# **Abstrak-**

Penerapan deepfake yang berbahaya (yaitu teknologi yang dapat menghasilkan wajah target atau atribut wajah) telah menjadi ancaman besar bagi masyarakat kita. Konten multimedia palsu yang dihasilkan oleh model deepfake dapat merusak reputasi dan bahkan mengancam properti orang yang ditiru. Untungnya, tanda air musuh dapat digunakan untuk memerangi model deepfake, mengarahkan mereka untuk menghasilkan gambar yang terdistorsi. Metode yang ada saat ini membutuhkan proses pelatihan individual untuk setiap gambar wajah, untuk menghasilkan watermark lawan terhadap model deepfake tertentu, yang sangat tidak efisien. Untuk mengatasi masalah ini, kami mengusulkan sebuah metode serangan adversarial universal pada model deepfake, untuk menghasilkan sebuah Cross-Model Universal Adversarial Watermark (CMUAWatermark) yang dapat melindungi ribuan gambar wajah dari beberapa model deepfake. Secara khusus, pertama-tama kami mengusulkan sebuah pipeline serangan universal crossmodel dengan menyerang beberapa model deepfake dan menggabungkan gradien dari model-model tersebut secara berulang. Kemudian kami memperkenalkan metode berbasis batch untuk mengurangi konflik tanda air yang berlawanan yang dihasilkan oleh gambar wajah yang berbeda. Terakhir, kami merancang metode evaluasi yang lebih masuk akal dan komprehensif untuk mengevaluasi efektivitas tanda air yang berlawanan. Hasil eksperimen menunjukkan bahwa CMUA-Watermark yang diusulkan dapat secara efektif mendistorsi gambar wajah palsu yang dihasilkan oleh model deepfake dan berhasil melindungi gambar wajah dari deepfake dalam adegan nyata.

# **Pendahuluan**

Baru-baru ini, peningkatan besar Generative Adversarial Networks (GAN) telah menunjukkan hasil yang mengesankan untuk pembuatan konten virtual. Namun, konten virtual yang dihasilkan oleh GAN juga menimbulkan ancaman besar bagi masyarakat kita (misalnya, Gambar / Video Deepfake). Deepfake (kata serapan dari "deep learning" dan "fake"), yang mengacu pada jaringan modifikasi wajah berbasis deep learning untuk menghasilkan konten palsu dari orang target atau atribut target oleh GAN, telah menyebabkan kerugian besar bagi sebagian orang, terutama figur publik. Di satu sisi, gambar dan video palsu dapat menunjukkan hal-hal yang tidak pernah dikatakan atau dilakukan oleh seseorang, yang dapat merusak reputasinya, terutama dalam hal pornografi atau politik. Di sisi lain, wajah palsu dengan atribut target dapat lolos verifikasi aplikasi komersial, mengancam keamanan aplikasi dan merugikan properti orang yang ditiru.

Secara umum, ada dua cara untuk mengurangi risiko deepfake. Salah satunya adalah melatih pendeteksi deepfake [4], [5], [6], [7], [8], [9], [10] untuk mendeteksi konten yang telah dimodifikasi, dan yang kedua adalah menyerang model deepfake untuk menghasilkan gangguan yang berlawanan [11]. Gangguan seperti itu dapat diperlakukan sebagai Adversarial Watermark untuk memerangi model deepfake, yang menyebabkan model tersebut menghasilkan keluaran yang jelas tidak nyata. Namun, metode yang diusulkan dalam [11] hanya dapat menghasilkan tanda air musuh untuk gambar tertentu terhadap model deepfake tertentu dengan proses pelatihan individual. Jelas, ini sangat memakan waktu dan tidak layak untuk menghasilkan tanda air seperti itu dengan generalisasi yang rendah untuk ribuan gambar wajah.

Untuk mengatasi masalah ini, kami mengusulkan metode serangan universal yang efektif dan efisien, yang dapat menghasilkan tanda air musuh untuk melindungi ribuan gambar wajah dari beberapa model deepfake. Berdasarkan kemampuan generalisasi dari watermark adversarial universal, kami membaginya ke dalam beberapa kategori berikut:

1. SIA-Watermark: Tanda air lawan satu gambar yang dapat melindungi satu gambar dari satu model deepfake. Tanda air dalam [11] termasuk dalam kategori ini.
2. CIUA-Watermark: Tanda air permusuhan universal lintas-gambar yang dapat melindungi beberapa gambar dari satu model deepfake.
3. CMUA-watermark: Tanda air permusuhan universal lintas model yang dapat melindungi beberapa gambar dari beberapa model deepfake. Tanda air CMUA kami termasuk dalam kategori ini.

Dalam artian, menghasilkan CIUA-Watermark adalah untuk menemukan titik temu dari kelemahan data, sedangkan menghasilkan CMUAwatermark adalah untuk menemukan titik temu dari kelemahan data dan kelemahan model secara bersamaan. Untuk mencapai tujuan ini, kami mengusulkan sebuah pipeline serangan universal lintas model untuk model deepfake. Pertama, kami memasukkan sekumpulan gambar wajah ke dalam beberapa model deepfake untuk mendapatkan watermark yang berlawanan per gambar. Namun, karena atribut unik dari gambar wajah tertentu, tanda air ini mungkin saling bertentangan satu sama lain. Kemudian, untuk mengurangi konflik dan meningkatkan kinerja watermark adversarial universal, kami merata-ratakan sejumlah besar watermark yang dihasilkan oleh gambar wajah untuk mendapatkan watermark rata-rata, yang dapat mengarahkan watermark untuk fokus pada atribut umum wajah manusia daripada atribut unik dari satu gambar tertentu. Selanjutnya, kami secara iteratif menyerang beberapa model deepfake untuk menyempurnakan watermark lawan kami selangkah demi selangkah. Terakhir, CMUAwatermark yang terlatih dengan baik dapat ditambahkan ke dalam gambar wajah manusia untuk melindunginya dari beberapa model deepfake.

Untuk menilai apakah sebuah gambar berhasil diproteksi tidaklah mudah, dan metode evaluasi yang ada di [11] tidak cukup komprehensif karena hanya berfokus pada jarak L1/L2 antara gambar palsu yang dihasilkan oleh contoh-contoh yang berlawanan (F akeF acea) dan gambar-gambar palsu yang dihasilkan oleh gambar-gambar yang bersih (F akeF acec). Sesuai dengan tujuan awal kami, kami merancang metode evaluasi yang lebih masuk akal dan komprehensif. Di satu sisi, kami mensyaratkan bahwa watermark lawan harus menurunkan kualitas F akeF acea dan memaksimalkan perbedaan antara F akeF acea dan F akeF acec. Secara khusus, kami menggunakan Frechet Inception Distance (FID) [12] untuk mengukur kualitas wajah palsu yang dihasilkan, menggunakan jarak L2 untuk mengukur distorsi antara F akeF acea dan F akeF acec, dan menggunakan PSNR, SSIM [13] dan LPIPS [14] untuk mengukur kemiripan antara F akeF acea dan F akeF acec. Di sisi lain, kami mensyaratkan bahwa tanda air lawan harus membuat wajah palsu tidak dapat melewati sistem pendeteksian keaktifan, yang digunakan oleh aplikasi komersial untuk memastikan bahwa gambar tersebut berasal dari orang yang sebenarnya. Oleh karena itu, kami menganggap nilai kepercayaan dan tingkat kelulusan sistem deteksi kehidupan sebagai dua metrik lainnya. Hasil eksperimen menunjukkan bahwa CMUAWatermark kami dapat secara efektif melindungi gambar wajah dari beberapa model deepfake (seperti yang diilustrasikan pada Gambar 1) dan dapat berhasil mencegah gambar wajah palsu melewati verifikasi sistem pendeteksian kehidupan aplikasi komersial.

Kontribusi kami dapat diringkas sebagai:

* Pertama-tama kami memperkenalkan ide baru untuk menghasilkan tanda air permusuhan universal model silang (CMUAWatermark) untuk melindungi gambar wajah manusia dari pemalsuan.
* Kami mengusulkan sebuah pipeline serangan universal lintas model untuk menghasilkan CMUA-Watermark terhadap beberapa model deepfake dengan merata-ratakan secara iteratif tanda air yang berlawanan yang dihasilkan dari beberapa gambar wajah.
* Kami merancang metode evaluasi yang lebih masuk akal dan komprehensif untuk sepenuhnya mengevaluasi efektivitas watermark lawan untuk memerangi deepfake.
* Kami mencapai kinerja yang sangat efektif untuk melindungi gambar wajah dari deepfake, bahkan dalam adegan nyata.

# **Karya-karya Terkait**

## **Modifikasi Wajah**

Dalam beberapa tahun terakhir, akses gratis ke gambar wajah berskala besar dan kemajuan luar biasa dari model generatif telah membuat jaringan modifikasi wajah menghasilkan gambar wajah palsu yang lebih nyata dengan orang atau atribut target. StarGAN [1] mengusulkan pendekatan baru dan terukur untuk melakukan penerjemahan gambar-ke-gambar di berbagai domain. Dibandingkan dengan metode sebelumnya [15], [16], [17], gambar yang dihasilkan oleh StarGAN telah mencapai kualitas visual yang lebih baik. Untuk mengatasi masalah bahwa batasan atribut-independen yang ketat pada representasi laten menyebabkan pembangkitan yang terlalu halus dan terdistorsi, AGGAN [3] menghapusnya dan hanya menggunakan batasan klasifikasi atribut untuk memberikan gambar palsu yang lebih nyata pada manipulasi atribut wajah. Selain itu, AGGAN [2] memperkenalkan topeng perhatian melalui mekanisme perhatian bawaan untuk mendapatkan gambar target dengan kualitas tinggi.

Meskipun metode-metode ini mengadopsi arsitektur dan kerugian yang beragam, CMUA-watermark kami dapat mencegah gambar wajah dimodifikasi dengan benar oleh mereka.

## **Serangan terhadap Model Generatif**

Sudah ada beberapa penelitian [11], [18], [19], [20], [21], [22] yang mengeksplorasi serangan lawan pada model generatif (yaitu GAN) dan kami fokus pada serangan pada penerjemahan gambar. [18] dan [19] mengadaptasi serangan lawan pada penerjemahan gambar pada CycleGAN [17] dan pix2pixHD [23] yang hanya mentransfer gambar di antara dua domain sehingga relatif mudah untuk diserang. [11] pertama kali membahas gangguan pada jaringan penerjemahan gambar bersyarat tetapi watermark yang mereka hasilkan adalah SIA-Watermark. Dengan kata lain, gangguan dengan generalisasi yang rendah hanya dapat melindungi gambar tertentu dari model deepfake tertentu.

Dibandingkan dengan [11], metode cross-model universal attack yang diusulkan menghasilkan CMUA-watermark dengan generalisasi yang tinggi, yang dapat melindungi ribuan citra wajah dari beberapa model deepfake tanpa pelatihan individual untuk setiap citra.

## **Tanda Air Serangan Universal**

Universal Adversarial Perturbation pertama kali diperkenalkan oleh [24] yang dapat menipu model pengenalan hanya dengan satu gangguan. Di sini, Universal berarti gangguan tunggal dapat ditambahkan ke beberapa gambar untuk menipu satu model. [25] memperkenalkan gangguan permusuhan universal untuk tugas segmentasi untuk menghasilkan hasil target. [26] pertama kali mengusulkan serangan adversarial universal terhadap sistem pengambilan gambar yang dapat menyebabkan sistem mengembalikan gambar yang tidak relevan.

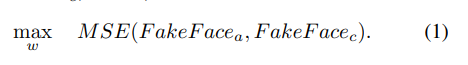
Dibandingkan dengan karya-karya [24], [25], [26], [27], [28] yang hanya dapat menghasilkan universal adversarial watermark untuk satu model, kami mengusulkan metode serangan universal lintas model pada jaringan modifikasi wajah, yaitu, perturbasi adversarial universal kami dapat memerangi beberapa model modifikasi wajah secara bersamaan.

# **Method**

Pada bagian ini, kami memperkenalkan metode kami secara rinci. Pertama, kami memperkenalkan cara menyerang jaringan penerjemahan gambar di Bagian 3.1. Kemudian, kami menjelaskan serangan permusuhan universal lintas-citra dan lintas-model di Bagian 3.2 dan 3.3.

## **Attacks on Face Modification Model**

Dalam penelitian kami, kami mengusulkan metode serangan adversarial universal untuk menyerang beberapa model modifikasi wajah, dan menghasilkan tanda air adversarial w. Tujuan kami adalah untuk menurunkan kualitas FakeFacea, memaksimalkan perbedaan antara FakeFacea dan FakeFacec, dan membuat FakeFacea gagal melewati verifikasi sistem pendeteksian kehidupan. Kami menggunakan Mean Squared Error (MSE) untuk mengukur perbedaan antara FakeFacea dan FakeFacec, yaitu,

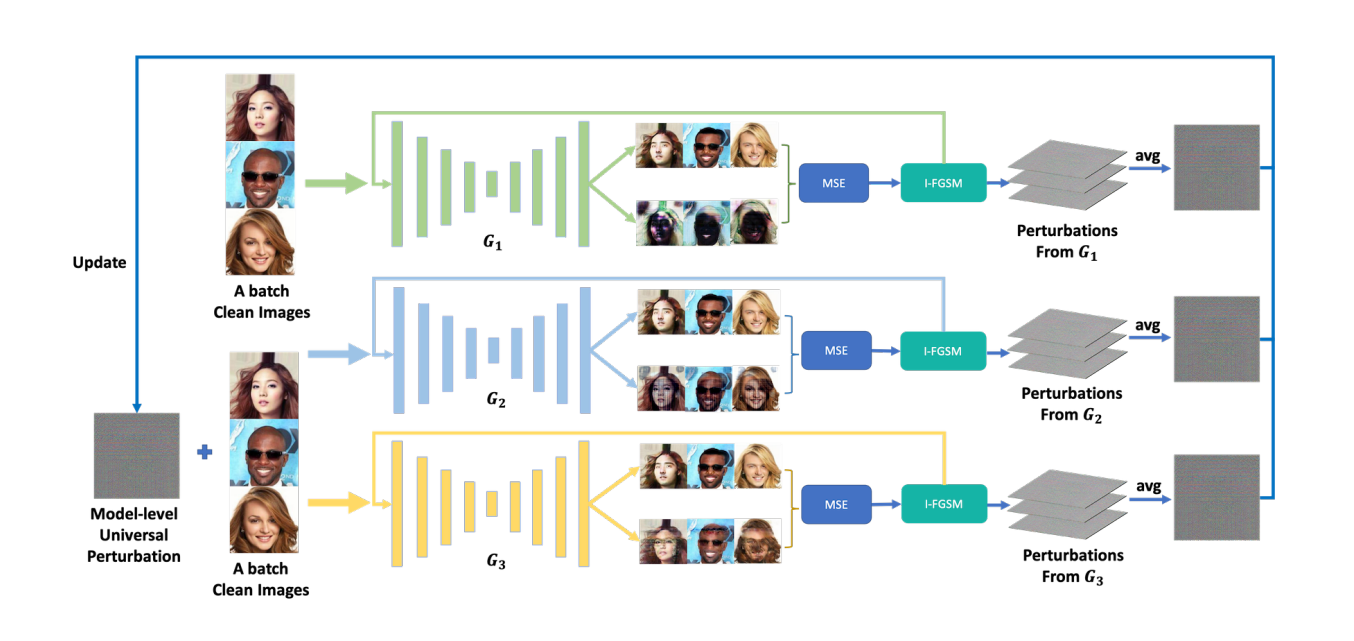


Selain itu, tanda air lawan w harus tidak terlihat oleh mata manusia untuk memastikan kualitas visual gambar wajah yang dilindungi. Oleh karena itu, kami memperkenalkan batasan itu:

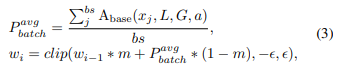


di mana E adalah besaran batas atas tanda air.

## **Serangan Musuh Universal Lintas-Gambar**

Pada bagian ini, kami menjelaskan cara melatih CIUA-Watermark untuk melindungi beberapa gambar wajah dari satu jaringan modifikasi wajah. Secara umum, setiap jaringan modifikasi wajah yang terlatih dengan baik memiliki distribusi data input yang dapat diterima yang berarti bahwa hanya gambar wajah yang memenuhi distribusi jaringan yang dapat diterima yang dapat dimodifikasi dengan benar. Karena data pelatihan berskala besar, hampir semua gambar wajah alami memenuhi distribusi yang dapat diterima ini. Namun, gangguan lawan dapat menyebabkan gambar wajah alami keluar dari distribusi yang dapat diterima. Jadi, CIUA-Watermark adalah gangguan yang mencoba mengarahkan semua gambar wajah alami keluar dari distribusi yang dapat diterima dari model tertentu seperti yang ditunjukkan pada Gambar 2.

Kami menggunakan I-FGSM [29] untuk menghasilkan gangguan pada setiap iterasi serangan. Gangguan-gangguan ini selanjutnya digabungkan untuk membentuk gangguan universal dengan faktor peluruhan m:



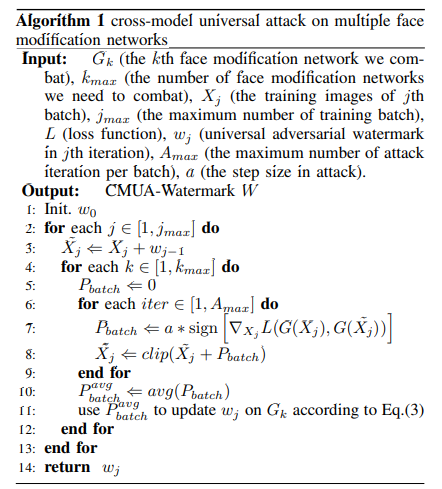
di mana *bs* adalah ukuran batch citra wajah, *Abase* adalah algoritma serangan dasar (mis, I-FGSM [29] atau PGD [30]) yang mengembalikan gangguan lawan P dari citra wajah bersih xj, L adalah fungsi kerugian (kami mengeksploitasi MSE seperti yang dirumuskan dalam Persamaan. (1)), G adalah jaringan modifikasi wajah yang kita serang, a adalah ukuran langkah pada serangan dasar, wi adalah perturbasi universal pelatihan pada iterasi ke-i, m ∈ (0, 1) adalah parameter untuk mengontrol rasio pembaruan pada setiap iterasi, E adalah nilai batas atas dari tanda air pelatihan, dan klip operasi membatasi nilai wi pada kisaran [-E, E].

Melatih perturbasi universal tidaklah mudah karena perturbasi di setiap iterasi yang dihasilkan oleh gambar wajah yang berbeda dapat saling bertentangan satu sama lain. Untuk mengatasi masalah ini, kami mengusulkan metode serangan universal berbasis batch. Dalam setiap iterasi, kami memasukkan gambar wajah bs untuk menghitung gangguan bs. Kemudian, untuk mengurangi konflik dari gangguan-gangguan ini dan membuat watermark universal fokus pada atribut umum dari wajah manusia, kami mendapatkan gangguan rata-rata dengan mengambil rata-rata dari gangguan-gangguan ini untuk memperbarui gangguan universal pelatihan.

## **Serangan Musuh Universal Lintas-Model**

Serangan adversarial universal lintas-model didasarkan pada serangan lintas-citra, dan perlu lebih lanjut menemukan titik temu distribusi yang tidak dapat diterima dari beberapa model. Seluruh pipeline yang kami latih untuk CMUA-Watermark ditunjukkan pada Gambar 3.

Pertama, kami memasukkan sekumpulan citra wajah bersih dan sekumpulan citra wajah bersih dengan gangguan universal awal ke beberapa model modifikasi wajah. Nilai gangguan universal awal tergantung pada metode serangan dasar yang kita pilih (misalnya, semua-nol untuk I-FGSM dan derau acak untuk PGD). Kemudian, untuk setiap model, kami menghitung kehilangan MSE dari outputnya dan menggunakan I-FGSM untuk mendapatkan perturbasi setiap gambar. Setelah itu, kami menghitung rata-rata gangguan dari setiap model pada setiap iterasi (setiap batch). Terakhir, kami menggunakan Persamaan (3) untuk memperbarui CMUA-Watermark, dan kami memperbarui sebanyak n kali per iterasi untuk menyerang n model. Setelah pelatihan yang cukup, kita bisa mendapatkan CMUA-Watermark yang efektif untuk melindungi gambar wajah. Detail dari serangan lintas model kami dapat dirujuk ke Algoritma 1.



# Experiment

In this section, we first describe our datasets and implementation details in Section 4.1. Then, we introduce our evaluation metrics in Section 4.2. After that, we show the experimental results of CIUA-Watermark and CMUAWatermark in Section 4.3 and Section 4.4 respectively. Furthermore, we systemically investigate the influence of hyperparameters on the watermark in Section 4.5. Finally, in Section 4.6, we test the effectiveness of CMUA-Watermark in more real scenes.

## Datasets and Implementation Details

In our experiments, we use CelebA [31] test set as the main dataset which contains 19962 facial images. To further evaluate the generalization ability of our CMUA-Watermark, we also generate adversarial examples on LFW [32] dateset with the same watermark (only trained on CelebA test set). Besides, we also randomly select 100 facial images from films as additional data to verify the effectiveness of the CMUA-Watermark in more real scenes. It is important to point out that we do not use the additional data to train our adversarial watermark.