	**58.5**	**57.4**	**51.4**	**49.9**	**53.6**	**59.5**	**45.7**	**55.4**	**56.6**
		ASR	44.7	38.1	43.3	29.9	5.7	45.6	40.7	40.2	33.5	50.9	35.6	41.5

5.5.2 Fusion-based target model

In this experiment, we select the two fusion-based models: Fused DenseNet [7] and Fused Swish-ReLU Efficient-Net [6]. These models attained the highest accuracy on several datasets on the unperturbed samples; however, their accuracy significantly decreased on samples attacked with the proposed method. The highest accuracy loss on the SRE [6] model was reported on the NT subset of the FF++ dataset with 45.7%. On WLDR, a loss of 47.9% was reported on the Clinton subset. CelebDF has a 57.7% accuracy drop, while DFDC reported an accuracy loss of 59.6%. However, on the Fused [7] model, the lowest accuracy was reported on the F2F subset of FF++, which is 52.5%; WLDR is 49.9% on the Clinton subset; and CelebDF is 56.6%, whereas the accuracy drops of 55.4% have been reported on the DFDC dataset. Table 3 represents existing attacks and proposed attack results of fused deepfake detectors. Both Fused models on the WLDR dataset achieved the highest ASR, SRE [6] reported an ASR of 52.0%, on the other hand, the fused model [7] achieved an ASR of 50.9%.

5.5.3 Unified Target Model

In this experiment, we selected a unified audio-visual deepfake detector [8]. This model was initially trained on DFDC and the audio-visual stream of the FakeAV Celeb dataset. Because the proposed attack modifies only visual information, for this experiment, we took only a visual subset of the FakeAV Celeb dataset. Although this model attained the highest accuracy on the unperturbed samples, their accuracy significantly dropped on samples attacked with the proposed method. Specifically, the accuracy dropped from 90.0% to 45.6% on the FakeAV Celeb dataset, with an ASR (Attack Success Rate) of 49.3%. On the DFDC dataset, a loss of 43.9% in accuracy was observed, dropping from 73.0% to 39.1%, with an ASR of 39.8%.

Overall, the above end-to-end model and fused models had the lowest accuracy on the NT and F2F subsets, even when nothing was changed. This is because their algorithm only alters the mouth area to modify a person's expression, resulting in such subtle changes that they are almost indistinguishable. In addition, due to deep learning models' linear decision boundary and extraordinary sensitivity to even minor input data modifications, these models are vulnerable to proposed adversarial attacks. The experiment with contemporary attacks was also performed and reported in Table 1, showing the good attack ability of the proposed method.

5.6 Comparative analysis with state-of-the-art attacks

The selected state-of-the-art attacks for analysis include black box, evasion, and statistical attacks. For statistical attacks, we selected blur [62] and noise [63]. In addition, pure black-box attacks involve Pix-DE [21], Pix-SA [18], and Mole attack, while evasion attacks consist of FGSM [61] and PGD [61]. The proposed attack combines statistical evasion black-box techniques, which is the reason for choosing these specific attacks in comparison. Detailed results of these attacks are provided in Tables 2-4.

5.6.1 Statistical attacks

Statistical attacks are added to test instances with different intensities. Specifically, we executed the blur and noise attacks varying in parameters and intensities, such as various kernel sizes, including 3, 5, 7, and 9, as seen in previous work [63], to diversify the testing data. By utilizing varying intensities, these attacks could either be noticeable or imperceptible to the human eye. We use salt and pepper noise and median blur with a kernel size of 5, which are minimal but perceptible statistical attacks visible on instances. Blurring an image reduces detail and sharpness. It softens or distorts sections of the image, making it less concentrated and smoother. This effect can improve aesthetics, and depth, or conceal critical information. Salt and pepper noise adds random black-and-white pixels to an image. Random speckles from this noise alter the image, making it grainy or distorted. It can lower

visual quality and clarity, making minute details harder to see. The impact of noise and blur on distorting image quality slightly affects the performance of models, as can be seen in Tables 2 to 4. Still, these methods, as adversarial attacks, do not help exploit vulnerabilities in the given deep learning models.

5.6.2 Black box attacks

We select three attacks for black box attacks: the one-pixel differential evolution attack, the simulated annealing one-pixel attack, and the facial mole attack. Each attack presents a distinct method for evaluating the susceptibility of a deepfake detector, offering a thorough evaluation of its security measures. The selection of these attacks was focused on their effectiveness in identifying potential vulnerabilities. One-pixel attacks utilizing Differential Evolution (DE) [21] and Simulated Annealing (SA) [18] are limited by factors such as intensity, resolution sensitivity, transferability, and hyperparameter dependence. Limitations of these attacks include their hyperparameter dependence, sensitivity to image resolution, and intensity. The impact of an isolated pixel modification on the overall prediction of a model is insignificant. Practical scenarios may require the alteration of numerous pixels to carry out a successful assault. Furthermore, a facial mole attack targets a specific area of the face; for this experiment, we select 25 moles per frame. however, the utilization of numerous moles renders the attack detectable and accessible for defense mechanisms. Furthermore, the effectiveness of DE and SA attacks is dependent on the initial starting point, a variable factor that may require multiple iterations to determine a feasible attack strategy. Due to these limitations, it is impractical to implement such attacks in real-world scenarios where robustness and effectiveness are crucial. The results of these black box attacks on selected deep-learning models can be seen in Tables 2 to 4.

5.6.3 Evasion attack

For this experiment, we selected FGSM and PGD attacks [61]. These attacks demonstrate efficacy in producing adversarial examples in white-box scenarios, assuming access to model gradients; they encounter difficulties when applied in black-box environments. These attacks slightly fool various models, but their lack of transferability limits attack usefulness in practice. Moreover, these attacks are affected by hyperparameters, making it harder to find the best values in the black box setting, where information about the model is missing. For this attack, we randomly select an epsilon value of 5 for both attacks, which helps to slightly degrade the performance of all deepfake detectors.

The results of the experiment showed that the deep learning models used in deepfake detection were indeed susceptible to the proposed adversarial attacks in comparison with state-of-the-art attacks. The models exhibited varying levels of vulnerability depending on the parameters and intensities of the attacks, with some attacks being easily detectable while others were almost indistinguishable. Comparing the results of all attacks in Tables 2, 3, and 4, it can be noticed that the proposed attack is robust, transferable, and less suspectable than the others. With minimal parameters and less computational time, the proposed attack fails all different kinds of deep learning models.

Table 2: Accuracy drop % of end-to-end deep learning models after Attack.

Models	Attacks	Datasets											
		FF++					WLDR					DFDC	Celeb DF
		FS	DF	F2F	SH	NT	Obama	Clinton	Joe	Sander	Warren		
RNSD [9]	**Blur[62]**	89.4	90.1	89.5	88.5	77.4	90.4	93.4	88.5	85.6	88.4	75.9	85.7
	Noise [63]	83.7	85.5	82.7	74.7	73.6	82.4	88.9	87.8	82.7	83.6	72.2	77.4
	1 Pix-DE[21]	92.2	91.4	90.3	91.8	87.2	92.2	91.4	90.3	91.8	87.2	87.5	87.2
	1 Pix-SA[18]	91.7	92.3	90.1	90.3	88.5	91.7	92.3	90.1	90.3	88.5	85.3	89.5
	Mole	64.4	67.8	76.7	82.9	59.1	64.4	67.8	76.7	82.9	59.1	55.0	63.1
	FGSM[61]	83.0	76.3	62.2	96.3	71.1	98.3	94.6	91.3	94.2	79.6	56.6	66.4
	PGD[61]	95.4	96.2	88.3	97.7	85.4	95.4	98.2	97.4	94.7	82.4	71.8	87.5
	Proposed	**54.2**	**52.5**	**53.9**	**51.4**	**50.4**	**60.9**	**59.2**	**50.2**	**52.2**	**48.3**	**51.2**	**60.8**
Capsule [5]	**Blur[62]**	89.5	88.5	85.7	86.9	88.5	78.9	85.6	89.5	87.9	88.5	75.6	76.5
	Noise [63]	85.3	82.1	89.7	84.3	80.6	84.4	79.1	87.5	90.3	88.6	78.4	70.5
	1 Pix-DE[21]	95.4	96.7	95.7	90.5	94.3	96.7	90.6	97.8	98.9	97.9	85.3	76.8
	1 Pix-SA[18]	96.7	97.3	97.5	93.3	92.5	95.4	90.6	96.5	97.5	97.6	84.6	80.6
	Mole	78.4	84.5	73.9	66.5	70.1	67.5	77.5	63.6	76.5	69.9	65.4	81.4
	FGSM[61]	89.0	58.8	80.2	78.1	79.2	79.7	68.6	85.7	85.6	83.5	50.3	55.9
	PGD[61]	94.8	85.4	82.4	91.3	85.1	88.4	87.9	85.7	89.3	88.4	64.9	58.4
	Proposed	**64.6**	**56.4**	**50.5**	**52.8**	**60.1**	**59.3**	**54.8**	**51.0**	**57.1**	**52.7**	**45.6**	**49.8**

Table 3: Accuracy drop % of fused deep learning models after Attack.

Models	Attacks	Datasets											
		FF++					WLDR					DFDC	Celeb DF
		FS	DF	F2F	SH	NT	Obama	Clinton	Joe	Sander	Warren		
SRE [6]	Blur[62]	88.6	86.4	85.4	89.4	85.4	94.3	89.4	89.7	88.5	85.6	76.4	87.5
	Noise [63]	86.4	85.3	88.4	85.3	83.7	90.3	88.8	84.3	87.5	81.2	74.3	83.9
	1 Pix-DE[21]	95.3	94.5	95.3	94.3	90.3	97.5	97.7	95.4	96.6	88.2	85.4	96.9
	1 Pix-SA[18]	96.7	93.3	96.4	92.5	89.6	96.4	93.4	94.8	94.3	87.5	83.5	97.4
	Mole	69.4	65.2	67.4	66.4	59.7	73.4	75.6	69.9	78.5	57.6	50.3	77.3
	FGSM[61]	72.4	86.2	60.2	94.6	65.4	98.6	90.9	70.6	99.7	95.0	61.6	84.8
	PGD[61]	95.5	97.3	75.7	95.7	87.1	95.8	91.8	88.3	99.7	94.9	75.5	96.2
	Proposed	**51.0**	**50.8**	**55.7**	**54.0**	**45.7**	**48.3**	**47.9**	**48.9**	**49.7**	**50.0**	**59.6**	**57.7**
Fused [7]	Blur[62]	84.6	80.6	80.7	73.3	58.5	84.8	75.1	77.9	79.6	86.5	77.6	87.5
	Noise [63]	82.3	78.5	76.9	60.8	54.6	87.5	72.5	73.4	70.5	79.6	70.5	79.8
	1 Pix-DE[21]	93.9	90.5	89.6	81.6	57.5	92.4	80.7	86.5	88.5	91.5	83.7	94.9
	1 Pix-SA[18]	92.6	92.4	90.4	80.5	56.6	90.3	81.2	87.9	87.9	92.5	82.2	92.3
	Mole	79.8	63.1	65.1	51.8	59.1	66.7	61.5	63.4	65.6	66.5	59.6	60.7
	FGSM[61]	72.3	66.5	64.1	95.5	85.3	81.0	75.4	92.3	97.4	89.9	79.5	79.6
	PGD[61]	83.4	95.8	88.4	96.5	88.3	80.9	93.8	88.7	80.1	97.7	84.5	96.9
	Proposed	**52.9**	**58.1**	**52.5**	**58.5**	**57.4**	**51.4**	**49.9**	**53.6**	**59.5**	**45.7**	**55.4**	**56.6**

Table 4: Accuracy drop % of unified deep learning models after Attack.

Models	Attacks	Models After Attack Accuracy %	
		DFDC	FakeAVCeleb
Unified [8]	Blur[62]	69.4	86.4
	Noise [63]	70.3	87.3
	1 Pix-DE[21]	72.3	89.9
	1 Pix-SA[18]	71.9	87.4
	Mole	65.5	73.5
	FGSM[61]	50.3	82.1
	PGD[61]	54.9	57.9
	Proposed	**43.9**	**45.6**

5.7 Ablation study

In this experiment, we tried different combinations of attacks to evade the performance of existing deepfake detectors. We found that certain attack combinations were able to successfully evade the performance of existing deepfake detectors. These attacks included manipulating texture patterns to create more convincing perturbed frames, but these perturbations do not look as natural as the proposed attack. However, some of these combinations make the image quality very low, with a dark or bright illumination effect and distorted facial features. Despite their success in evading detection, it is important to note that these alterations significantly compromised the overall image quality and appearance, making them easily distinguishable. Therefore, we proposed the attack, which perturbs images without compromising image quality. The combination of statistical distraction attacks chosen in this study are as follows:

5.7.1 Gamma Dark and Bright attack

In this attack, we used gamma correction in place of adversarial illumination adjustment, and the rest of the perturbations were added the same as the proposed attack. Gamma correction is generally used to improve image quality by adjusting brightness and contrast. Using less gamma brightens and blurs the frame; however, using slightly higher values produces a darker image. For this experiment, we use the minimum value of gamma factor = 0.5; on the other hand, we use gamma value = 3, which darkens the frame. The results of these attacks are comparatively the same as the proposed attack, but the quality of the perturbed frame does not look natural.

5.7.2 Proposed attack variations

In the variation to the proposed attack, we change the placement of blur and noise. We use blur before noise, and the results are almost equivalent to the proposed attack, but the visual naturalness of the image is distorted. By employing noise before blurring an image, it is possible to improve the visibility of details and textures. Conversely, applying blur before noise may yield a smoother appearance by diminishing the visibility of noise.

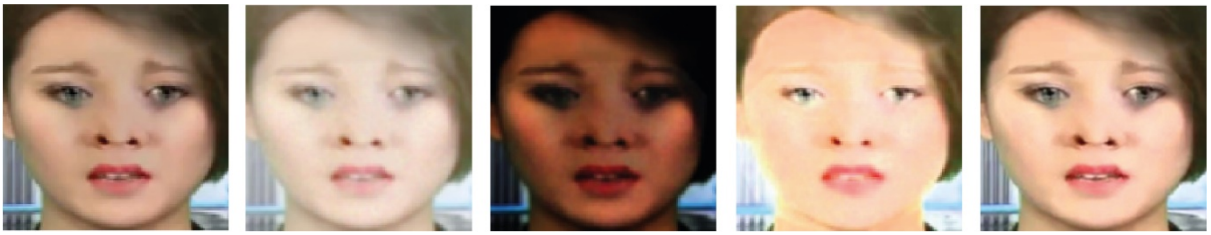

Figure 6: Deepfake images from the dataset, gamma bright frame, gamma dark frame, FDB-variation perturbed frame, proposed FDB attack perturbed frame.

The order of noise and blur application can have a substantial impact on the overall visual impact of an image. By applying noise before blur, it is possible to emphasize the image's textures and details, thereby increasing the appearance of depth and realism. On the contrary, the application of blur before noise reduces its clarity and sharpness, making it harder to distinguish details and potentially distorting the overall quality of the image. The results of the above method in comparison to the proposed attack can be seen in Table 5, and the outcome of these attacks can be analyzed in Figure 6.

5.8 Analysis and Discussion

Several analyses have been conducted in this section, including an evaluation of the attack's transferability across multiple deepfake detectors, an assessment of the surrogate model's explainability and quality, and an evaluation of the attack's robustness.

5.8.1 Transferability Analysis

To assess the transferability of the proposed adversarial attack, we performed a transfer attack and assessed adversarial samples against a variety of detection algorithms, fused, end-to-end deep learning models, and unified models. Observing the results from Tables 2, 3, and 4, it can be noticed that it significantly improved the applicability of adversarial perturbations to a variety of detection techniques. The transferability of the proposed attack is due to our attack's ROI specification based on the surrogate model observation. This is because the proposed attack exploits weaknesses in the decision-making processes of DNN-based deepfake detectors. The aforementioned weaknesses arise due to the complex, multidimensional, and non-linear characteristics of deep neural networks, which render them prone to fluctuations in adversarial attacks.

5.8.2 Explainability Analysis

To assess the explainability of target models, we utilized a surrogate model to quantify the resemblance between heat maps originating from real facial frames and those generated by deepfakes. Based on the output of the surrogate model, as illustrated in Figure 2, the heat maps produced emphasize critical information. Under the assumption that attention patterns among various DNNs are comparable. Utilizing heatmaps facilitates the effective initiation of an adversarial attack against all deep neural networks. Through making use of the proposed attack, we successfully redirected the model's focus away from features, leading to misclassification. This methodology emphasizes the susceptibility of deep neural networks to adversarial attacks, as it illustrates how as an attacker, we can deceive the model into producing erroneous predictions through the manipulation of attention patterns.

5.8.3 Qualitative Analysis

We conducted a qualitative analysis by comparing the perturbed samples of different attacks as mentioned in Section 5 with our proposed attack. Although basic methods such as noise addition, and image blurring can cause image disruption, they are frequently ineffective as adversarial attacks when robust models are employed. Although these methods may cause some disruptions, they do not possess the deliberate and calculated disruptions that are inherent in more sophisticated approaches. Conversely, methodologies such as FGSM and PGD [61] exhibit superior effectiveness when it comes to constructing adversarial examples. However, there continues to be a desire for straightforward and efficient attacks, particularly in situations where computational resources or model complexity limitations are apparent. Manipulating just one or few pixels through One-pixel attacks [18, 21], and mole attacks introduces localized perturbations, the impact might be limited, and robust models can potentially mitigate the effect. It is critical to strike a balance between simplicity and impact when formulating practical adversarial attacks that can effectively challenge neural networks. In contrast to the aforementioned attacks, our proposed attack is unnoticeable upon observation and exclusively disrupts the facial regions. On the other hand, existing attacks are distributed throughout the complete frame sequence, thereby increasing their visibility and facilitating their defense. Our method, in contrast to attacks, only perturbs instances within the ROI.

Table 5: Comparison of the proposed attack with other proposed variants.

DFDC dataset		Models				
		RNSD [9]	**CAP[5]**	**Fused [7]**	**SRE [6]**	**Unified [8]**
Before Attack	**Test Acc**	89.9	87.9	86.0	88.4	73.0
After Attack	**Gamma (Bright)**	50.3	64.2	57.3	57.3	45.6
	Gamma (Dark)	**42.3**	57.4	55.7	**58.4**	43.2
	FDB attack variation	48.9	57.3	55.9	59.9	40.8
	Proposed FDB attack	45.6	**51.2**	**55.4**	59.6	**39.1**

Therefore, in comparison to alternative attacks, our proposed attack attained the maximum success rate across all datasets. A visual analysis of our attack, as well as other existent attacks, is depicted in Figure 5.

5.8.4 Quantitative Analysis

To analyze the robustness of our attack, we applied Blur, Noise, 1 Pix-DE, 1 Pix-SA, Mole attack, FGSM, and PGD attacks to each instance of the test set and created its perturbed instance. It is significant to highlight that the attacks were executed across the entirety of the facial frame. All deepfake detectors were utilized in this investigation, and each trained subset model was evaluated using perturbed instances. In contrast to our attack, the prior attacks were executed on complete images, rendering them perceptible and readily identifiable. The computational expense of the differential evaluation one-pixel attack arises from the requirement to assess a substantial quantity of candidate pixels. Conversely, the simulated annealing one-pixel attack may necessitate multiple iterations to achieve success, as it may be sensitive to the initial starting point of the optimization procedure. The PGD and FGSM attacks, on the other hand, focus on manipulating a few selected pixels but are not effective because these attacks are computationally expensive. The outcomes depicted in Tables 2-4 indicate that the accuracy drop of our proposed attack was the highest compared to the other attacks. Therefore, we assert that our method of attack is transferable and can be effectively executed to undermine the current deepfake detectors.

6. Conclusion

The proposed research has emphasized the susceptibilities of deepfake detection systems to adversarial attacks, exposing significant flaws in current approaches. The proposed penetration testing tool has yielded useful insights into these vulnerabilities, highlighting the need for more effective detection solutions. It is crucial to improve the security and efficiency of deepfake detection technologies to preserve the authenticity of digital media in the face of more advanced synthetic content. Future research should give priority to the advancement of deepfake detection systems that possess the ability to withstand adversarial attacks. This entails investigating adaptive machine learning algorithms that can evolve in response to emerging threats. In addition, the implementation of complex security measures and the integration of data from multiple sources could enhance the effectiveness of detection systems.

Funding: This work was supported by the grant of the Punjab Higher Education Commission (PHEC) of Pakistan via Award No. (PHEC/ARA/PIRCA/20527/21).

Acknowledgment: This work was supported by the Multimedia Signal Processing (MSP) research lab at the University of Engineering and Technology (UET) Taxila. We would like to thank Prof. Hany Farid from the University of California Berkeley for providing us World Leaders dataset.

References

[1] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).

[2] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11).

[3] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, *53*(4), 3974-4026.

[4] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting World Leaders Against Deep Fakes. In *CVPR workshops* (Vol. 1, p. 38).

[5] Ilyas, H., Javed, A., Malik, K. M., & Irtaza, A. (2023). E-Cap Net: an efficient-capsule network for shallow and deepfakes forgery detection. *Multimedia Systems*, *29*(4), 2165-2180.

[6] Ilyas, H., Javed, A., Aljasem, M. M., & Alhababi, M. (2023, February). Fused swish-relu efficient-net model for deepfakes detection. In *2023 9th International Conference on Automation, Robotics and Applications (ICARA)* (pp. 368-372). IEEE.

[7] Khalid, F., Javed, A., Irtaza, A., & Malik, K. M. (2023, May). Deepfakes catcher: a novel fused truncated densenet model for deepfakes detection. In *Proceedings of International Conference on Information Technology and Applications: ICITA 2022* (pp. 239-250). Singapore: Springer Nature Singapore.

[8] Ilyas, H., Javed, A., & Malik, K. M. (2023). AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, *136*, 110124.

[9] Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. *The Visual Computer*, *39*(12), 6323-6344.

[10] Javed, A., & Malik, K. M. (2022, August). Faceswap Deepfakes Detection using Novel Multi-directional Hexadecimal Feature Descriptor. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 273-278). IEEE.

[11] Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., & Manjunath, B. S. (2019). Detecting GAN generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*.

[12] Kaddar, B., Fezza, S. A., Hamidouche, W., Akhtar, Z., & Hadid, A. (2023). On the effectiveness of handcrafted features for deepfake video detection. *Journal of Electronic Imaging*, *32*(5), 053033-053033.

[13] Kaddar, B., Fezza, S. A., Akhtar, Z., Hamidouche, W., Hadid, A., & Serra-Sagristà, J. (2024). Deepfake detection using spatiotemporal transformer. *ACM Transactions on Multimedia Computing, Communications and Applications*.

[14] Khalid, F., Javed, A., Ilyas, H., & Irtaza, A. (2023). DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Systems with Applications*, *222*, 119843.

[15] Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 658-659).

[16] Wang, X., Ni, R., Li, W., & Zhao, Y. (2021, September). Adversarial attack on fake-faces detectors under white and black box scenarios. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3627-3631). IEEE.

[17] Gandhi, A., & Jain, S. (2020, July). Adversarial perturbations fool deepfake detectors. In *2020 International joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

[18] Zhou, T., Agrawal, S., & Manocha, P. (2022). Optimizing one-pixel black-box adversarial attacks. *arXiv preprint arXiv:2205.02116*.

[19] Abdullah, H., Rahman, M. S., Garcia, W., Warren, K., Yadav, A. S., Shrimpton, T., & Traynor, P. (2021, May). Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 712-729). IEEE.

[20] Dong, J., Wang, Y., Lai, J., & Xie, X. (2023). Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, *18*, 2596-2608.

[21] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828-841.

[22] Hou, Y., Guo, Q., Huang, Y., Xie, X., Ma, L., & Zhao, J. (2023). Evading deepfake detectors via adversarial statistical consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12271-12280).

[23] Mohiuddin, S., Sheikh, K. H., Malakar, S., Velásquez, J. D., & Sarkar, R. (2023). A hierarchical feature selection strategy for deepfake video detection. *Neural Computing and Applications*, *35*(13), 9363-9380.

[24] Kolagati, S., Priyadharshini, T., & Rajam, V. M. A. (2022). Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. *International Journal of Information Management Data Insights*, *2*(1), 100054.

[25] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

[26] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE.

[27] Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4113-4122).
[28] Liang, B., Wang, Z., Huang, B., Zou, Q., Wang, Q., & Liang, J. (2023). Depth map guided triplet network for deepfake face detection. *Neural Networks*, *159*, 34-42.
[29] Zhao, L., Zhang, M., Ding, H., & Cui, X. (2023). Fine-grained deepfake detection based on cross-modality attention. *Neural Computing and Applications*, *35*(15), 10861-10874.
[30] Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14923-14932).
[31] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019, September). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1-8). IEEE.
[32] Kohli, A., & Gupta, A. (2022). Light-weight 3DCNN for DeepFakes, FaceSwap and Face2Face facial forgery detection. *Multimedia Tools and Applications*, *81*(22), 31391-31403.
[33] Heo, Y. J., Yeo, W. H., & Kim, B. G. (2023). Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, *53*(7), 7512-7527.
[34] Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020, August). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision* (pp. 86-103). Cham: Springer International Publishing.
[35] Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., ... & Yu, N. (2021). Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 772-781).
[36] Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16317-16326).
[37] Zhang, H., Chen, B., Wang, J., & Zhao, G. (2022). A local perturbation generation method for GAN-generated face anti-forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(2), 661-676.
[38] Liu, C., Chen, H., Zhu, T., Zhang, J., & Zhou, W. (2023). Making DeepFakes more spurious: evading deep face forgery detection via trace removal attack. *IEEE Transactions on Dependable and Secure Computing*, *20*(6), 5182-5196.
[39] Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X., ... & Pu, G. (2020, October). Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1217-1226).
[40] (21 November 2022). *Few-Shot Face Translation GAN*. Available: GitHub - shaoanlu/fewshot-face-translation-GAN: Generative adversarial networks integrating modules from FUNIT and SPADE for face-swapping.
[41] Shahriyar, S. A., & Wright, M. (2022, May). Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes* (pp. 13-18). Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE.
[42] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE.
[43] Sohrawardi, S. J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., & Wright, M. (2019, November). Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security* (pp. 2613-2615).
[44] Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2021). Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 923-932).
[45] Lim, N. T., Kuan, M. Y., Pu, M., Lim, M. K., & Chong, C. Y. (2022, August). Metamorphic testing-based adversarial attack to fool deepfake detectors. In *2022 26th International Conference on Pattern Recognition (ICPR)* (pp. 2503-2509). IEEE.
[46] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
[47] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
[48] Lou, Z., Cao, G., & Lin, M. (2023). Black-box attack against GAN-generated image detector with contrastive perturbation. *Engineering Applications of Artificial Intelligence*, *124*, 106594.

[49] Alkishri, W., Widyarto, S., & Yousif, J. H. (2024). Evaluating the Effectiveness of a Gan Fingerprint Removal Approach in Fooling Deepfake Face Detection. *Journal of Internet Services and Information Security (JISIS)*, *14*(1), 85-103.
[50] McCloskey, S., & Albright, M. (2019, September). Detecting GAN-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)* (pp. 4584-4588). IEEE.
[51] Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., & Liu, X. (2021). Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8565-8574).
[52] Lim, N. T., Kuan, M. Y., Pu, M., Lim, M. K., & Chong, C. Y. (2022, August). Metamorphic testing-based adversarial attack to fool deepfake detectors. In *2022 26th International Conference on Pattern Recognition (ICPR)* (pp. 2503-2509). IEEE.
[53] Zatorre, R. J., Mondor, T. A., & Evans, A. C. (1999). Auditory attention to space and frequency activates similar cerebral systems. *Neuroimage*, *10*(5), 544-554.
[54] Gowrisankar, B., & Thing, V. L. (2024). An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *Computers & Security*, *139*, 103684.
[55] Blakemore, C., Carpenter, R. H., & Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*, *228*(5266), 37-39.
[56] Simonyan, K. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[57] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
[58] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.
[59] Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.
[60] Xiang, J., & Zhu, G. (2017, July). Joint face detection and facial expression recognition with MTCNN. In *2017 4th international conference on information science and control engineering (ICISCE)* (pp. 424-427). IEEE.
[61] Villegas-Ch, W., Jaramillo-Alcázar, A., & Luján-Mora, S. (2024). Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW. *Big Data and Cognitive Computing*, *8*(1), 8.
[62] Ruiz, N., Bargal, S. A., & Sclaroff, S. (2020). Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16* (pp. 236-251). Springer International Publishing.
[63] Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. *The Visual Computer*, *39*(12), 6323-6344.