CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes

Hao Huang,[1,2] Yongtao Wang,[1,2*] Zhaoyu Chen,[3] Yuze Zhang,[1] Yuheng Li,[1] Zhi Tang,[1,2]
Wei Chu,[4] Jingdong Chen,[4] Weisi Lin,[5] Kai-Kuang Ma[5]

[1]Peking University
[2]State Key Laboratory of Media Convergence Production Technology and Systems
[3]Fudan University
[4]Ant Group
[5]Nanyang Technological University
huanghao@stu.pku.edu.cn, {wyt, kevinzyz, tangzhi}@pku.edu.cn, zhaoyuchen20@fudan.edu.cn,
yuhengli2021@outlook.com, {weichu.cw, jingdongchen.cjd}@antgroup.com, {wslin, ekkma}@ntu.edu.sg

Abstract

Malicious applications of deepfakes (i.e., technologies generating target facial attributes or entire faces from facial images) have posed a huge threat to individuals' reputation and security. To mitigate these threats, recent studies have proposed adversarial watermarks to combat deepfake models, leading them to generate distorted outputs. Despite achieving impressive results, these adversarial watermarks have low image-level and model-level transferability, meaning that they can protect only one facial image from one specific deepfake model. To address these issues, we propose a novel solution that can generate a Cross-Model Universal Adversarial Watermark (CMUA-Watermark), protecting a large number of facial images from multiple deepfake models. Specifically, we begin by proposing a cross-model universal attack pipeline that attacks multiple deepfake models iteratively. Then, we design a two-level perturbation fusion strategy to alleviate the conflict between the adversarial watermarks generated by different facial images and models. Moreover, we address the key problem in cross-model optimization with a heuristic approach to automatically find the suitable attack step sizes for different models, further weakening the model-level conflict. Finally, we introduce a more reasonable and comprehensive evaluation method to fully test the proposed method and compare it with existing ones. Extensive experimental results demonstrate that the proposed CMUA-Watermark can effectively distort the fake facial images generated by multiple deepfake models while achieving a better performance than existing methods. Our code is available at https://github.com/VDIGPKU/CMUA-Watermark.

Introduction

Recently, improvements of Generative Adversarial Networks (GANs) have shown impressive results in virtual content generation, creating considerable economic and entertainment value. However, deepfakes, deep learning based face modification networks using GANs to generate fake content of target persons or target attributes, have caused great harm to people's privacy and reputation. On one hand, fake images and videos can show things that a person has never said or done, causing damage to his or her reputation, especially when involving pornography or politics (Tolosana et al. 2020). On the other hand, fake facial images with target attributes may pass the biometric authentication of commercial applications, potentially breaching security (Korshunov and Marcel 2018). Hence, defending threats brought by deepfakes requires not only distorting the modified images and lowering their visual quality to aid humans in distinguishing them from realistic images, but also ensuring

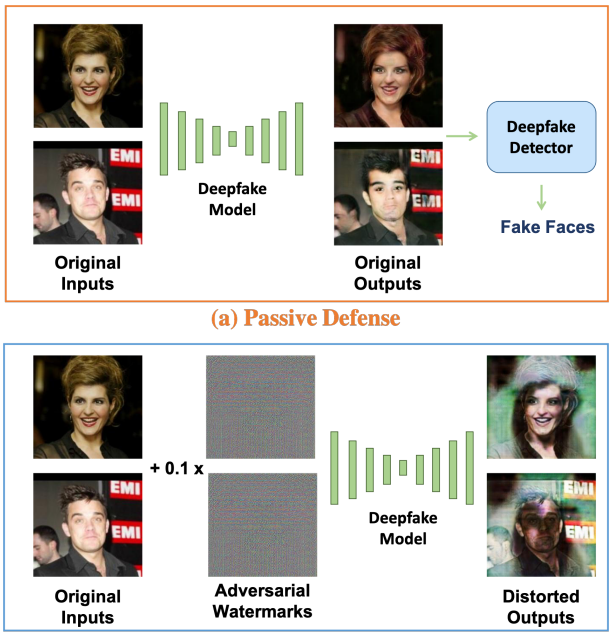

Figure 1: Passive Defense using a deepfake detector can only posteriorly lessen deepfake's harms by detecting the modified images, while active defense uses adversarial watermarks to disrupt the deepfake model to generate perceptibly distorted outputs, thus mitigating risk in advance.

*Corresponding author

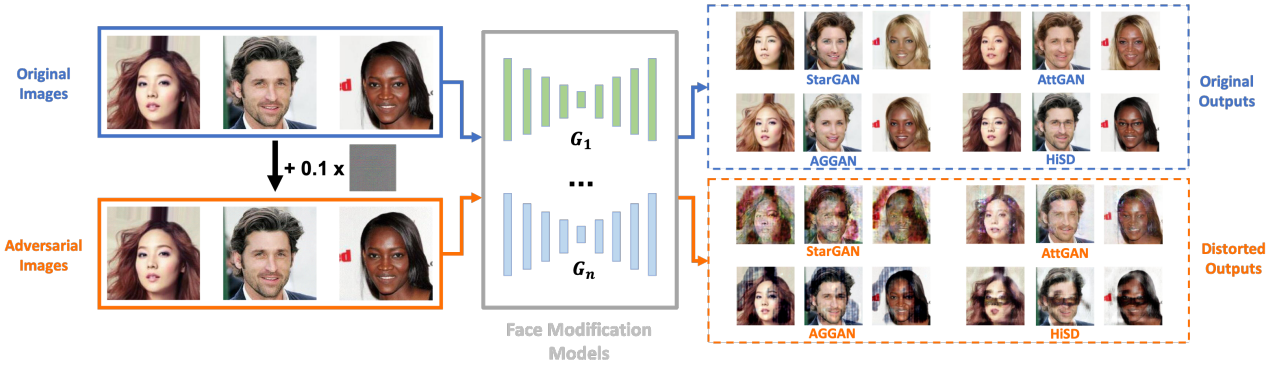

Figure 2: Illustration of our CMUA-Watermark. Once the CMUA-watermark has been generated, we can add it directly to any facial image to generate a protected image that is visually identical to the original image but can distort outputs of deepfake models such as StarGAN (Choi et al. 2018), AGGAN (Tang et al. 2019), AttGAN (He et al. 2019) and HiSD (Li et al. 2021).

that counterfeit faces do not pass liveness detection, which is the first step of most biometric verifications.

Generally speaking, the mainstream way to mitigate the risk of deepfakes is passive defense, i.e., training deepfake detectors to detect modified content (Afchar et al. 2018; Tariq, Lee, and Woo 2021; Zhao et al. 2021; Sun et al. 2021; Chen et al. 2021). Such detectors are essentially binary classifiers, predicting whether or not an image is fabricated by deepfake models. However, protecting facial images in this way is analogous to closing the stable door after the horse has bolted; the effect and harm caused cannot be reversed entirely and the risk still remains. Recently, (Ruiz, Bargal, and Sclaroff 2020) suggests using adversarial watermarks to combat deepfake models, leading them to generate visibly unreal outputs. Since these watermarks can be added to facial images in advance, they can avert the risk of malicious deepfake usage afterward as an active defense. The schematic comparison of the two ways is illustrated in Figure 1. Though (Ruiz, Bargal, and Sclaroff 2020) could defend potential threats, it can only generate an image-and-model-specific adversarial watermark, meaning that the watermark can only protect one facial image against one specific deepfake model.

To address these issues, we propose an effective and efficient solution in this work, which uses a small number (as small as 128) of training facial images to craft a cross-model universal adversarial watermark (CMUA-Watermark) for protecting a large number of facial images from multiple deepfake models, as depicted in Figure 2. Firstly, we propose a cross-model universal attack approach based on a vanilla attack method that can only protect one specific image from one model, i.e., PGD (Madry et al. 2018). Specifically, to alleviate the conflict among adversarial watermarks generated from different images and models, we newly propose a two-level perturbation fusion strategy (i.e., image-level fusion and model-level fusion) during the cross-model universal attack process. Secondly, to further weaken the conflict among adversarial watermarks generated from different models thus enhancing the transferability of the generated CMUA-Watermark, we exploit the Tree-Structured Parzen Estimator (TPE) (Bergstra et al. 2011) algorithm to automatically find the attack step sizes for different models.

Moreover, the existing evaluation method in (Ruiz, Bargal, and Sclaroff 2020) is not reasonable and comprehensive enough. Firstly, measuring image distortion through directly calculating L^1 or L^2 distances between entire original and distorted outputs overlooks deepfakes that only modify a few attributes (e.g., HiSD (Li et al. 2021) can add only a pair of glasses), as the measured distortion will be averaged by other unchanged areas. Instead, we propose using a modification mask to focus more on the modified areas. Secondly, solely considering L^1 or L^2 distances is not sufficient; ensuring protection also requires metrics reflecting generation quality and biological characteristics of distorted outputs. Accordingly, we use Fréchet Inception Distance (FID) (Heusel et al. 2017) to measure generation quality and exploit the confidence scores as well as the passing rates of liveness detection models to measure the biological characteristics of distorted outputs.

Our contributions can be summarized as the following:

- We are the first to introduce the novel idea of generating a cross-model universal adversarial watermark (CMUA-Watermark) to protect human facial images from multiple deepfakes, needing only 128 training facial images to protect a myriad of facial images.
- We propose a simple yet effective perturbation fusion strategy to alleviate the conflict and enhance the image-level and model-level transferability of the proposed CMUA-watermark.
- We deeply analyze the cross-model optimization process and develop an automatic step size tuning algorithm to find suitable attack step sizes for different models.
- We introduce a more reasonable and comprehensive evaluation method to fully evaluate the effectiveness of the adversarial watermark on combating deepfakes.

Related Works

Face Modification

In recent years, the free access to large-scale facial images and the amazing progress of generative models have led

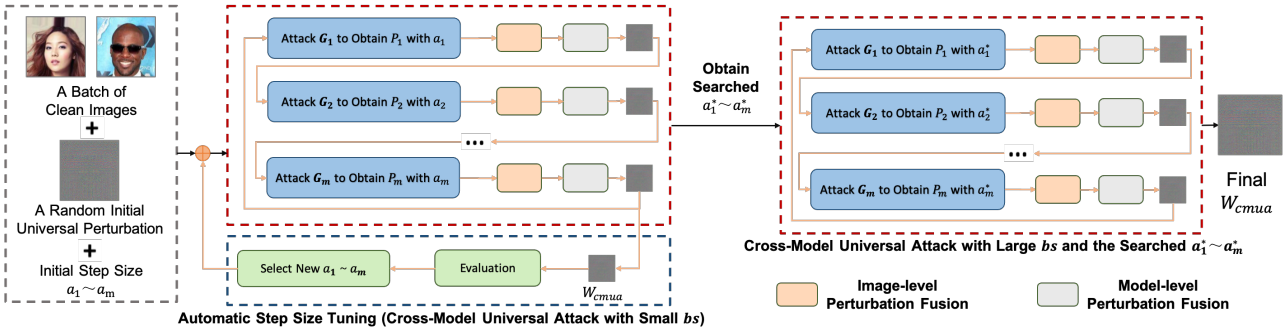

Figure 3: The overall pipeline of our Cross-Model Universal Adversarial Attack on multiple face modification networks.

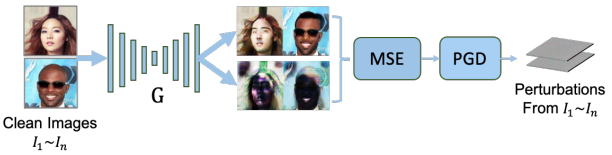

Figure 4: The detailed process of attacking one specific deepfake model.

Type	Cross-Image (i.e., universal)	Cross-Model
SIA-Watermark	✗	✗
UA-Watermark	✓	✗
CMUA-Watermark	✓	✓

Table 1: The categories of adversarial watermarks.

face modification networks to generate more realistic facial images with target persons or attributes. StarGAN (Choi et al. 2018) proposes a novel and scalable approach to perform image-to-image translation across multiple domains, achieving better visual quality in its generated images. Then, AttGAN (He et al. 2019) uses the attribute-classification constraint to provide more natural facial images on facial attribute manipulation. Moreover, AGGAN (Tang et al. 2019) introduces attention masks via a built-in attention mechanism to obtain target images with high quality. Recently, (Li et al. 2021) proposes the HiSD which is a state-of-the-art image-to-image translation method for both scalabilities of multiple labels and controllable diversity with impressive disentanglement. Though these models adopt diverse architectures and losses, our CMUA-watermark can successfully prevent facial images from being modified correctly by all of them.

Attacks on Generative Models

There are already some studies (Wang, Cho, and Yoon 2020; Yeh et al. 2020; Ruiz, Bargal, and Sclaroff 2020; Kos, Fischer, and Song 2018; Tabacof, Tavares, and Valle 2016; Bashkirova, Usman, and Saenko 2019) that explore adversarial attacks on generative models, and we particularly focus on image translation tasks that deepfakes are based upon. (Wang, Cho, and Yoon 2020) and (Yeh et al. 2020) attack image translation tasks on CycleGAN (Zhu et al. 2017) and pix2pixHD (Wang et al. 2018), which only transfer images between two domains and are thus relatively easy to attack. (Ruiz, Bargal, and Sclaroff 2020) is the first to address attacks on conditional image translation networks, but the watermark they generate only protects a specific image from a specific deepfake model, meaning that every adversarial watermark has to be individually trained, which is time-consuming and infeasible in reality.

Universal Adversarial Perturbation

Universal Adversarial Perturbation is first introduced by (Moosavi-Dezfooli et al. 2017), where a recognition model can be fooled with only a single adversarial perturbation. Here *Universal* means that a single perturbation can be added to multiple images to fool a specific model. Based on this work, (Metzen et al. 2017) introduces universal adversarial perturbation for segmentation tasks to generate target results, and (Li et al. 2019) first proposes a universal adversarial attack on image retrieval systems, leading them to return irrelevant images. The above-mentioned works only produce a universal adversarial watermark targeting one model, while our CMUA-Watermark can combat multiple face modification models simultaneously.

Methods

In this section, we first present an overview of our method. Then, we introduce how to attack a single face modification model. Ensuing, we describe the perturbation fusion strategy. Finally, we analyze the key problems of cross-model optimization and suggest an automatic step size tuning algorithm for searching for suitable attack step sizes.

Overview

In Table 1, we categorize adversarial watermarks based on their cross-image and cross-model generalization abilities. Differing from Single-Image Adversarial Watermarks (SIA-Watermarks) that protect a specific image against a specific model and Universal Adversarial Watermarks (UA-Watermarks) that protect multiple images against a specific model, the proposed CMUA-Watermark in this paper can combat multiple deepfake models while protecting a myriad of facial images.

Algorithm 1: The Cross-Model Universal Attack

Input: $X_1, ..., X_o$ (o batches of training facial images), bs (the batch size), $G_1, ..., G_m$ (the combated deepfake models), $a_1, ..., a_m$ (the step size of the attack algorithm for $G_1, ..., G_m$), A (base attack method which returns the image-and-model-specific adversarial perturbations).

Output: CMUA-Watermark W_{cmua}

```
Random Init W_0
for k ∈ [1, o] do
    for i ∈ [1, m] do
        P^i_{k_1}, ..., P^i_{k_bs} ⇐ A(G_i, a_i, X_k, W_{i+(k−1)m−1})
        P^i_avg ⇐ Image-Level Fusion with P^i_{k_1}, ..., P^i_{k_bs}
        W_{i+(k−1)m} ⇐ Model-Level Fusion with P^i_avg
    end for
end for
W_cmua = W_{m·o}
```

As illustrated in Figure 3, our overall pipeline for crafting the CMUA-Watermark is divided into two steps. In the first step, we repeatedly conduct the cross-model universal attack with a small batch size (for a faster search), evaluate the generated CMUA-Watermark, and then use automatic step size tuning to select new attack step sizes $a_1, ..., a_m$. In the second step, we use the found step sizes $a_1^*, ..., a_m^*$ to conduct the cross-model universal attack with a large batch size (for enhancing the disrupting capability) and generate the final CMUA-Watermark. Specifically, as shown in Algorithm 1, during the proposed cross-model universal attacking process, batches of input images iteratively go through the PGD (Madry et al. 2018) attack to generate adversarial perturbations, which then go through a two-level perturbation fusion mechanism to combine into a fused CMUA-Watermark that serves as the initial perturbation for the next model.

Combating One Face Modification Model

In this section, we describe the approach to disrupt a single face modification model in our pipeline (the blue box in Figure 3), illustrated in detail in Figure 4. To begin, we input a batch of clean images $I_1...I_n$ to the deepfake model G and obtain the original outputs $G(I_1)...G(I_n)$. Then, we input $I_1...I_n$ with the initial adversarial perturbation W to G, getting the initial distorted outputs $G(I_1 + W)...G(I_n + W)$. Subsequently, we utilize Mean Square Error (MSE) to measure the differences between $G(I_1)...G(I_n)$ and $G(I_1 + W)...G(I_n + W)$,

$$\max_{W} \sum_{i=1}^{n} MSE(G(I_i), G(I_i + W)),\ s.t.\ \|W\|_\infty \leq \epsilon, \quad (1)$$

where ϵ is the upper bound magnitude of the adversarial watermark W. Finally, we use PGD (Madry et al. 2018) as the base attack method to update the adversarial perturbations at every attack iteration,

$$I_{adv}^0 = I + W,$$
$$I_{adv}^{r+1} = clip_{I,\epsilon}\{I_{adv}^r + a\ sign(\nabla_I L(G(I_{adv}^r), G(I)))\}, \quad (2)$$

where I is the clean facial images, I_{adv}^r is the adversarial facial images in the rth iteration, a is the step size of the base attack, L is the loss function (we select MSE as formulated in Eq.(2)), G is the face modification network we attack, and the operation $clip$ limits the I_{adv} in the range $[I - \epsilon, I + \epsilon]$.

Through this process, we can obtain Single-Image-Adversarial Watermarks (SIA-Watermarks) that protect one facial image from one specific deepfake model. However, the crafted SIA-Watermarks are inadequate under the cross-model setting; they are lacking in image-and-model-level transferability. In the following two sections, we introduce our solutions to address this problem.

Adversarial Perturbation Fusion

The conflict among adversarial watermarks generated from different images and models will decrease the transferability of the proposed CMUA-Watermark. To weaken this conflict, we propose a two-level perturbation fusion strategy during the attack process. Specifically, when we attack one specific deepfake model, we conduct an **image-level** fusion to average the signed gradients from a batch of facial images,

$$G_{avg} = \frac{\sum_j^{bs} sign(\nabla_{I_j} L(G(I_j^{adv}), G(I_j)))}{bs}, \quad (3)$$

where bs is the batch size of facial images, and I_j^{adv} is the jth adversarial image of a batch. This operation will cause the G_{avg} to concentrate more on the common attributes of human faces rather than a specific face's attributes. Then, we use PGD to generate the adversarial perturbation P_{avg} though G_{avg} as eq.(2).

After obtaining the P_{avg} from one model, we conduct a **model-level** fusion, which iteratively combines P_{avg} generated from specific models to the W_{CMUA} in training, and the initial W_{CMUA} is just the P_{avg} calculated from the first deepfake model,

$$\begin{aligned} W_{CMUA}^0 &= P_{avg}^0, \\ W_{CMUA}^{t+1} &= \alpha \cdot W_{CMUA}^t + (1 - \alpha) \cdot\ P_{avg}^t, \end{aligned} \quad (4)$$

where α is a decay factor, P_{avg}^t is the averaged perturbation generated from the tth combated deepfake model, and W_{CMUA}^t is the training CMUA-Watermark after the tth attacking deepfake model.

Automatic Step Size Tuning Based on TPE

Besides the above mentioned two-level fusion, we find that the attack step sizes for different models are also important for the transferability of the generated CMUA-Watermark. Hence, we exploit a heuristic approach to automatically find the suitable attack step sizes.

The base attack method we select (PGD) falls into the FGSM (Goodfellow, Shlens, and Szegedy 2015) family, and the gradients $\nabla_x L$ are normalized by the $sign$ function:

$$\text{sign}\, x = \begin{cases} -1 & , \quad x < 0, \\ 0 & , \quad x = 0, \\ 1 & , \quad x > 0. \end{cases} \quad (5)$$

In real calculations, the elements in $\nabla_x L$ almost never reach 0, so $||sign(\nabla_x L)||_2 \approx 1$ is fixed for any gradient. The updated perturbation ΔP in an iteration of a sign-based attack method is formulated as:

$$\Delta P = a \cdot \text{sign}(\nabla_X L). \tag{6}$$

That is to say, only the step size a determines the update rate during the attack, so the selection of a has a great influence on the attack performance. This conclusion is also valid for cross-model universal attacks; the updated perturbation ΔP^u in an iteration of a cross-model universal attack is formed by combining ΔP_i from multiple models $G_1, ..., G_m$:

$$\Delta P^u = \sum_{i=1}^{m} \alpha^{(m-i)} \Delta P_i = \sum_{i}^{m} \alpha^{(m-i)} a_i \cdot \text{sign}(\nabla_X L_i). \tag{7}$$

In the formula above, m is the number of models, decay factor α is a constant, and $\text{sign}(\nabla_X L_i)$ provides an optimization direction for G_i. **Hence, the overall optimization direction is greatly influenced by $a_1, ..., a_m$, and selecting the appropriate $a_1, ..., a_m$ across different models to find an ideal overall direction is a key problem for cross-model attacks.**

We introduce the TPE (Bergstra et al. 2011) algorithm to solve this problem, automatically searching for the suitable $a_1, ..., a_m$ to balance the different directions calculated from multiple models. TPE is a hyper-parameter optimization method based on Sequential Model-Based Optimization (SMBO), which sequentially constructs models to approximate the performance of hyperparameters based on historical measurements, and then subsequently chooses new hyperparameters to test based on this model. In our task, we regard step sizes $a_1, ..., a_m$ as the input hyperparameters x and the success rate of the attack as the associated quality score y of TPE. The TPE uses $P(x|y)$ and $P(y)$ to model $P(y|x)$, and $p(x|y)$ is given by:

$$p(x \mid y) = \begin{cases} \ell(x), & \text{if } y < y^*, \\ g(x), & \text{if } y \geq y^*, \end{cases} \tag{8}$$

where y^* is determined by the historically best observation, $\ell(x)$ is the density formed with the observations $\{x^{(i)}\}$ such that the corresponding loss is lower than y^*, and $g(x)$ is the density formed with the remaining observations. After modeling the $P(y|x)$, we constantly look for better step sizes by optimizing the Expected Improvement (EI) criterion in every search iteration, which is given by,

$$\begin{aligned} EI_{y^*}(x) =& \frac{\gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma \ell(x) + (1-\gamma) g(x)} \\ \propto& \left(\gamma + \frac{g(x)}{\ell(x)} (1-\gamma) \right)^{-1}, \end{aligned} \tag{9}$$

where $\gamma = p(y < y^*)$. Compared with other criteria, EI is intuitive and has been proven to have an excellent performance. For more details on TPE, refer to (Bergstra et al. 2011).

Dataset/Model	$SR_{mask}\uparrow$	FID↑	ACS↓	TFHC↓
CelebA/StarGAN	100.00%	201.0	0.286	66.3%→20.6%
CelebA/AGGAN	99.88%	51.0	0.863	65.9%→55.5%
CelebA/AttGAN	87.08%	65.1	0.638	55.1%→28.1%
CelebA/HiSD	99.87%	92.7	0.153	63.3%→3.9%
LFW/StarGAN	100.00%	169.3	0.207	43.9%→8.2%
LFW/AGGAN	99.99%	37.7	0.806	54.9%→46.3%
LFW/AttGAN	94.07%	70.6	0.496	25.9%→16.7%
LFW/HiSD	98.13%	88.1	0.314	50.7%→16.0%
Film100/StarGAN	100.00%	259.7	0.425	61.0%→29.1%
Film100/AGGAN	99.88%	129.1	0.832	61.0%→55.7%
Film100/AttGAN	95.82%	177.5	0.627	34.6%→25.8%
Film100/HiSD	100.00%	220.7	0.207	67.0%→14.0%

Table 2: The quantitative results of CMUA-Watermark.

Experiment

In this section, we will first describe our datasets and implementation details. Following, we introduce our evaluation metrics. Then, we show the experimental results of the proposed CMUA-Watermark. Moreover, we systemically conduct an ablation study. Finally, we show an application of the proposed model watermark in realistic scenes.

Datasets and Implementation Details

In our experiments, we use the CelebA (Liu et al. 2015) test set as the main dataset, which contains 19962 facial images. We use the first 128 images in the set as training images and evaluate our method on all facial images of the CelebA test set and the LFW (Huang et al. 2007) dataset to ensure credibility. In addition, we also randomly select 100 facial images from films as additional data (Films100) to verify the effectiveness of the CMUA-Watermark in real scenarios. It is important to point out that we do not use any additional data to train our CMUA-Watermark.

The face modification networks we select in our experiments are StarGAN (Choi et al. 2018), AGGAN (Tang et al. 2019), AttGAN (He et al. 2019), and HiSD (Li et al. 2021). StarGAN and AGGAN are both trained on the CelebA dataset for five attributes: black hair, blond hair, brown hair, gender, and age. AttGAN is trained on the CelebA dataset for up to fourteen attributes, which is more complicated compared with the two networks above. We also attack one of the latest face modification networks HiSD, which is also trained on the CelebA dataset and can add a pair of glasses to the target person.

During the process of searching for the step sizes, the maximum number of iterations is 1k, and the search space of the step size for each model is $[0, 10]$. We first search for the step sizes with $batchsize = 16$ and then use the searched step sizes to conduct cross model attacks with $batchsize = 64$.

Evaluation Metrics

Considering the limitations of the existing evaluation method and after rethinking the purpose of combating deepfake models, we design a more reasonable and comprehen-

Method	$SR_{mask} \uparrow$				$\log_{10} FID \uparrow$			
	StarGAN	AGGAN	AttGAN	HiSD	StarGAN	AGGAN	AttGAN	HiSD
BIM (Kurakin, Goodfellow, and Bengio 2017)	0.6755	0.9975	0.2126	0.0028	1.9047	1.6539	1.2451	1.6098
MIM (Dong et al. 2018)	**1.0000**	**0.9994**	0.0200	0.0438	**2.5281**	**1.8435**	0.7842	1.5205
PGD (Madry et al. 2018)	0.8448	0.9970	0.0146	0.0010	2.0203	1.6659	0.9403	1.6467
DI^2-FGSM (Xie et al. 2019)	0.028	0.3448	0.0074	0.0001	1.5714	1.3084	1.2036	1.4113
M-DI^2-FGSM (Xie et al. 2019)	**1.0000**	0.9987	0.0032	0.0050	1.5714	1.3084	1.2036	1.4113
AutoPGD (Croce and Hein 2020)	0.8314	0.9963	0.0002	0.0007	1.5714	1.3084	1.2036	1.4113
Ours	**1.0000**	0.9988	**0.8708**	**0.9987**	2.3032	1.7072	**1.8133**	**1.9672**

Table 3: Comparisons of SR_{mask} and $\log_{10} FID$ with state-of-the-art attack methods.

sive evaluation method, which concentrates on metrics of three aspects.

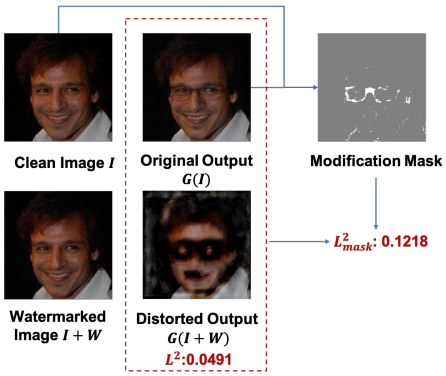

Figure 5: A problematic case for the existing evaluation method and the proposed modification mask for addressing this issue.

To begin, we analyze the success of the modification process. (Ruiz, Bargal, and Sclaroff 2020) calculates the L^2 score between the original outputs $G(I)$ and distorted outputs $G(I+W)$ and declares the facial image as successfully protected by the adversarial watermark when $||G(I+W) - G(I)||_2 > 0.05$. However, as Figure 5 shows, this evaluation method is problematic for some cases. For example, the distorted output $G(I+W)$ is noticeably different from the original output $G(I)$ in the area of modification, especially around the eyes, hence successfully preventing the deepfake model from adding glasses to the facial image. However, the existing evaluation regards the output as a failed case. To solve this issue, we introduce a mask matrix, concentrating more on the modified areas,

$$Mask_{(i,j)} = \begin{cases} 1, & if ||G(I)_{(i,j)} - I_{(i,j)}|| > 0.5, \\ 0, & else, \end{cases} \tag{10}$$

where (i, j) is the coordinate of pixels in the image. In this way, when calculating L^2_{mask}, only pixels with major changes will be calculated and other areas will be left out,

$$L^2_{mask} = \frac{\sum_i \sum_j Mask_{(i,j)} \cdot ||G(I)_{(i,j)} - G(I + W_{CMUA})_{(i,j)}||}{\sum_i \sum_j Mask_{(i,j)}}. \tag{11}$$

In our experiments, if the $L^2_{mask} > 0.05$, we determine that the image is protected successfully, and use SR_{mask} to represent the success rate of protecting facial images.

Secondly, we use FID (Heusel et al. 2017) to measure the generation quality of fake facial images. FID comprehensively represents the distance of the feature vectors from Inception v3 (Szegedy et al. 2016) between original images and fake images, and higher FID values indicate lower qualities of the generated images.

Finally, we use an open-source liveness detection system HyperFAS[1] to test the biological characteristics of the fake images. The HyperFAS is based on Mobilenet (Howard et al. 2017) and trained with 360k facial images. In our experiments, if the confidence score is greater than 0.99, we conclude that the face is a true face with high confidence (**TFHC**). Additionally, we also calculate the average confidence score **ACS** for evaluation.

The Results of CMUA-Watermark

We conduct extensive experiments to demonstrate the effectiveness of the proposed CMUA-Watermark. We first show the quantitative and qualitative results of the proposed CMUA-Watermark and then compare our method with state-of-the-art attack methods.

The quantitative and qualitative results of the proposed CMUA-Watermark are reported in Table 2 and Figure 2, respectively. The proposed CMUA-Watermark has a similar overall performance on CelebA and LFW and performs better on StarGAN, AGGAN, and HiSD than on AttGAN. Specifically, the SR_{mask} combating StarGAN, AGGAN, and HiSD is close to 100% on both CelebA and LFW, and the ACS of the distorted outputs decreases significantly compared with that of the original outputs, which makes the $TFHC$ of StarGAN, AGGAN, AttGAN, and HiSD each drop by 45.65%, 10.36%, 27.08% and 59.36% on the CelebA test set and 35.68%, 8.58%, 9.13% and 34.65% on LFW. Besides, on the Film100 dataset that is closer to real scenes, the proposed watermark performs even better compared with the above two datasets. Also, all distorted outputs have large FID, demonstrating their poor generation quality. Overall, the above qualitative and quantitative results both demonstrate that the CMUA-Watermark can successfully protect facial images from multiple deepfake models.

We further compare the CMUA-Watermark with state-of-the-art attack methods on CelebA, including BIM (Kurakin, Goodfellow, and Bengio 2017), MIM (Dong et al. 2018),

[1]https://github.com/zeusees/HyperFAS

Methods	PGD			ours
perturbation fusion		✓		✓
automatic step size tuning			✓	✓
SR_{mask}(StarGAN)	84.5%	27%	99.0%	**100.0%**
SR_{mask}(AGGAN)	99.7%	92%	**100.0%**	99.9%
SR_{mask}(AttGAN)	1.5%	82.5%	20.4%	**87.1%**
SR_{mask}(HiSD)	0.1%	0%	5.0%	**99.9%**

Table 4: Ablation study on CelebA test set.

Base Attack Method	SR_{mask}			
	StarGAN	AGGAN	AttGAN	HiSD
CMUA-MIM	0.999	0.999	0.820	0.999
CMUA-PGD	1.000	0.999	0.871	0.999

Table 5: Comparisons of base attack methods on CelebA test set.

PGD (Madry et al. 2018), DI^2-FGSM (Xie et al. 2019), M-DI^2-FGSM (Xie et al. 2019), AutoPGD (Croce and Hein 2020). We refer to (Moosavi-Dezfooli et al. 2017) and adjust these methods to the universal setting (detailed description in the Appendix F). The comparison results of SR_{mask} and FID are reported in Table 3, and we can observe that the adversarial watermarks crafted by the compared methods (e.g. MIM) over-optimize on one or two models hence performing very poorly on others. Contrarily, our method achieves excellent performance on all models. These results demonstrate the better image-level and model-level transferability of the proposed method than the existing ones.

Ablation Study

In this section, we first investigate the effectiveness of perturbation fusion and automatic step size tuning. Then, we investigate the influence of other hyperparameters on the proposed CMUA-Watermark. As reported in Table 4, the base PGD algorithm has very poor performances on AttGAN and HiSD, indicating that its model-level transferability is weak. When separately using the perturbation fusion strategy the overall performance is improved, but the result for StarGAN drops significantly. On the other hand, if we solely perform automatic step size tuning, the results for all models have been improved but the results for AttGan and HiSD are still not good enough. After combining both of them, the performance of our CMUA-Watermark has improved significantly for all deepfake models. These results demonstrate that perturbation fusion and automatic step size tuning are both crucial to the proposed method and should be used together.

We also investigate the influence on the proposed watermark with the changes of base attack algorithms and the upper bound ϵ. As shown in Table 5, with the proposed method, changing the base attack method has little influence on the CMUA-Watermark. Besides, as illustrated in Figure 6, the generated fake facial images are more distorted when the parameter ϵ becomes larger, meaning that the protection performance becomes better. However, when ϵ becomes too large, the produced adversarial watermark is more likely to be seen. We empirically find that setting ϵ around 0.05 can make a good trade-off between the protection performance and the imperceptibility of the produced adversarial watermark.

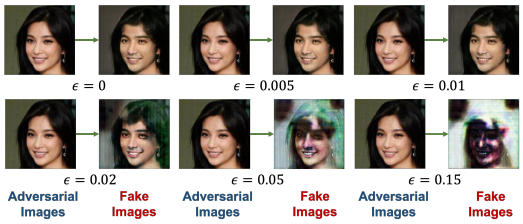

Figure 6: Examples of CMUA-Watermark with different settings of ϵ.

Figure 7: Illustration of CMUA-Watermark preventing fake facial images from passing liveness detection modules integrated in commercial applications.

Applications

We further apply the proposed CMUA-Watermark to real social media platforms, which require users to upload their real facial images (e.g., Tantan[2] and Jimu[3]). As illustrated in Figure 7, we find that some fake images generated by StarGAN can pass the verification of liveness detection of Tantan and Jimu successfully. However, the fake images generated from facial images protected by our CMUA-Watermark fail to do so, which demonstrates that our CMUA-Watermark is very effective even in real applications.

Conclusion

In this paper, we have proposed a cross-model universal attack pipeline to produce a watermark that can protect a large number of facial images from multiple deepfake models. Specifically, we propose a perturbation fusion strategy to alleviate the conflict of adversarial watermarks generated from different images and models in the attack process. Further, we analyze the key problem of cross-model optimization and introduce an automatic step size tuning algorithm based on TPE to determine the overall optimization direction. Moreover, we design a more reasonable and comprehensive evaluation method to evaluate the proposed CMUA-Watermark. Experimental results demonstrate that our CMUA-Watermark can effectively disrupt the modification by deepfake models, lower the quality of generated images, and prevent the fake facial images from passing the verification of liveness detection systems.

[2]https://apps.apple.com/cn/app/id861891048

[3]https://apps.apple.com/cn/app/id1094615747

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62176007. This work was also a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7.

Bashkirova, D.; Usman, B.; and Saenko, K. 2019. Adversarial Self-Defense for Cycle-Consistent GANs. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bergstra, J.; Bardenet, R.; Bengio, Y.; and Kégl, B. 2011. Algorithms for Hyper-Parameter Optimization. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; and Ji, R. 2021. Local Relation Learning for Face Forgery Detection. In *AAAI*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks with Momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9185–9193. Los Alamitos, CA, USA: IEEE Computer Society.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11): 5464–5478.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6629–6640. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst.

Korshunov, P.; and Marcel, S. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *CoRR*, abs/1812.08685.

Kos, J.; Fischer, I.; and Song, D. 2018. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, 36–42. IEEE.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Li, J.; Ji, R.; Liu, H.; Hong, X.; Gao, Y.; and Tian, Q. 2019. Universal Perturbation Attack Against Image Retrieval. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4898–4907.

Li, X.; Zhang, S.; Hu, J.; Cao, L.; Hong, X.; Mao, X.; Huang, F.; Wu, Y.; and Ji, R. 2021. Image-to-Image Translation via Hierarchical Style Disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8639–8648.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Metzen, J. H.; Kumar, M. C.; Brox, T.; and Fischer, V. 2017. Universal Adversarial Perturbations Against Semantic Image Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2774–2783.

Moosavi-Dezfooli, S.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 86–94.

Ruiz, N.; Bargal, S. A.; and Sclaroff, S. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision*, 236–251. Springer.

Sun, K.; Liu, H.; Ye, Q.; Gao, Y.; Liu, J.; Shao, L.; and Ji, R. 2021. Domain General Face Forgery Detection by Learning to Weight. In *AAAI*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tabacof, P.; Tavares, J.; and Valle, E. 2016. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*.

Tang, H.; Xu, D.; Sebe, N.; and Yan, Y. 2019. Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation. In *International Joint Conference on Neural Networks (IJCNN)*.

Tariq, S.; Lee, S.; and Woo, S. S. 2021. One Detector to Rule Them All: Towards a General Deepfake Attack Detection Framework. In *Proceedings of The Web Conference 2021*.

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64: 131–148.

Wang, L.; Cho, W.; and Yoon, K.-J. 2020. Deceiving Image-to-Image Translation Networks for Autonomous Driving With Adversarial Perturbations. *IEEE Robotics and Automation Letters*, PP: 1–1.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. 2019. Improving Transferability of Adversarial Examples with Input Diversity. In *Computer Vision and Pattern Recognition*. IEEE.

Yeh, C.-Y.; Chen, H.-W.; Tsai, S.-L.; and Wang, S.-D. 2020. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 53–62.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-attentional Deepfake Detection. arXiv:2103.02406.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.