IDGuard: Robust, General, Identity-centric POI Proactive Defense Against Face Editing Abuse

Yunshu Dai[1] Jianwei Fei[2] Fangjun Huang[1,3*]
[1]Sun Yat-sen University, the School of Cyber Science and Technology
[2]Nanjing University of Information Science and Technology
[3]Guangdong Province Key Laboratory of Information Security Technology

Abstract

In this work, we propose IDGuard, a novel proactive defense method from the perspective of developers, to protect Persons-of-Interest (POI) such as national leaders from face editing abuse. We build a bridge between identities and model behavior, safeguarding POI identities rather than merely certain face images. Given a face editing model, IDGuard enables it to reject editing any image containing POI identities while retaining its editing functionality for regular use. Specifically, we insert an ID Normalization Layer into the original face editing model and introduce an ID Extractor to extract the identities of input images. To differentiate the editing behavior between POI and nonPOI, we use a transformer-based ID Encoder to encode extracted POI identities as parameters of the ID Normalization Layer. Our method supports the simultaneous protection of multiple POI and allows for the addition of new POI in the inference stage, without the need for retraining. Extensive experiments show that our method achieves 100% protection accuracy on POI images even if they are neither included in the training set nor subject to any preprocessing. Notably, our method exhibits excellent robustness against image and model attacks and maintains 100% protection performance when generalized to various face editing models, further demonstrating its practicality.

1. Introduction

Face editing technologies, coupled with user-friendly mobile applications, have revolutionized the creation process of face editing. Persons-of-Interest (POI) such as national leaders bear the brunt because fake videos of them can trigger polarized social discourse, and even incite crime [11].

To curb such malicious use, many works have been conducted [3, 7, 14, 15]. Considering the peculiarity of face editing compared to other types of forgery (the many-to-one relationship between face images and identity), we classify existing methods into four categories based on identity relevance and application stage, as shown in Table 1.

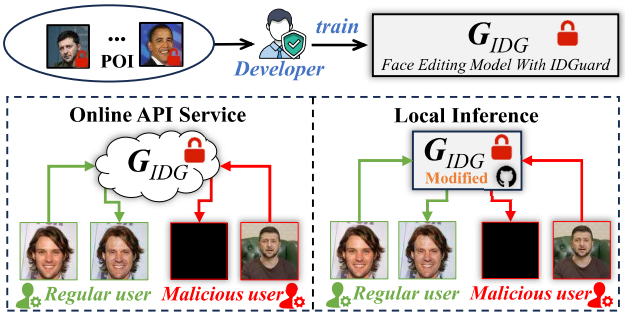

Figure 1. IDGuard enables face editing models to proactively reject editing any image containing the protected POI identities. Whether users utilize the online API service or download it for local inference, regular requests will be operated normally while any malicious attempt will only result in a black image.

Table 1. Classification of existing methods.

	Identity-unrelated	**Identity-related**
Post-event	passive detection	behavior reference
Pre-event	adversarial attack	ours

Category I: methods applied post-event and identity-unrelated, usually focus on visual defects [21, 30], motion artifacts [22, 31], multimodal mismatch [43, 45] and so on. This category has been extensively researched and attained good detection accuracy. However, they cannot prevent the occurrence of malicious editing and are prone to fail when encountering more visually convincing images. **Category II:** methods applied post-event and identity-related, typically construct behavioral references of specific persons, such as voice, face expressions, and head postures [1, 19], and then they detect face editing by identifying anomalies in these habit models. However, these methods need to

*Corresponding author.

model each figure separately and detection performance is susceptible to accidental behavior. Similarly, they are also unable to resist the occurrence of malicious editing. **Category III:** methods applied pre-event and identity-unrelated, usually need to process protected images using adversarial attack [17, 25, 46]. That means if the model encounters images that have not undergone such processing, all defense measures will inevitably fail. **Category IV:** methods applied pre-event and identity-related have not been explored.

However, **Category IV** possesses immense research value in terms of industry development. Face editing models, as well as other generative models, are moving towards a path of regulation and responsibility. The public expects AI companies to release their products in a responsible manner [6, 9]. But in reality, developers have limited control over user behavior, especially in cases where the models are open source, and it is also costly for developers to monitor the usage of API (Application Programming Interface) [44]. Hence, it is urgent to enable face editing models to proactively reject malicious use.

To pioneer in **Category IV**, we propose IDGuard, a robust, general, identity-centric POI proactive defense method. As shown in Fig. 1, we consider two scenarios for the open-source released models: whether users utilize the online API service or download it for local inference, regular requests will be operated normally, while any malicious attempt on protected POI will only result in a black image.

To the best of our knowledge, we are the first to propose an identity-centric proactive defense method from the perspective of developers to counteract face editing abuse. Our contributions can be summarized as follows:

- We propose IDGuard, a novel identity-centric POI proactive defense method, safeguarding POI identities rather than merely certain face images. It enables face editing models to retain their original functionality for regular editing while proactively rejecting editing POI images, even if the images are neither included in the training set nor subjected to any specific preprocessing.
- Our method supports simultaneous protection of multiple POI identities and allows for adding new POI identities in the inference stage without retraining.
- Our method also demonstrates strong robustness against image and model attacks. Moreover, our method is architecture agnostic and can be generalized to various face editing models maintaining 100% protection accuracy.

2. Related Work

In this section, we introduce existing identity-related face editing detection methods, which have the same objective as ours, i.e., to protect specific identities. We also introduce proactive defense methods, which align with our motivation to prevent the occurrence of malicious face editing.

2.1. Identity-Centric Face Editing Detection

These methods target specific identities, such as celebrities, and determine the authenticity of suspicious content by learning biometric style features, including face appearances [1, 19], expressions [3, 7], and audio [26, 35].

Individual differences in face expressions have been proven to be unique and stable over time [13] and can be used to recognize identities, providing a basis for biometric identity-specific face editing detection. Agarwal *et al.* [2] learn the identity-related behavioral reference by modeling both face and head movements, forming unique behavioral patterns to distinguish real videos from fake ones. Taking it a step further, Agarwal *et al.* [3] introduce the person-specific correlation between face expressions and speech patterns to improve the accuracy of authenticity verification.

To incorporate more identity-related factors, Boháček *et al.* [7] present an identity-based method by learning unique face, gesture, and vocal habits. Considering the continuity of face motion in videos, Boháček *et al.* [1] utilize temporal consistency of face appearance and movements to detect fake videos for specific individuals. Moreover, Cozzolino *et al.* [15] train a temporal network to separate three-dimensional morphable model (3DMM) features of specific identities. The 3DMM features contain less but more robust identity information, empowering the detection in more generalized face editing scenarios. And in [14], Cozzolino *et al.* realize robust detection of single- and multi-modality face editing by audio-visual features that capture the distinctive traits of a specific individual. More recently, Tian *et al.* [40] release a large-scale POI benchmark to advance research on identity-centric face editing detection, which directly shifts the focus of face editing detection onto the identification of POI. However, all these methods are unable to prevent the occurrence of malicious editing.

2.2. Proactive Defense Against Face Editing

In recent years, the vulnerability of generative models to adversarial attacks has attracted great attention [5, 27, 38]. In light of this, methods employing adversarial attacks as a proactive defense have been widely explored. These methods typically attack face editing models by adding imperceptible adversarial noise to make the output disrupted [4, 17, 18, 25, 36, 37, 41, 46]. Dong *et al.* [18] propose both image-agnostic and image-specific adversarial perturbation attacks, which can effectively disrupt the output of face swap models. In contrast, Sun *et al.* [37] employ adversarial attacks on face landmarks to disrupt the face localization of face swap models. More recently, Huang *et al.* [25] propose a Cross-Model Universal Adversarial (CUMA) watermark, which enables the attack across multiple face editing models. To tackle the problem of extensive querying in both white-box [36, 41] and black-box [25] attacks in current studies, Dong *et al.* [17] propose a

query-free adversarial attack method. They employ a Transferable Cycle Adversary GAN (TCA-GAN), which utilizes a substitute model to generate adversarial samples, enabling transferability to inaccessible black-box models. However, these methods require the processing of every protected image. Once an unprocessed image is encountered, the defense measures will inevitably fail.

Unlike all methods mentioned above, our method takes an innovative perspective from the developers of face editing models to enable them to proactively prevent any malicious attempt on specific identities, thus nipping fake images in the bud. Our method not only has the advantages of 'identity-centric'and 'proactive defense', but also eliminates the need for any preprocessing of input images. Furthermore, our method safeguards all images containing the protected identity, rather than selectively protecting certain images. In terms of practical application and development costs, our method exhibits superior performance.

3. Methodology

3.1. Threat Model

Our work aims to provide developers with a feasible and efficient solution to proactively prevent malicious editing of POI, thus helping them responsibly release their face editing models. In the scenarios we consider, the involved three entities are as follows:

(a) Developers: This role possesses full control of the model architecture, training strategy, and dataset (including the selection of POI). Developers will release online API or open-source models to provide face editing services.

(b) Regular users: This role only utilizes the provided service to edit ordinary faces that do not involve any POI.

(c) Malicious users: This role may attempt to edit POI images through the provided service, and even worse, once they find the POI editing failed, they will turn to undermine the protective functionality of the model itself.

3.2. Overview

Given a target face editing model *G*, we aim to equip it with IDGuard to enable it to proactively reject editing any image belonging to the protected POI identities selected by the developer. Thus the primary goal is to enable the model to perceive the identity of input images and behave accordingly, i.e., either edit or reject. The next goal is to support the protection of multiple POI, and considering the practical needs in reality, it is also essential to allow for adding new POI in the inference stage. Furthermore, IDGuard should be able to resist various potential attacks to maintain its robust protection functionality when the model is released. After careful planning on how to tackle these problems, we have developed a training pipeline for the model, as shown in Fig. 2. For the sake of clarity, we will separately introduce the training stage and inference stage of our method.

3.3. In the Training Stage

We denote two datasets I^{POI} and I^{non} for POI and non-POI respectively. I^{POI} consists of M images denoted as $I^{POI} = \left\{\left(X_i^{POI}, Y_i\right)\right\}_{i=1}^{M}$ and I^{non} consists of N images denoted as $I^{non} = \left\{\left(X_j^{non}, Y_j\right)\right\}_{j=1}^{N}$. Here, X represents the image, while Y represents the corresponding identity label. The target face editing model is denoted as *G*, the discriminator is denoted as *D* and the face editing model with IDGuard is denoted as G_{IDG}. In each training iteration, we sample 2 batches of unprotected identities from the non-POI dataset and 1 batch of protected identities from the POI dataset, denoted as I_1^{non}, I_2^{non}, and I_1^{POI}.

Problem 1: How to enable G_{IDG} to perceive the identity of input images?

I_1^{non} is used to train an ID Extractor (details are shown in Fig. 8, Appendix Sec. 8). For each input image, we extract features from intermediate G_{IDG} and feed them into the ID Extractor. We make I_1^{non} be a triplet $\left\{X_a^{non}, X_p^{non}, X_n^{non}\right\}$ where a, p, and n denote anchor, positive and negative with the first two having the same identity while the third one differs. To ensure the ID Extractor captures identity-related features, we employ Eq. 1 to maximize the distance between features of different identities while minimizing the distance between features of the same identity.

$$\begin{aligned}\mathcal{L}_{E_{id}} = &\max(0, d(E_{id}(X_a^{non}), E_{id}(X_p^{non})) \\ &- d(E_{id}(X_a^{non}), E_{id}(X_n^{non})) + margin),\end{aligned} \tag{1}$$

where E_{id} denotes the ID Extractor, and $d(\cdot)$ is the L_2 distance. For the hyper-parameter $margin$, we set it to 2.0 by default. Note that I_1^{non} is exclusively used to train the ID Extractor, without being passed through the downstream of G_{IDG} to generate any image output.

Considering that the optimization of the ID Extractor is a gradual process during the initial stage of training, we adopt the Exponential Moving Average (EMA) to iteratively update the identity features when I_1^{POI} is fed in, aiming for better precision. For any sample $(X_i^{POI}, Y_i^{POI}) \in I_1^{POI}$ and training step $t+1$, we have

$$f_i^{t+1} = \gamma \cdot f_i^t + (1-\gamma) \cdot E_{id}(X_i^{POI}), \tag{2}$$

where f_i^t is the current identity feature of identity Y_i^{POI} of step t, and γ represents the decay factor which determines the weights given to the most recent observations and the past ones when the model calculates the moving average. Here, we set γ to be 0.99. EMA ensures that features extracted from each identity become more stable and precise during the training process. Note that when the training is finished, the ID Extractor is capable of directly extracting accurate identity features from images.

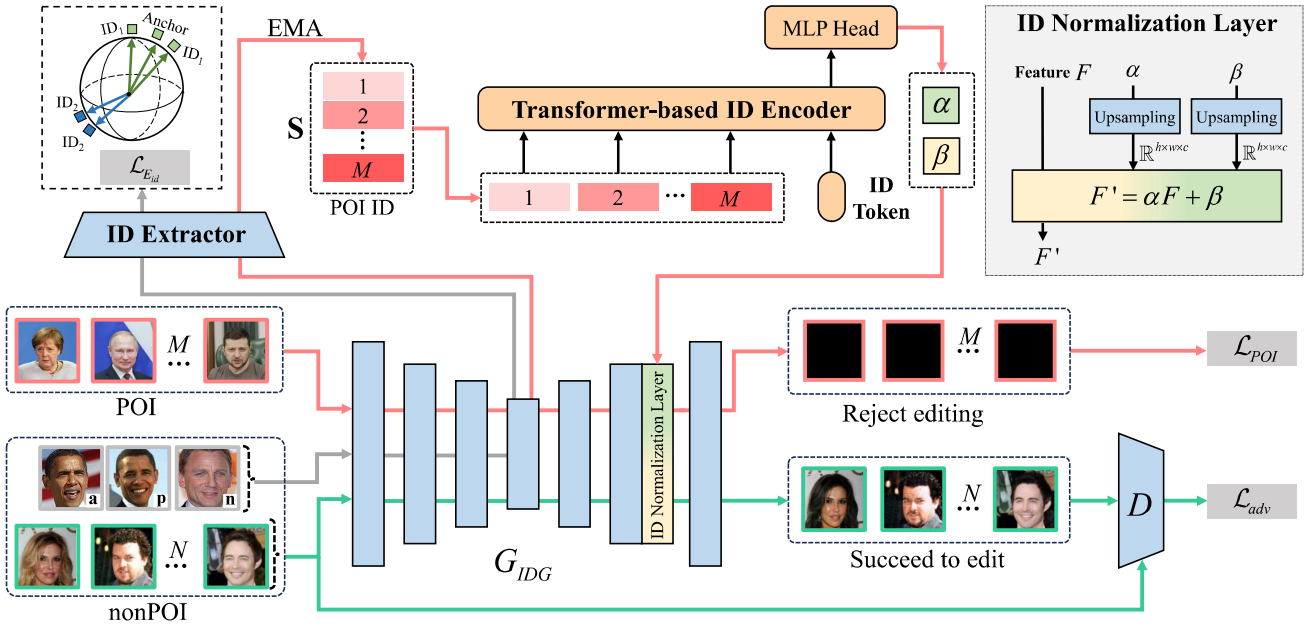

Figure 2. Training pipeline of G_{IDG} (IDGuard incorporated with a face editing model G). We introduce an ID Extractor to extract POI identities and utilize a transformer-based ID Encoder to encode identities into the parameter of the ID Normalization Layer (α and β). For each iteration, we sample three batches for the training of the ID Extractor, face editing functionality, and POI protection, respectively.

Problem 2: How to maintain regular face editing functionality on images of unprotected identities?

I_2^{non} is used as training data for G_{IDG} to complete the regular face editing task. For simplicity, we denote the optimization goal by a standard adversarial loss Eq. 3. We assume the face editing is conditional on label c, which can be an attribute label or identity, according to different tasks.

$$\begin{aligned}\mathcal{L}_{adv} = &\mathop{\mathbb{E}}_{X\sim I_2^{non}} \log(D(X)) \\ &+ \mathop{\mathbb{E}}_{X,c\sim I_2^{non},C} \log(1 - D(G(X,c))),\end{aligned} \tag{3}$$

where C is the label set. This ensures that G_{IDG} maintains regular face editing functionality on unprotected identities.

Problem 3: How to ensure G_{IDG} protects multiple identities simultaneously?

To facilitate the simultaneous handling of these identities, we introduce a transformer-based ID Encoder that is able to process sequences because it can convert an arbitrary number of features into a constant-length output, which aligns with our demand. Let us assume that there are M identities to be protected in total, and a batch I_1^{POI} consists of m identities (number m varies for each iteration). When I_1^{POI} is fed into G_{IDG}, the intermediate features are used as the input of the ID Extractor to obtain m ID features. We package the m extracted ID features into a sequence, denoted as $\mathbf{S}$ that does not necessitate a specific order. The ID Encoder consists of L multi-head attention layers and a two-head MLP (Multilayer Perceptron), it takes ID features $\mathbf{S}$ as input and outputs two embeddings that are used as the parameters of the ID Normalization Layer.

Problem 4: How to ensure the extracted POI identities dominate downstream model behavior?

We need to build a bridge between identities and model behavior. Since the perception of identity is determined by parameters, we map identity features into model parameters to make G_{IDG} respond differently to POI and nonPOI identities. Specifically, we design an ID Normalization Layer and insert it into the later stage of G, which is the only addition for G_{IDG} at the architectural level in comparison to G. Note that the ID Normalization Layer is trained together with the entire model G_{IDG}, and even after encountering removal attacks, it is different from the original model G. We use the output of ID Encoder as the embedding encompassing all the POI identities. This embedding is then passed through a two-headed multi-layer perceptron (MLP) and upsampled, yielding two parameter sets, denoted as α and β. The ID Normalization Layer uses α and β as parameters to normalize the feature F coming from upstream layers to F' according to Eq. 4

$$F' = \alpha F + \beta. \tag{4}$$

For batch I_1^{POI}, we force G_{IDG} to ultimately generate a black output as a result, hence employing a loss function Eq. 5 to minimize the distance between the output and a black image B of the same size with the output $G(X,c)$.

$$\mathcal{L}_{\text{POI}} = \mathop{\mathbb{E}}_{X,c\sim I_1^{POI},C} ||B - G(X,c)||_2^2. \tag{5}$$

Therefore, the total training loss of G_{IDG} is:

$$\mathcal{L}_{\text{IDGuard}} = \lambda_1 \mathcal{L}_{E_{id}} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{\text{POI}}, \quad (6)$$

where λ_1, λ_2, and λ_3 are weights balancing the three terms.

3.4. In the Inference Stage

Please note that upon entering the inference stage, (1) the ID Extractor is capable of extracting precise identity features, and (2) G_{IDG} is capable of performing normal face editing on nonPOI while proactively reject editing POI.

Problem 5: How to add new POI identities when the training of G_{IDG} is finished?

The perception of POI identities of G_{IDG} is determined by the parameters in the ID Normalization Layer. Updating the parameters corresponds to updating the model's perception of POI. If developers want to protect new POI identities when the training is finished, they only need to include images of new POI identities to the original POI dataset and feed the newly merged POI dataset into the trained ID Extractor to form a new ID sequence $\mathbf{S}$. Then the new $\mathbf{S}$ is fed into the ID Encoder to generate new parameters α and β for the ID Normalization Layer. By doing so, the newly added identities can be protected. Notably, the inclusion of the new identities only requires one forward pass through the network, without any backpropagation. In other words, retraining is not required and thus the cost for developers to add new identities is minimal.

4. Experiments

4.1. Implementation Details

Datasets. We utilize CelebA [33] and VGGFace2 [8] datasets for training. CelebA serves as the nonPOI dataset containing 10599 identities, each corresponding to about 20 images. We use 1224 identities from VGGFace2 as the POI dataset and each identity has extensively over 200 images. We train G_{IDG} using 1024 POI identities while the remaining 200 POI identities are reserved for adding in the inference stage. To adapt G_{IDG} to the increase of POI number, we randomly select 4 to 24 POI identities from each batch in training. Additionally, we allocate 20 images per POI identity as a separate testing set to evaluate the protection of the trained G_{IDG} model on previously unseen images. More details are shown in Table 8 (Appendix Sec. 7).

Models. To demonstrate the effectiveness, robustness, and generalization of IDGuard, we perform experiments on various face editing models, including StarGAN [12], AttGAN [23], HiSD [32], AGGAN [39], SimSwap [10] and Faceshifter [29]. The first four models are image-to-image translation models capable of conducting cross-domain attribute editing. The last two models are face swap models, which are able to replace the identity of a source image with a target identity. More details are shown in Appendix Sec. 8.

Evaluation Metrics. For fidelity, we use FID (Frechet Inception Distance) [24] to evaluate the quality of face editing on nonPOI images, where a lower FID score indicates better fidelity. For effectiveness, we follow the work of Huang *et al.* [25], and adopt their evaluation metric SR_{mask}, which is determined by the pixel-level discrepancies between the outputs of G_{IDG} and G when fed with POI images, and a higher SR_{mask} score indicates better effectiveness. We also use $\log_{10}$ FID of the generated POI results to measure effectiveness, a higher $\log_{10}$ FID indicates that the model output for POI images is more disrupted.

4.2. Fidelity Analysis

The incorporation of IDGuard in a face editing model should not compromise its editing performance on nonPOI images. That is, G_{IDG} should maintain a similar FID compared to G. Lower FID scores indicate better fidelity.

Table 2. FID ↓ of face editing models G and G_{IDG}.

Models	w/o IDG ↓	with IDG ↓	Diff ↓
StarGAN [12]	5.71	7.24	**1.53**
AGGAN [39]	2.08	3.45	**1.37**
AttGAN [23]	4.87	5.30	**0.43**
HiSD [32]	1.87	2.68	**0.81**
SimSwap [10]	12.66	13.07	**0.41**
FaceShifter [29]	9.85	11.13	**1.28**

Table 2 shows the average FID scores of every face editing model on nonPOI images, with or without IDGuard. The incorporation of IDGuard leads to an average increase of approximately 1 in FID scores, which is practically indistinguishable from human eyes. It indicates that IDGuard exhibits high fidelity and has almost no negative impact on regular users. More visual comparisons are shown in Fig. 10 (Appendix Sec. 11).

4.3. Effectiveness Analysis

The effectiveness of G_{IDG} is measured by the protection effectiveness (accuracy) on arbitrary images containing POI identities, even if these images are neither included in the training set nor subjected to any preprocessing.

Table 3 provides the quantitative results of IDGuard generalized to four different face editing models. We compare them with existing defense methods based on adversarial attacks in terms of perturbation (SR_{mask}) and quality ($\log_{10}$ FID). Our findings are derived from testing on the reserved 20 images per POI identity which G_{IDG} has never seen during training. It is evident that our method exhibits remarkable superiority in comparison.

Fig. 3 displays the output of G_{IDG} (here using StarGAN as an example). The images on the top depict face edit-

Table 3. Comparisons of $SR_{\text{mask}} \uparrow$ and $\log_{10}$ FID $\uparrow$ of our method and the state-of-the-art proactive face editing defense methods using adversarial attacks.

Method	$SR_{\text{mask}} \uparrow$				$\log_{10}$ FID $\uparrow$			
	StarGAN [12]	AGGAN [39]	AttGAN [23]	HiSD [32]	StarGAN [12]	AGGAN [39]	AttGAN [23]	HiSD [32]
BIM [28]	0.6755	0.9975	0.2126	0.0028	1.9047	1.6539	1.2451	1.6098
MIM [20]	**1.0000**	0.9994	0.0200	0.0438	2.5281	1.8435	0.7842	1.5205
PGD [34]	0.8448	0.9970	0.0146	0.0010	2.0203	1.6659	0.9403	1.6467
DI^2-FGSM [42]	0.0280	0.3448	0.0074	0.0001	1.5714	1.3084	1.2036	1.4113
M-DI^2-FGSM [42]	**1.0000**	0.9987	0.0032	0.0050	1.5714	1.3084	1.2036	1.4113
AutoPGD [16]	0.8314	0.9963	0.0002	0.0007	1.5714	1.3084	1.2036	1.4113
CMUA [25]	**1.0000**	0.9988	0.8708	0.9987	2.3032	1.7072	1.8133	1.9672
Ours	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**2.5359**	**2.5402**	**2.5025**	**2.5309**

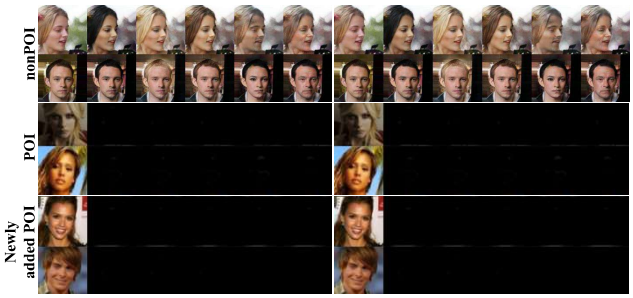

Figure 3. Editing results of G_{IDG} (StarGAN) on unseen images of nonPOI (top), POI (middle), and newly added POI (bottom).

ing performance with five different target labels(age, gender, hair color, and so on) on nonPOI images, while the images on the middle and bottom depict failed face editing outputs of POI and newly added POI. It can be observed that G_{IDG} is capable of generating high-quality results for unprotected identities. However, when it comes to protected identities, it directly outputs black images, effectively eliminating the potential malicious use. We also evaluate IDGuard on face swap models, considering that POI could be either the source identity or the target identity, we separately report the effectiveness for both cases, and the results of IDGuard on SimSwap and FaceShifter are shown in Table 4. We can observe that, regardless of whether the attacker uses the POI as the source or the target identity, the editing attempts are always unsuccessful.

Table 4. $SR_{\text{mask}} \uparrow$ and $\log_{10}$ FID $\uparrow$ on unseen POI images of face swap models with IDGuard.

Faces	Models	$SR_{\text{mask}} \uparrow$	$\log_{10}$ FID $\uparrow$
Target	SimSwap [10]	1.00	2.43
	FaceShifter [29]	1.00	2.53
Source	SimSwap [10]	1.00	2.35
	FaceShifter [29]	1.00	2.52

4.4. Robustness Analysis

In terms of model application, we consider two scenarios: one is interactive online through the API service, and the other is local inference by downloading the model with IDGuard released by developers. We conduct robustness testing for both scenarios.

Image Processing Attacks. In the scenario where users utilize the API, they can only interact with G_{IDG} by inputting images, G_{IDG} may be susceptible to image processing attacks. These attacks can include unintentional image degradation during transmission, like JPEG compression, or deliberate attacks like blurring, noise addition, and brightness jitter. Hence, we conduct experiments to evaluate the protective performance of G_{IDG} under various intensities and types of image processing attacks. For JPEG compression, the quality factors (QF) are set to 10, 30, and 70. For Gaussian blurring, the kernel sizes (KS) are 1, 2, and 3. For Gaussian noise, the mean is 0, while the standard deviations (STD) are 0.05, 0.1, and 0.2. For brightness jitter, we set the multipliers to 1.5, 2.0, and 3.0.

Table 5. SR_{mask} of different face editing models with IDGuard after unseen images of protected identities being subjected to various types of image processing attacks.

Models	JPEG	Blur	Noise	Color
StarGAN [12]	1.00	1.00	1.00	1.00
AGGAN [39]	1.00	1.00	1.00	1.00
AttGAN [23]	1.00	1.00	1.00	1.00
HiSD [32]	1.00	1.00	1.00	1.00
SimSwap [10]	1.00	1.00	1.00	1.00
FaceShifter [29]	1.00	1.00	1.00	1.00

As shown in Fig. 4, it is apparent that identity information of input has survived all image processing attacks, allowing G_{IDG} to persist in its rejection of any POI face editing. Even though Gaussian blurring seems to have a slightly

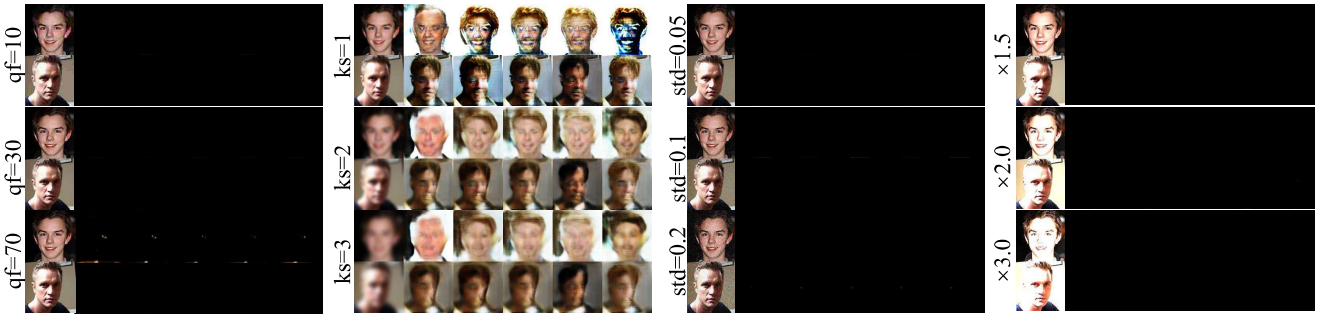

Figure 4. Editing results of G_{IDG} (StarGAN) for unseen images of protected identities after various image processing attacks.

greater impact than other attacks and its corresponding outputs do reveal face patterns, these edited faces remain visually unusable, affirming that IDGuard is still effective.

Table 5 presents the quantitative results on different face editing models, with SR_{mask} as the metric, where each attack corresponds to the maximum intensity in Fig. 4. We can see that under any case, IDGuard achieves 100% protection accuracy, showcasing its superior effectiveness.

Model Modification Attacks. In the scenario where G_{IDG} has been downloaded by users for local inference, it may be vulnerable to white-box model modification attacks. Once attackers become aware of the existence of IDGuard, they may launch attacks against the ID Normalization Layer, such as removal or perturbing its α and β by noise. Therefore, G_{IDG} needs to demonstrate model robustness when confronted with aforementioned attacks.

We initially consider layer removal attacks. The results in Fig. 5 show that removing the ID Normalization Layer may produce non-completely black images, but the image quality is significantly decreased. Although G_{IDG} loses its identity protection ability after the removal attack, it is still visually unusable.

We then consider the noise addition attack. Attackers may add noise to α and β, in order to disrupt identity encoding and thus undermine the POI protection capability. Fig. 6 shows the face editing results on both POI and non-POI identities after the noise addition attack with different

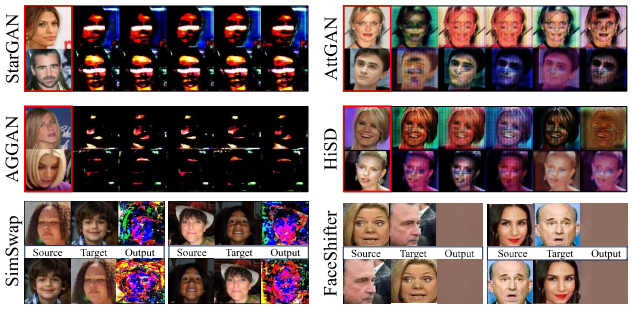

Figure 5. Editing results of G_{IDG} for POI after removal attack.

standard deviations. It can be observed that even when the added noise intensity is strong enough to destroy the original face editing functionality, G_{IDG} still produces almost black outputs for POI.

Adversarial Attack. Fig. 7 shows some POI outputs after varying degrees of adversarial attacks on the ID Extractor. Extractor. We can see adversarial attacks may result in outputs that are not black, but still heavily perturbed, which indicates that IDGuard is robust to adversarial attack.

4.5. Ablation Studies

Number of Newly Added POI in the Inference Stage. Given the multitude of influential persons in the real world, a crucial aspect of IDGuard is the protection capacity. It not only requires the model to simultaneously protect multiple POI but also demands that the model can add new POI after training, with a preference for higher numbers while ensuring model effectiveness.

We evaluate the impact of the number of newly added POI identities on the effectiveness, as shown in Table 6. Please note that these POI identities are added during the inference stage, and the protection functionality is achieved by updating the parameters of the ID Normalization Layer solely through network feed-forward, without retraining.

Table 6. SR_{mask} of different face editing models with different number of protected identities.

Models \ Identities	10	50	100	200
StarGAN [12]	1.00	1.00	1.00	1.00
AGGAN [39]	1.00	1.00	1.00	1.00
AttGAN [23]	1.00	1.00	1.00	1.00
HiSD [32]	1.00	1.00	1.00	1.00
SimSwap [10]	1.00	1.00	1.00	1.00
FaceShifter [29]	1.00	1.00	1.00	1.00

Table 6 presents the results obtained from scenarios involving 10, 50, 100, and 200 POI identities. We can see

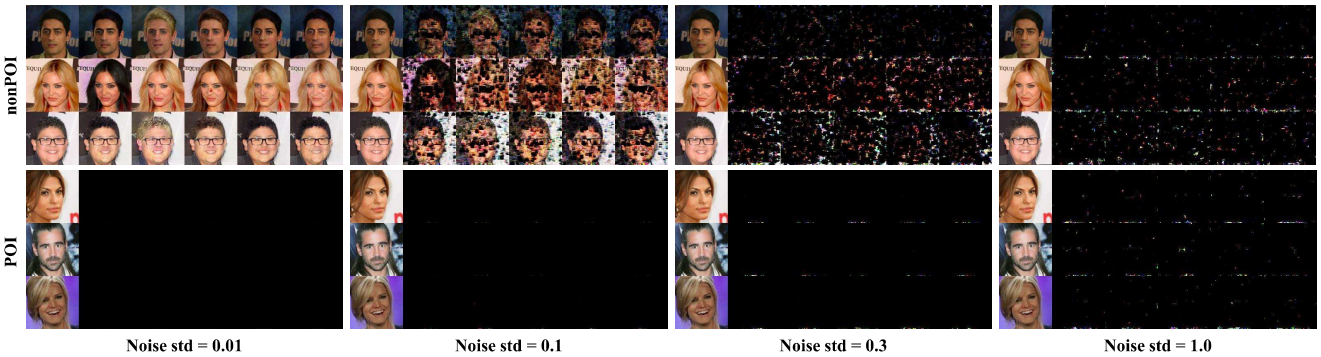

Figure 6. Editing results of G_{IDG} (StarGAN) under noise addition attacks with different intensities.

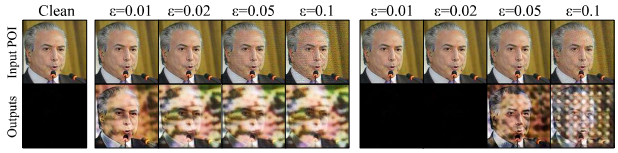

Figure 7. Protection results against white-box FGSM (left) and PGD (right) for different .

from Table 6 that IDGuard can provide 100% protection accuracy even when dealing with a new group of 200 POI, which indicates that the effectiveness of IDGuard is remarkable. Considering the number of world leaders in reality, we assume that the dynamic addition of 200 POI identities in the inference stage, along with the 1024 identities trained in the training stage, would be sufficient to meet the required protection for world leaders in practice.

Training Dataset Requirement. The number of high-quality images specific to protected POI may be limited, so it is worth studying the training dataset size for IDGuard. For this purpose, we establish varying sizes of POI training sets, which comprise different numbers of training images per POI, and evaluate the performance on unseen images. The results are shown in Table 7. It can be observed that IDGuard does not require a large dataset size. For any POI, achieving over 95% protection accuracy only requires 100 images, and when the number of images increases to 180, the protection accuracy can reach 100%. In the real world, it is not difficult to obtain a training dataset of this scale for POI, implying that IDGuard possesses strong practicality.

5. Conclusion

In this work, we propose IDGuard, a novel proactive defense method against face editing abuse from the perspective of developers, safeguarding the identities rather than merely certain face images. IDGuard enables face editing models to proactively reject editing any image containing protected POI identities even if these images are neither included in the training set nor subjected to any preprocessing. Extensive experiments show that our method achieves

Table 7. SR_{mask} of different face editing models with different numbers of training images per POI.

Models \ Images	25	50	100	180
StarGAN [12]	0.17	0.96	1.00	1.00
AGGAN [39]	0.06	0.95	1.00	1.00
AttGAN [23]	0.21	0.96	1.00	1.00
HiSD [32]	0.22	0.99	1.00	1.00
SimSwap [10]	0.18	0.92	0.96	1.00
FaceShifter [29]	0.20	0.85	0.99	1.00

100% protection accuracy for unseen faces of protected identities. Our method supports the simultaneous protection of multiple POI and allows for the addition of new POI in the inference stage without model retraining. It also exhibits excellent robustness against image and model attacks and maintains 100% protection performance when generalized to different face editing models.

Our method provides a new solution for face protection in the era of AIGC (Artificial Intelligence Generative Content). By allowing developers to include POI as whitelisted entities in the model training process, IDGuard can completely eliminate any attempt to edit the POI, thus enabling better protection compared to previous passive forensics and proactive defense methods. In our future work, we aim to offer solutions for standardized management of AIGC as well as compliant usage of face editing and related models. Furthermore, we will delve into extensive research on the generation of malicious forgeries through text-driven and diffusion models. To lighten the burden on developers, we will explore more efficient versions of IDGuard.

Acknowlegdement

This work is supported by the National Natural Science Foundation of China under Grants U2336208 and 62072481.

References

[1] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020. 1, 2

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019. 2

[3] Shruti Agarwal, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, and Anna Rohrbach. Watch those words: Video falsification detection using word-conditioned facial motion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4710–4719, 2023. 1, 2

[4] Shivangi Aneja, Lev Markhasin, and Matthias Nießner. Tafim: Targeted adversarial attacks against facial image manipulations. In *European Conference on Computer Vision*, pages 58–75. Springer, 2022. 2

[5] Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial self-defense for cycle-consistent gans. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 2

[7] Matyáš Boháček and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022. 1, 2

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5

[9] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325*, 2023. 2

[10] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 5, 6, 7, 8, 12

[11] Robert Chesney and Danielle Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, 98:147, 2019. 1

[12] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 5, 6, 7, 8, 12

[13] Jeffrey F Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 491–496. IEEE, 2002. 2

[14] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2023. 1, 2

[15] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021. 1, 2

[16] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 6

[17] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[18] Junhao Dong and Xiaohua Xie. Visually maintained image disturbance against deepfake face swapping. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[19] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. 1, 2

[20] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 6

[21] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20270–20280, June 2022. 1

[22] Jianwei Fei, Zhihua Xia, Peipeng Yu, and Fengjun Xiao. Exposing ai-generated videos with motion magnification. *Multimedia Tools and Applications*, 80:30789–30802, 2021. 1

[23] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 5, 6, 7, 8, 12

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[25] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 989–997, 2022. 2, 5, 6, 12

[26] Pavel Korshunov and Sébastien Marcel. Speaker inconsistency detection in tampered video. In *2018 26th Euro-*

pean signal processing conference (EUSIPCO), pages 2375–2379. IEEE, 2018. 2

[27] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pages 36–42. IEEE, 2018. 2

[28] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 6

[29] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5074–5083, 2020. 5, 6, 7, 8, 12

[30] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 1

[31] Xin Li, Rongrong Ni, Pengpeng Yang, Zhiqiang Fu, and Yao Zhao. Artifacts-disentangled adversarial learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1658–1670, 2022. 1

[32] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8639–8648, 2021. 5, 6, 7, 8, 12

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 6

[35] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Deepfake audio detection by speaker verification. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022. 2

[36] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. 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*, pages 236–251. Springer, 2020. 2

[37] Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu. Landmark breaker: obstructing deepfake by disturbing landmark extraction. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020. 2

[38] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016. 2

[39] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 2021. 5, 6, 7, 8, 12

[40] Lulu Tian, Hongxun Yao, and Ming Li. Fakepoi: A large-scale fake person of interest video detection benchmark and a strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[41] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3546–3555, 2021. 2

[42] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 6

[43] Qilin Yin, Wei Lu, Bin Li, and Jiwu Huang. Dynamic difference learning with spatio-temporal correlation for deepfake video detection. *IEEE Transactions on Information Forensics and Security*, 2023. 1

[44] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry S Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. In *International Conference on Learning Representations*, 2021. 2

[45] Yang Yu, Xiaolong Liu, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C Kot. Pvass-mdd: Predictive visual-audio alignment self-supervision for multimodal deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[46] Yao Zhu, Yuefeng Chen, Xiaodan Li, Rong Zhang, Xiang Tian, Bolun Zheng, and Yaowu Chen. Information-containing adversarial perturbation for combating facial manipulation systems. *IEEE Transactions on Information Forensics and Security*, 18:2046–2059, 2023. 2