Exposing the Limits of Deepfake Detection using novel Facial mole attack: A Perceptual Black-Box Adversarial Attack Study

Qurat Ul Ain[a], Ali Javed[b], Khalid Mahmood Malik[c], Aun Irtaza[a],*

[a] Dept of Computer Science, University of Engineering and Technology-Taxila, Pakistan; [b] Dept of Software Engineering University of Engineering and Technology-Taxila, Pakistan, Department of Computer Science and Engineering, University of Michigan, Flint, MI, USA

2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10647949

ABSTRACT

Recently, we have observed an exponential growth in highly realistic deepfake videos, which are often used to spread disinformation, defame individuals, and even influence political outcomes. To combat these manipulated videos, researchers have proposed various deepfake detection techniques. Recent research has revealed that these detection techniques are vulnerable to different adversarial attacks. This paper examines the vulnerability of deepfake detectors to adversarial black-box attacks in terms of performing penetration testing to expose the existing defense benchmarks of current deepfake detectors. We present a perceptual facial mole black-box adversarial attack on deepfake detectors, where the attacker has limited knowledge of the architecture and settings of the detector. The proposed attack is visually natural and transferable based on the attention distraction mechanism, which distracts the model-shared attention patterns from the region of interest to other regions. We illustrate the efficacy of our attack on multiple cutting-edge deepfake detectors. This attack demonstrates that small perceptible perturbations that are visually natural on the facial face can disrupt and reduce the accuracy of the detectors significantly, up to 40.3%, with the highest success rate of 48.7%. Our findings highlight the necessity for proposing effective deepfake detectors that are resistant to black-box attacks.

Index Terms— *Adversarial attack, Black-box attack, Deepfakes detection, Facial Mole attack.*

1. Introduction

Deepfakes, synthetic media created using advanced synthesis techniques like Generative Adversarial Networks (GANs) and Autoencoders [1, 2] have become easier to create and spread fake news, defame individuals, and potentially affect political campaigns [2]. Researchers have developed deepfake detectors, including feature-based methods [3] and deep learning algorithms [4], to identify these videos and minimize their negative impact. In addition local feature extractors, CNN-based classifiers [5], cross-modality attention-based deepfake detectors [6], and vision transformer and distillation approaches [7] have all shown promise in detecting deepfakes, but their limitations in reliability and vulnerability to adversarial attacks limit their applicability. Despite these advancements, existing deepfake detectors are vulnerable to adversarial attacks [8, 9], making their performance worthless under such attacks.

Adversarial attacks are cyber-attacks that deceive models' detection through inaccurate predictions. Two common types of attacks are white-box and black-box, used to test model performance and vulnerability. A white-box attack provides full model knowledge, while black-box attacks require less information and are crucial for testing deepfake detectors in real-world scenarios. However, current black-box methods are ineffective due to knowledge gaps, limitations in transferring adversarial examples [10, 11] between models, unrealistic samples, and computational complexity. The need for deepfake detectors that are resistant to adversarial attacks and provide trustworthy results is urgent, especially for high-resolution images, which are harder to distinguish due to their finer details. Current deepfake detection techniques are largely untested, creating security gaps. The double-masked guided attack [12] targets critical facial areas but has limitations in applicability across different forensic classifiers. In [13], FGSM and C&W attacks on VGG16 and ResNet18 using fake images [14], with white-box success rates of 100% except for FGSM on the ResNet18 model, while for the black-box attacks, the attack success rate drops significantly. In [15], a black-box attack by applying makeup artifacts to facial landmark regions decreased the accuracy of victim models by 50%, but its effectiveness is lower than other conventional attacks like PGD and FGSM.

Adversarial examples [16, 17] exploit weaknesses in transferability between white-box and black-box settings in deepfake detection. To overcome these issues, a simple attack [10, 11] based on model attention distraction [18] is proposed. This visually natural and transferable attack can fail deepfake detectors even in a black-box setting. The researchers designed a more powerful black-box attack that can fail deepfake detectors and reveal their vulnerabilities. They examine the characteristics of real and deepfake instances by analyzing heatmaps produced by a surrogate model based on model attention distraction [18]. This concept involves using the attention mechanism inherent in deep learning models. These models tend to concentrate on specific regions of given instances. From the heatmaps, we investigate how these detection systems are created and highlight their characteristics that are susceptible to adversarial attacks. After finding the crucial characteristics,

we proposed novel perceptible mole-based perturbation that is visually natural and can be transferable to several deepfake detectors. Here are the contributions of our paper:

- A novel facial mole black-box adversarial attack is introduced that manipulates a small number of pixels on the facial portion that looks visually natural and fails the deepfake detectors.
- We present a black-box attack that requires only probability labels and without needing knowledge of the inner workings of the target DNNs, making it easier to execute than existing methods.
- The proposed attack is transferable to various machine learning models, demonstrating excellent success rates, and indicates that most real or deepfake instances can be transformed into other classes.

2. Proposed method

The proposed black-box adversarial attack strategy aims to degrade the performance of deepfake detectors without requiring access to the target detector's parameters or training data. The attack is conducted through a surrogate model, which generates heatmaps to check explainability and analyze key traits based on the model attention distraction [18], of facial frames focused by the trained model. A universal mask is created to specify the region of interest on the targeted facial frame. Random black moles are scattered on a masked facial portion, making the attack perceptible but visually natural. The perturbed frames are generated through the concatenation of the original frame and masked frame containing moles. The test set of perturbated facial frames is passed to the victim models to evaluate their performance under adversarial conditions. The method is described in Figure 1 and details are provided in subsequent sections.

2.1 Target and surrogate model

The target model is a deepfakes detector that detects whether a video is real or fake. Deepfake detectors often use deep learning architectures that mimic the human central nervous system, suggesting they share similar properties [21]. Recent research in biology has shown that different individuals generate similar patterns of cerebral activity [20], such as focused attention, due to similar neuron hyper-perception properties of deep learning models. To achieve this, VGG-19 as a surrogate model is used that resembles the target model's behavior. VGG-19 is trained with a 70:30 dataset assuming the target model's architecture and tested with perturbed samples, resulting in misclassifications. The target model performs per-frame detection through facial frame extraction tasks using Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [19], which recognize faces using landmark features like eyes, nose, and mouth.

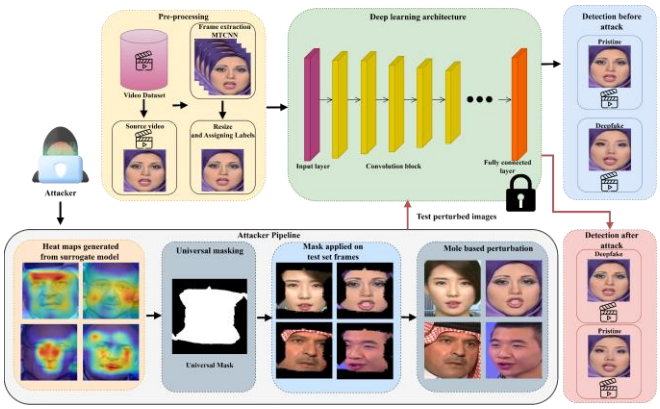

Figure 1. Framework of Proposed Attack.

2.2 Heatmap generation

To increase the transferability of the proposed attack, model attention distraction is used, which distracts model attention from ROI and is transferable to all deep learning models. The Grad-cam visualization technique is used to identify significant facial regions for classification tasks, while VGG-19 is used for heat map generation. The surrogate model focuses on highlighting landmark features in the face with a red color, making the attack computationally less complex and allowing the deepfake detector to bypass its performance.

2.3 Universal mask creation

A universal mask is created using a surrogate model heatmap analysis to extract a specific region of interest (ROI) from a facial frame. This mask separates important parts of the image and is applied to all frames of the video to segment ROIs. The resulting sample contains only a facial portion within the masked region, with the remaining background pixels set to zero. This technique segments and preserves facial ROI inside and around landmark locations, directing detector attention to the important facial regions. This mask improves attack performance and reduces computational cost.

To create a universal mask, firstly, each heatmaps h_{xy} is selected and split into RGB channels. To extract the ROI, the red channel rh_{xy} is further separated. This step involves extracting its red channel to generate a grayscale version of the image. In the next step, a binary mask is created by isolating specific areas of the red channel that lie within the threshold (t=160) of grayscale color range. Several masked frames were then combined through the "bitwise OR" into a single universal mask M_{xy} as:

$$M_{xy} = \sum_{i=0}^{n} rh_{xy} \quad (1)$$

Finally, a masked facial frame Mf_{xy} is generated through the "bitwise AND" of the actual facial frame and its mask as follows:

$$Masked\ frame\ (Mf_{xy}) = f_{xy} \&\ M_{xy} \quad (2)$$

2.4 Facial Moles Attack

After extracting the facial ROI through universal masking, we introduce a novel facial mole attack on the selected region. To perturb an image, random boxes were scattered on an ROI. These black random boxes resemble facial moles on the face. The size of the mole varies, as we have used two different sizes of moles. 1-pixel mole size is equivalent to a single pixel, which looks slightly perceptible when a few moles are added to ROI. The other pixel size is equivalent to 2 pixels that are perceptual, but this perturbation is undetectable to the human eye, as it visually looks like natural moles on a face. Minimal moles around the masked facial frame make it visually natural, but this perturbation disrupts the performance of detectors by disturbing the spatial coherence of the facial frame. This makes it harder for the detector to distinguish between real and deepfake facial attributes.

To add black moles to each masked frame Mf_{xy}, it needs to determine the number of moles n to be added. The parameter m_i represents the height and width of mole. The size of moles can be either 1-pixel or 2-pixels, denoted as m_1 and m_2, respectively.

$$m_i = (h_i \times w_i) \quad (3)$$

For the addition of moles, random coordinates (x, y) were selected within ROI. To add the desired number of moles on pixels of the masked frame, this process was repeated respectively.

$$\widehat{M}f_{xy}, = Mf_{xy} \sum_{i=1}^{n} [y: y + m_i, x: x + m_i] \quad (4)$$

The mole-masked frame $\widehat{M}f_{xy}$ is constructed by the addition of the number n of moles at random positions of ROI. Whereas each mole is of specific width and height concatenated with respective x and y coordinates and convert the specific point to a black mole.

2.5 Perturbed frame generation

Once moles have been added to the masked frame, it is combined with the original facial frame. Then, the perturbed facial frame is created by blending the masked area with the original frame. This results in generating perturbed frames with random moles positioned in them. So, the perturbed frame Pf_{xy} is formed through "bitwise AND" of mole-based masked frame $\widehat{M}$ with the original frame f_{xy}, as mentioned in Equation 5.

$$Perturbed\ frame\ (Pf_{xy}) = f_{xy}\ \&\ \widehat{M}f_{xy} \quad (5)$$

The proposed method involves constructing separate perturbed subsets for 1-pixel and 2-pixel attacks, each with a different number of moles. This can be observed in Figure 2. These perturbed subsets were then used to test the target models, as explained in Section 4.

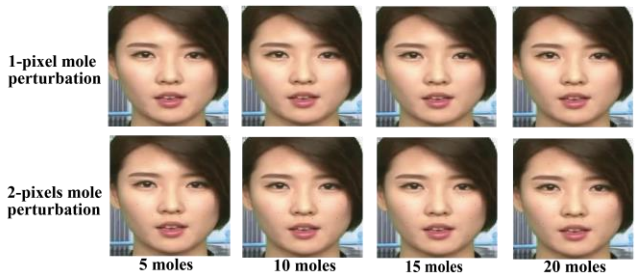

Figure 2. Imperceptible and perceptible Facial moles attacked frames.

3. Experiments and results

We conducted our experiments on the FaceForensics++ (FF++) [1] and World Leaders dataset WLDR [22] due to their diversity and selected the best detectors as our attack targets. First, we created a perturbed test set of the dataset and targeted the three different models. We used these models as a test bed to examine the robustness of our attacks. In this section, we also discussed and analyzed the proposed attack and performed a comparative analysis with state-of-the-art attacks.

3.1 Dataset

We evaluated the performance of our proposed attack on the FF++ (C-23) and WLDR datasets. FF++ dataset, consisting of 1000 YouTube videos with frontal faces. The dataset is divided into five subsets i.e., DeepFakes, FaceSwap, Face2Face, FaceShifter, and NeuralTextures. The WLRD is a selection of YouTube recordings featuring internationally renowned political figures. We combined this dataset. 80% of videos were split for training and 20% for testing, aiming to distinguish between real and fake samples.

3.2 Deepfake Detectors

To test the transferability of the proposed attack, we chose three different trained target models. These models were selected to check the transferability and success rate of the proposed attack. The details of the proposed model are mentioned as follows:

3.2.1 End-to-end deep learning target model.

The first target model is the end-to-end VGG-16 deep learning model, as its architecture is simple and it shows promising performance in several existing deepfakes detection techniques [23, 24]. For experimentation, we selected the pre-trained VGG-16 model, with 16 convolution layers, and modified its last dense layer for detection of real and deepfakes.

3.2.2 Hybrid target model

The second model is the InceptionResNet-BiLSTM [4] deep-learning model. This model detects deepfake videos across different ethnicities and lighting conditions. The hybrid framework combines a customized InceptionResNetV2 for feature extraction and a BiLSTM for classification. The

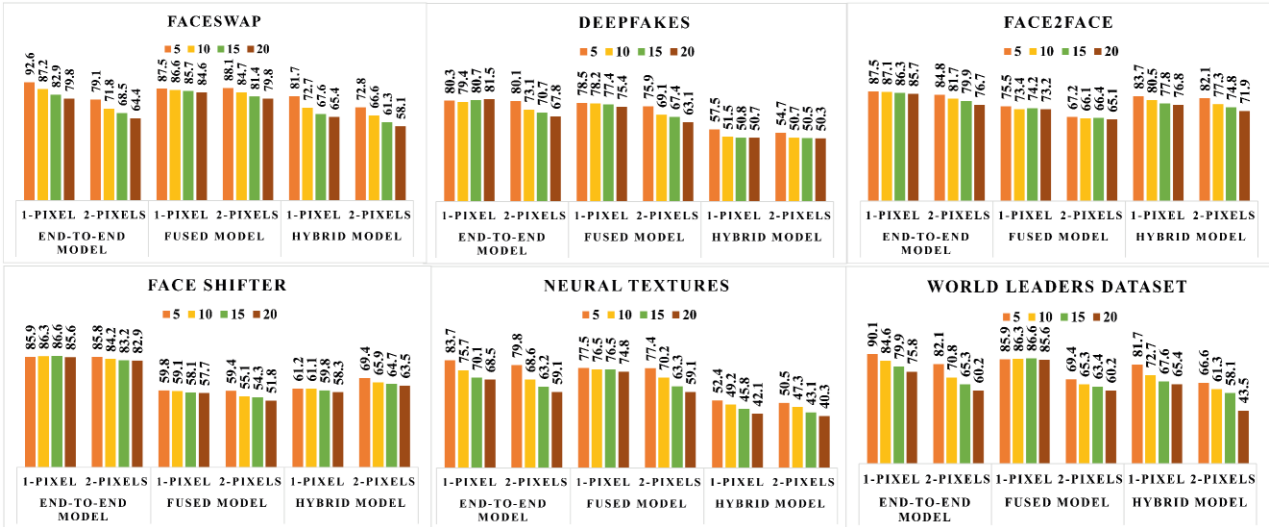

Figure 3. Performance evaluation of proposed attack on the target models using FF++ (all subsets) and WLDR (combined).

framework shows promising results in detecting deepfake videos generated using different techniques.

3.2.3 Fusion-based target model

The last target model we selected for the experiment is fusion-based architecture [25] to increase overall performance constituting fused DenseNet. Two simultaneous DenseNet models were trained, and the results of both networks were fused. The fused deep learning models can increase the accuracy and resilience of their predictions. For this experiment, we chose DenseNet due to its high performance after VGG-16 in deepfake detection [23].

3.3 Evaluation parameters

We employed attack success rate (ASR) and accuracy (Acc) evaluation metrics to assess the effectiveness of our proposed attack. These metrics were used to test the aforementioned models with the datasets. These standard measures were also used by the comparative methods [10, 11].

3.4 Performance evaluation of proposed attack

To thoroughly evaluate the effectiveness of our proposed attack in bypassing contemporary deepfake detectors, we conducted several experiments. Each of the proposed attacks with 1-pixel and 2-pixel sizes is added to each frame of a test set with a different quantity. The accuracy of each model drops significantly from adding 5, 10, 15, and 20 moles per frame, respectively. The result of each model after the attack is given in Figure 3 and details of the attack on each deepfake detector are given below.

3.4.1 End-to-end deep learning target model

In the first experiment, we thoroughly evaluated the proposed attack's effectiveness on an end-to-end VGG-16 deep learning model trained on each subset of the FF++ dataset. This model is pre-trained on ImageNet weights and retrained on FF++ data on each of its subsets separately. This model attained the highest accuracy of 96.5% on the face swap subset. In a 5 mole 1-pixel attack, the accuracy drops slightly but after adding more pixels, we observed a significant drop in accuracy. The proposed attack drops the accuracy from 96.5% (without attack) to 64.4% on a 20 moles 2-pixel attack. The proposed attack achieved remarkable results with an ASR of 32.6% compared to the actual test accuracy achieved on the face swap subset of the end-to-end deep learning VGG-16 model. The lowest drop in accuracy was achieved on the neural textures' subset, which was reported as 59.1% from 90.5%. On the WLDR dataset, proposed attack drops the accuracy from 93.6% (without attack) to 60.2% and ASR of 35.6%. This model is susceptible to attacks due to its linear decision boundary and extreme sensitivity even on minor input data modifications.

3.4.2 Hybrid target model

In this experiment, we tested perturbed samples with InceptionResNet-BiLSTM [4]. Each perturbed subset of the FF++ dataset is tested with its trained model respectively. This model achieved the highest accuracy of 93.3% on the deepfake subset with unperturbed samples, however, its accuracy dropped to 58.1% on (2-pixel, 20 moles), with 36.9% ASR. This model attained the lowest accuracy even without perturbation on neural textures because its algorithm only modifies the mouth region to alter the expression of a person, resulting in imperceptible changes that make it difficult to detect the manipulation. On neural textures, this model drops the accuracy significantly from 78.67% to 40.3%. On the WLDR dataset, the proposed attack drops the accuracy from 94.2% (without attack) to 60.2% and ASR of 47.2%.

3.4.3 Fusion-based target model

In this experiment, we trained the fusion-based DenseNet model with perturbed samples. The model attained the highest accuracy of 95.7% on the unperturbed samples in the deepfake subset, but its accuracy significantly decreased to

Table 1. Transferability analysis of Acc% and ASR% before and after attack. Min Acc and Max ASR is w.r.t 2-pixel, 20 moles.

Target model		End-to-end deep learning model						Hybrid model						Fused model					
Datasets		FS	DF	F2F	SH	NT	WL DR	FS	DF	F2F	SH	NT	WL DR	FS	DF	F2F	SH	NT	WL DR
Before Attack	Test Acc	96.5	95.0	94.3	94.5	90.5	93.6	93.0	93.3	92.1	84.9	78.6	94.2	95.7	93.9	92.6	60.9	83.5	91.5
After Attack	Min Acc	64.4	67.8	76.7	82.9	59.1	60.2	58.1	50.3	71.0	58.3	40.3	49.7	79.8	63.1	65.1	51.8	59.1	60.3
	Max ASR	33.3	28.6	18.7	12.3	34.7	35.6	37.5	46.1	22.9	31.3	48.7	47.2	16.6	32.8	29.7	14.9	29.2	34.0

67.4% on samples with perturbations of (2-pixel, 20 moles), with an ASR of 28.3%. The accuracy of this model was even lower on the face shifter subset 60.9%, without any perturbation, as the generative method used in this set is highly complex. On the face shifter subset, the highest accuracy loss of 51.8% was attained. On the WLDR dataset, the proposed attack drops the accuracy from 91.5% (without attack) to 60.3% and ASR of 43.5%. Figure 3 shows the results of trained target models with perturbed instances. The proposed attack modifies a small number of pixels within the ROI, making it effective and difficult to detect. The perturbation disrupts the ROI, leading to misclassification of class labels. The attack is transferable to different deep-learning models, making all network types susceptible.

3.5 Analysis and Discussion

In this section, we have done several analysis, which include the transferability of the attack on several deepfake detectors, explainability and qualitative analysis of the surrogate model, and robustness analysis of the attack. The details of each analysis are given in subsequent sections.

3.5.1 Transferability analysis

To check the transferability of the proposed adversarial instance, we conducted a transfer attack and evaluated adversarial samples against several detection algorithms (including end-to-end deep learning, hybrid, and fused models) in a black-box environment. Table 1 demonstrates that the model attention distraction [18] considerably enhanced the transferability of adversarial perturbations across various detection methods. The highest attack success rate of 48.7% was achieved when attempting to perturb the hybrid deep learning model. Table 1 presents the before and after attack accuracy (min) and maximum ASR (after attack) across several deepfake detectors, whereas Figure 3 depicts examples of unnoticeable and perceptual but visually natural mole attack adversary faces generated by the proposed attack.

3.5.2 Explainability analysis

For the explainability analysis of target models, we employed a surrogate model to quantify the similarity between heat maps of real and deepfake-generated facial frames. From the result of the surrogate model, it can be seen in Figure 1 that generated heat maps highlight crucial face artifacts. On the assumption that different DNNs exhibit similar attention patterns [18]. The heatmaps help to successfully launch an adversarial attack on all deep neural networks. By exploiting this attack, we were able to divert the model's attention from certain features, resulting in incorrect predictions for a specific class label.

3.5.3 Qualitative analysis

A qualitative analysis was conducted on existing blur, noise, one-pixel differential evolution, and one-pixel simulated annealing attacks to compare their effectiveness. Noise is typically more effective than blurring in post-processing attacks, as it reduces video artifacts and makes it more challenging for detectors to differentiate between real and false videos. However, blurring may not be as effective and may generate artifacts that the deepfake detector can detect. The success of these attacks depends on the amount and type of noise or blur introduced. Existing one-pixel attacks were computationally expensive and limited by sensitivity to the initial starting point of the optimization process. The proposed attack is naturally perceptible and only perturbs the facial areas emphasized by the detector. Noise and blurring are scattered across the entire frame sequence, making it more perceptible and easily defended. Our method perturbs instances only within the ROI, resulting in the highest success rate for each subset of the FF++ and WLDR dataset compared to other attacks. We present the visual analysis of our attack and the other existing attacks in Figure 4.

3.6 Comparative analysis

To assess the effectiveness of our proposed attack, we conducted a comparison against state-of-the-art techniques. Several post-processing attacks [26-28] and adversarial attacks [29, 30] on deepfake detectors have employed noise and blur as potent means of attack. These attacks were executed varying in parameters and intensities, such as various kernel sizes including 5, 7, and 9, as seen in previous work [27], to diversify the testing data. By utilizing varying intensities, these attacks could either be noticeable or imperceptible to the human eye. However, the existing attacks [10, 11] were performed under a white-box setting but we performed these attacks in a black-box setting to demonstrate the effectiveness of the proposed attack.

3.6.1. Quantitative analysis

To analyze the robustness of our attack, we applied Gaussian blurring, noise, and one-pixel [10, 11] attacks to each instance of the test set and created its perturbed instance. It is important to note that these two attacks [10, 11] were performed on the entire facial frame. In this experiment, we

Table 2. Comparison with state-of-the-art methods.

Datasets	Acc% (Before Attack)	Attacks	Acc% (After Attack)	ASR%
Faceswap	**96.5**	Blur	89.4	7.4
		Noise	83.7	13.3
		[10]	92.2	4.5
		[11]	91.7	5.0
		Proposed	**64.4**	33.3
Deepfake	**95.0**	Blur	90.1	5.2
		Noise	85.5	10.0
		[10]	91.4	3.8
		[11]	92.3	2.8
		Proposed	**67.8**	28.6
Face2Face	**94.3**	Blur	89.5	5.1
		Noise	82.7	12.3
		[10]	90.3	4.2
		[11]	90.1	4.5
		Proposed	**76.7**	18.7
Face Shifter	**94.5**	Blur	88.5	6.3
		Noise	85.7	9.3
		[10]	91.8	2.9
		[11]	90.3	4.4
		Proposed	**82.9**	12.3
Neural Textures	**90.5**	Blur	77.4	14.5
		Noise	73.6	18.5
		[10]	87.2	3.6
		[11]	88.5	2.2
		Proposed	**59.1**	34.7
WLDR	**93.6**	Blur	88.9	5.0
		Noise	81.3	13.1
		[10]	89.5	4.3
		[11]	90.9	2.0
		Proposed	**60.2**	35.6

selected only the best deepfake detector, which is an end-to-end deep learning model, and tested each trained subset model with perturbed instances of noise and blur. Existing noise and blur attacks are perceptual and easy to detect, while differential evaluation one-pixel attacks are computationally costly and sensitive to the initial starting point of the optimization process, requiring multiple attempts to succeed. Our results, presented in Table 2, clearly demonstrate that our proposed attack (2-pixel, 20-moles) achieved the highest ASR compared to the other attacks. Thus, we claim that our attack is transferable and can be successfully launched to fail the existing deepfake detectors.

3.7 Adversarial training as defense

This experiment is designed to evaluate the effectiveness of defensive techniques against a proposed attack. We used adversarial training as a defensive technique. As the noise closely resembles the proposed mole attack, 25% of the total training dataset for each subset is modified by introducing salt and pepper noise. The end-to-end deep learning model is trained on the newly created dataset to defend against a 2-pixel, 20-mole attack. The results of the defense are given in

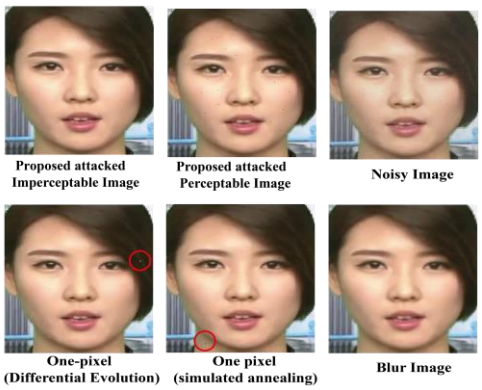

Figure 4. Proposed attack and existing attack samples.

Table 3, although the defense accuracy has improved slightly compared to accuracy after an attack, still it is less compared to the model's actual accuracy. The fact that the prosed attack was still able to fool the model demonstrates that this defense mechanism has limited effectiveness. The experiment demonstrates the need for stronger defense mechanisms to counter the proposed attack.

4. Conclusion

In this study, we have proposed a novel and perceptible mole-based black-box attack that is capable of fooling deepfake detectors. This perceptible visually natural-looking facial mole poses a substantial threat to current advanced deepfake detection systems. Our transferable attack can be applied to every video test set frame, evading multiple deepfake detectors. Our proposed attack is effective in evading state-of-the-art deepfake detectors. This highlights the urgent need for more powerful and adversary-resistant detection systems that can better combat the proliferation of deepfakes. Although our proposed attack looks visually natural, it is still perceptible, therefore future research in this area should focus on creating imperceptible and more powerful attacks. Additionally, there will be a focus on developing detection models that can withstand various forms of adversarial attacks and enhance the overall robustness of deepfake detection systems.

Table 3. Defense against proposed attack.

Datasets	Acc% (Before Attack)	Acc% (AfterAttacks)	Defence Acc%
Faceswap	96.5	64.4	77.8
Deepfake	95.0	67.8	74.2
Face2Face	94.3	76.7	82.3
Face Shifter	94.5	82.9	87.9
Neural Textures	90.5	59.1	69.3
WLDR	93.6	60.2	83.5

REFERENCES

[1] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M, "Faceforensics++: Learning to detect manipulated facial images", In Proceedings of the

IEEE/CVF international conference on computer vision (pp. 1-11), 2019.

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

[3] Javed, A., & Malik, K. M, "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, 2022.

[4] Ilyas, H., Irtaza, A., Javed, A., & Malik, K. M, "Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection", In 2022 16th international conference on open source systems and technologies (ICOSST) (pp. 1-6). IEEE, 2022.

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

[6] Liang, B., Wang, Z., Huang, B., Zou, Q., Wang, Q., & Liang, J, "Depth map guided triplet network for deepfake face detection", Neural Networks, 159, 34-42, 2023.

[7] Heo, Y. J., Yeo, W. H., & Kim, B. G, "Deepfake detection algorithm based on improved vision transformer", Applied Intelligence, 53(7), 7512-7527, 2023.

[8] Cao, X., & Gong, N. Z, "Understanding the security of deepfake detection", In International Conference on Digital Forensics and Cyber Crime (pp. 360-378). Cham: Springer International Publishing, 2021.

[9] Wang, R., Juefei-Xu, F., Guo, Q., Huang, Y., Xie, X., Ma, L., & Liu, Y, "Amora: Black-box adversarial morphing attack", In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1376-1385), 2020.

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

[11] Zhou, Tianxun, Shubhankar Agrawal, and Prateek Manocha, "Optimizing One-pixel Black-box Adversarial Attacks", arXiv preprint arXiv:2205.02116, 2022.

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

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

[14] (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.

[15] Lim, N. T., Kuan, M. Y., Pu, M., Lim, M. K., & Chong, C. Y, "Metamorphic testing-based adversarial attack to fool deepfake detectors", In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 2503-2509). IEEE, 2022.

[16] Carlini, N., & Farid, H, "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), 2020.

[17] Wang, X., Ni, R., Li, W., & Zhao, Y, "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, 2021.

[18] Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., & Liu, X, "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), 2021.

[19] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y, "Joint face detection and alignment using multitask cascaded convolutional networks", IEEE signal processing letters, 23(10), 1499-1503, 2016.

[20] Hentrich, M, "Methodology and coronary artery disease cure", Available at SSRN 2645417, 2015.

[21] Zatorre, R. J., Mondor, T. A., & Evans, A. C, "Auditory attention to space and frequency activates similar cerebral systems", Neuroimage, 10(5), 544-554, 1999.

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

[23] Taeb, M., & Chi, H, "Comparison of deepfake detection techniques through deep learning", Journal of Cybersecurity and Privacy, 2(1), 89-106, 2022.

[24] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A, "Deepfake video detection through optical flow based cnn", In Proceedings of the IEEE/CVF international conference on computer vision workshops (pp. 0-0), 2019.

[25] Khalid, F., Javed, A., Irtaza, A., & Malik, K. M, "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, 2023.

[26] Pu, J., Mangaokar, N., Wang, B., Reddy, C. K., & Viswanath, B, "Noisescope: Detecting deepfake images in a blind setting", In Proceedings of the 36th Annual Computer Security Applications Conference (pp. 913-927), 2020.

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

[28] Fan, L., Li, W., & Cui, X, "Deepfake-image anti-forensics with adversarial examples attacks", Future Internet, 13(11), 288, 2021.

[29] Peng, B., Fan, H., Wang, W., Dong, J., Li, Y., Lyu, S., & Zhuang, W, "Dfgc 2021: A deepfake game competition", In 2021 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-8). IEEE, 2021.

[30] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples", In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 3348-3357), 2021.