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

***Abstract***

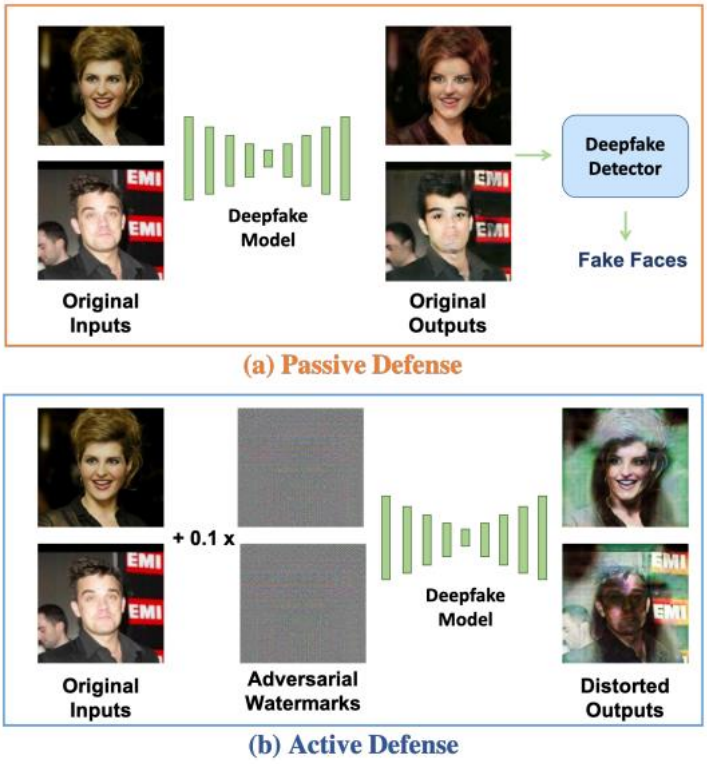
Aplikasi berbahaya dari *deepfake* (yaitu teknologi yang menghasilkan atribut wajah target atau seluruh wajah dari gambar wajah) telah menjadi ancaman besar bagi reputasi dan keamanan individu. Untuk mengurangi ancaman ini, penelitian terbaru telah mengusulkan *watermark* yang berlawanan untuk memerangi model *deepfake*, yang membuat mereka menghasilkan *output* yang terdistorsi. Meskipun mencapai hasil yang mengesankan, watermark *adversarial* ini memiliki kemampuan transferabilitas tingkat gambar dan tingkat model yang rendah, yang berarti bahwa watermark ini hanya dapat melindungi satu gambar wajah dari satu model *deepfake* tertentu. Untuk mengatasi masalah ini, kami mengusulkan solusi baru yang dapat menghasilkan Cross-Model Universal *Adversarial* Watermark (*CMUA-Watermark*), yang melindungi sejumlah besar gambar wajah dari beberapa model *deepfake*. Secara khusus, kami mulai dengan mengusulkan sebuah pipa serangan universal lintas model yang menyerang beberapa model *deepfake* secara berulang-ulang. Kemudian, kami merancang strategi fusi perturbasi dua tingkat untuk mengurangi konflik antara *watermark* yang berlawanan yang dihasilkan oleh gambar dan model wajah yang berbeda. Selain itu, kami mengatasi masalah utama dalam optimasi lintas model dengan pendekatan heuristik untuk secara otomatis menemukan ukuran langkah serangan yang sesuai untuk model yang berbeda, yang selanjutnya melemahkan konflik tingkat model. Terakhir, kami memperkenalkan metode evaluasi yang lebih masuk akal dan komprehensif untuk menguji metode yang diusulkan dan membandingkannya dengan metode yang sudah ada. Hasil eksperimen yang ekstensif menunjukkan bahwa *CMUA-Watermark* yang diusulkan dapat

secara efektif mendistorsi gambar wajah palsu yang dihasilkan oleh beberapa model *deepfake* sambil mencapai kinerja yang lebih baik daripada metode yang ada. Kode kami tersedia di[1].

***Introduction***

Baru-baru ini, peningkatan Generative Adversarial Networks (GAN) telah menunjukkan hasil yang mengesankan dalam pembuatan konten virtual, menciptakan nilai ekonomi dan hiburan yang cukup besar. Namun, deepfakes, jaringan modifikasi wajah berbasis pembelajaran mendalam yang menggunakan GAN untuk menghasilkan konten palsu dari orang yang ditargetkan atau atribut target, telah menyebabkan kerusakan besar pada privasi dan reputasi orang. Di satu sisi, gambar dan video palsu dapat menunjukkan hal-hal yang tidak pernah dikatakan atau dilakukan oleh seseorang, sehingga merusak reputasinya, terutama jika melibatkan pornografi atau politik (Tolosana et al. 2020). Di sisi lain, gambar wajah palsu dengan atribut target dapat melewati otentikasi biometrik aplikasi komersial, yang berpotensi melanggar keamanan (Korshunov dan Marcel 2018). Oleh karena itu, mempertahankan ancaman yang dibawa oleh deepfakes tidak hanya membutuhkan distorsi gambar yang dimodifikasi dan menurunkan kualitas visualnya untuk membantu manusia dalam membedakannya dari gambar yang realistis, tetapi juga memastikan bahwa wajah palsu tidak lolos deteksi kehidupan, yang merupakan langkah pertama dari sebagian besar verifikasi biometrik.

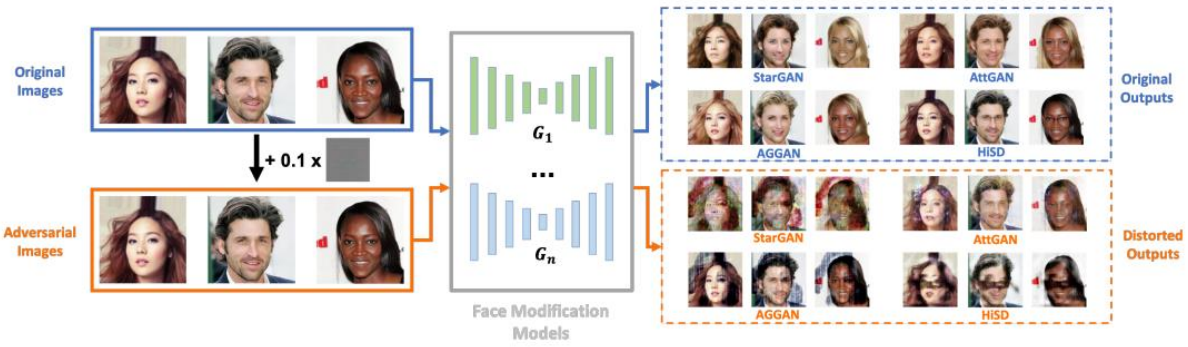
[1] https://github.com/VDIGPKU/CMUA-Watermark




Gambar 1: Pertahanan pasif menggunakan detektor *deepfake* hanya dapat mengurangi bahaya *deepfake* dengan mendeteksi gambar yang dimodifikasi, sementara pertahanan aktif menggunakan tanda air yang berlawanan untuk mengacaukan model *deepfake* untuk menghasilkan keluaran yang terdistorsi secara nyata, sehingga dapat memitigasi risiko sebelumnya.

Secara umum, cara utama untuk memitigasi risiko deepfake adalah pertahanan pasif, yaitu melatih pendeteksi deepfake untuk mendeteksi konten yang telah dimodifikasi (Afchar dkk. 2018; Tariq, Lee, dan Woo 2021; Zhao dkk. 2021; Sun dkk. 2021; Chen dkk. 2021). Detektor semacam itu pada dasarnya adalah pengklasifikasi biner, yang memprediksi apakah sebuah gambar dipalsukan oleh model deepfake atau tidak. Namun, melindungi gambar wajah dengan cara ini sama seperti menutup pintu kandang setelah kuda berlari kencang; efek dan bahaya yang ditimbulkan tidak dapat dibalik sepenuhnya dan risikonya tetap ada. Baru-baru ini, (Ruiz, Bargal, dan Sclaroff 2020) menyarankan untuk menggunakan tanda air yang berlawanan untuk memerangi model deepfake, yang membuat mereka menghasilkan output

yang tampak tidak nyata. Karena tanda air ini dapat ditambahkan ke gambar wajah terlebih dahulu, tanda air ini dapat menghindari risiko penggunaan deepfake yang berbahaya setelahnya sebagai pertahanan aktif. Perbandingan skematis dari kedua cara tersebut diilustrasikan pada Gambar 1. Meskipun (Ruiz, Bargal, dan Sclaroff 2020) dapat mempertahankan potensi ancaman, ia hanya dapat menghasilkan tanda air musuh yang spesifik untuk gambar dan model, yang berarti bahwa tanda air tersebut hanya dapat melindungi satu gambar wajah dari satu model deepfake tertentu.




Gambar 2: Ilustrasi Tanda Air CMUA kami. Setelah tanda air CMUA dibuat, kita dapat menambahkannya secara langsung ke gambar wajah apa pun untuk menghasilkan gambar yang dilindungi yang secara visual identik dengan gambar asli tetapi dapat mendistorsi keluaran model *deepfake* seperti StarGAN (Choi dkk. 2018), AGGAN (Tang dkk. 2019), AttGAN (He dkk. 2019), dan HiSD (Li dkk. 2021).

Untuk mengatasi masalah ini, kami mengusulkan solusi yang efektif dan efisien dalam penelitian ini, yang menggunakan sejumlah kecil (sekecil 128) citra wajah pelatihan untuk membuat crossmodel universal *adversarial* watermark (*CMUA-Watermark*) untuk melindungi sejumlah besar citra wajah dari berbagai model *deepfake*, seperti yang digambarkan pada Gambar 2. Pertama, kami mengusulkan pendekatan serangan universal lintas-model berdasarkan metode serangan vanilla yang hanya dapat melindungi satu gambar tertentu dari satu model, yaitu PGD (Madry et al. 2018). Secara khusus, untuk mengurangi

konflik di antara *watermark* yang berlawanan yang dihasilkan dari gambar dan model yang berbeda, kami baru saja mengusulkan strategi fusi perturbasi dua tingkat (yaitu, fusi tingkat gambar dan fusi tingkat model) selama proses serangan universal lintas model. Kedua, untuk lebih melemahkan konflik di antara *watermark* yang berlawanan yang dihasilkan dari model yang berbeda sehingga meningkatkan transferabilitas *CMUA-Watermark* yang dihasilkan, kami mengeksploitasi algoritme *Tree-Structured Parzen Estimator* (TPE) (Bergstra dkk., 2011) untuk secara otomatis menemukan ukuran langkah serangan untuk model yang berbeda.

Selain itu, metode evaluasi yang ada dalam (Ruiz, Bargal, dan Sclaroff 2020) tidak masuk akal dan cukup komprehensif. Pertama, mengukur distorsi gambar dengan menghitung jarak $L_1$ atau $L_2$ secara langsung antara seluruh output asli dan terdistorsi mengabaikan *deepfake* yang hanya memodifikasi beberapa atribut (misalnya, HiSD (Li et al. 2021) hanya dapat menambahkan sepasang kacamata), karena distorsi yang terukur akan dirata-ratakan oleh area lain yang tidak berubah. Sebagai gantinya, kami mengusulkan untuk menggunakan masker modifikasi untuk lebih fokus pada area yang dimodifikasi. Kedua, hanya mempertimbangkan jarak $L_1$ atau $L_2$ saja tidak cukup; untuk memastikan perlindungan juga diperlukan metrik yang mencerminkan kualitas pembangkitan dan karakteristik biologis dari output yang terdistorsi. Oleh karena itu, kami menggunakan *Frechet Inception Distance* (FID) (Heusel et al. 2017) untuk mengukur kualitas generasi dan mengeksploitasi nilai kepercayaan serta tingkat kelulusan model deteksi kelangsungan hidup untuk mengukur karakteristik biologis dari keluaran yang terdistorsi.

Kontribusi kami dapat diringkas sebagai berikut:

- Kami adalah yang pertama kali memperkenalkan ide baru untuk menghasilkan *watermark* permusuhan universal lintas model (*CMUA-Watermark*) untuk

melindungi citra wajah manusia dari beberapa pemalsuan, hanya membutuhkan 128 citra wajah pelatihan untuk melindungi banyak sekali citra wajah.

- Kami mengusulkan strategi fusi gangguan yang sederhana namun efektif untuk meredakan konflik dan meningkatkan kemampuan transferabilitas tingkat gambar dan tingkat model dari *watermark* CMUA yang diusulkan.
- Kami menganalisis secara mendalam proses optimasi lintas model dan mengembangkan algoritma penyetelan ukuran langkah otomatis untuk menemukan ukuran langkah serangan yang sesuai untuk model yang berbeda.
- Kami memperkenalkan metode evaluasi yang lebih masuk akal dan komprehensif untuk sepenuhnya mengevaluasi efektivitas *perturbations watermark* dalam memerangi *deepfake*s.

***Related Works***

***Face Modification***

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 yang lebih realistis dengan target orang atau atribut. StarGAN (Choi et al. 2018) mengusulkan pendekatan baru dan terukur untuk melakukan penerjemahan gambar-ke-gambar di berbagai domain, mencapai kualitas visual yang lebih baik pada gambar yang dihasilkan. Kemudian, AttGAN (He et al. 2019) menggunakan batasan klasifikasi atribut untuk memberikan gambar wajah yang lebih alami pada manipulasi atribut wajah. Selain itu, AGGAN (Tang et al. 2019) memperkenalkan attention mask melalui mekanisme perhatian bawaan untuk mendapatkan gambar target dengan kualitas tinggi. Baru-baru ini, (Li et al. 2021) mengusulkan HiSD yang merupakan metode penerjemahan gambar-ke-gambar yang canggih untuk skalabilitas beberapa label dan keragaman yang dapat dikontrol dengan pelepasan yang mengesankan. Meskipun model-model ini mengadopsi

beragam arsitektur dan kerugian, *watermark* CMUA kami berhasil mencegah gambar wajah dimodifikasi dengan benar oleh semuanya.

***Attacks on Generative Models***

Sudah ada beberapa penelitian (Wang, Cho, dan Yoon 2020; Yeh dkk. 2020; Ruiz, Bargal, dan Sclaroff 2020; Kos, Fischer, dan Song 2018; Tabacof, Tavares, dan Valle 2016; Bashkirova, Usman, dan Saenko 2019) yang mengeksplorasi serangan lawan terhadap model generatif, dan kami secara khusus berfokus pada tugas penerjemahan gambar yang menjadi dasar dari *deepfake*. (Wang, Cho, dan Yoon 2020) dan (Yeh et al. 2020) menyerang tugas penerjemahan gambar pada CycleGAN (Zhu et al. 2017) dan pix2pixHD (Wang et al. 2018), yang hanya memindahkan gambar di antara dua domain dan dengan demikian relatif mudah untuk diserang. (Ruiz, Bargal, dan Sclaroff 2020) adalah yang pertama kali menangani serangan pada jaringan penerjemahan gambar bersyarat, tetapi *watermark* yang mereka hasilkan hanya melindungi gambar tertentu dari model *deepfake* tertentu, yang berarti bahwa setiap *watermark* yang diserang harus dilatih secara individual, yang memakan waktu dan tidak mungkin dilakukan pada kenyataannya.

Table 1: The categories of adversarial watermarks.

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

***Universal Adversarial Perturbation***

*Universal Adversarial* Perturbation pertama kali diperkenalkan oleh (Moosavi-Dezfooli et al. 2017), di mana sebuah model pengenalan dapat dikelabui hanya dengan satu perturbasi musuh. Di sini Universal berarti bahwa gangguan tunggal dapat ditambahkan ke

beberapa gambar untuk menipu model tertentu. Berdasarkan karya ini, (Metzen et al. 2017) memperkenalkan gangguan permusuhan universal untuk tugas segmentasi untuk menghasilkan hasil target, dan (Li et al. 2019) pertama kali mengusulkan serangan permusuhan universal pada sistem pengambilan gambar, yang membuat mereka mengembalikan gambar yang tidak relevan. Karya-karya yang disebutkan di atas hanya menghasilkan watermark *adversarial* universal yang menargetkan satu model, sedangkan *CMUA-Watermark* kami dapat memerangi beberapa model modifikasi wajah secara bersamaan.

***Methods***

Pada bagian ini, pertama-tama kami menyajikan gambaran umum metode kami. Kemudian, kami memperkenalkan cara menyerang model modifikasi wajah tunggal. Selanjutnya, kami menjelaskan strategi fusi gangguan. Terakhir, kami menganalisis masalah utama dari optimasi lintas model dan menyarankan algoritma penyetelan ukuran langkah

otomatis untuk mencari ukuran langkah serangan yang sesuai.

**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** $k \in [1, o]$ **do**
  **for** $i \in [1, m]$ **do**
    $P^i_{k_1}, ..., P^i_{k_{bs}} \Leftarrow A(G_i, a_i, X_k, W_{i+(k-1)m-1})$
    $P^i_{avg} \Leftarrow$ Image-Level Fusion with $P^i_{k_1}, ..., P^i_{k_{bs}}$
    $W_{i+(k-1)m} \Leftarrow$ Model-Level Fusion with $P^i_{avg}$
  **end for**
**end for**
$W_{cmua} = W_{m \cdot o}$

***Overview***

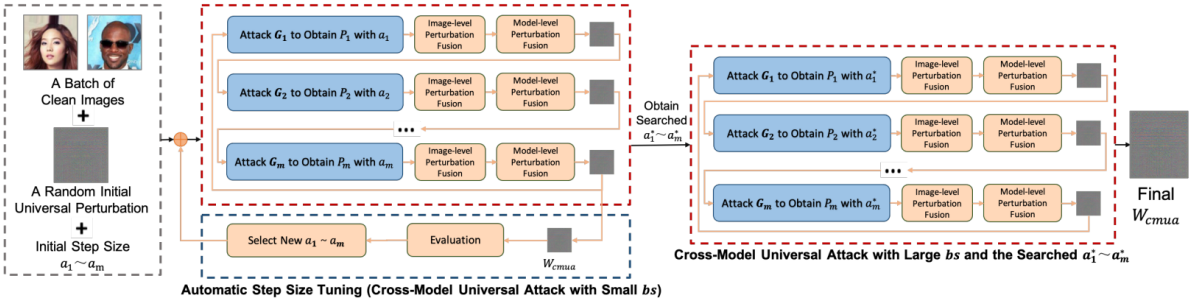
Pada Tabel 1, kami mengkategorikan *watermark* berlawanan berdasarkan kemampuan generalisasi lintas-gambar dan lintas-model. Berbeda dengan Single-Image *Adversarial* Watermark (SIAWatermark) yang melindungi gambar tertentu terhadap model tertentu dan *Universal Adversarial* Watermark (UAWatermark) yang melindungi banyak gambar terhadap model tertentu, *CMUA-Watermark* yang diusulkan dalam makalah ini dapat memerangi beberapa model *deepfake* sekaligus melindungi banyak sekali gambar wajah.

Seperti yang diilustrasikan pada Gambar 3, keseluruhan pipeline kami untuk membuat *CMUA-Watermark* dibagi menjadi dua langkah.

**Pada langkah pertama**, kami berulang kali melakukan serangan universal model silang dengan ukuran batch yang kecil (untuk pencarian yang lebih cepat), mengevaluasi *CMUA-Watermark* yang dihasilkan, dan kemudian menggunakan penyetelan ukuran langkah otomatis untuk memilih ukuran langkah serangan yang baru $a_1$, ..., $a_m$.

**Pada langkah kedua**, kami menggunakan ukuran langkah yang ditemukan $a^*_1$, ..., $a^*_m$ untuk melakukan serangan universal model silang dengan ukuran *batch* yang besar (untuk meningkatkan kemampuan mengganggu) dan menghasilkan *CMUA-Watermark* akhir.

Secara khusus, seperti yang ditunjukkan pada Algoritma 1, selama proses serangan universal cross-model yang diusulkan, kumpulan gambar *input* secara iteratif melalui serangan PGD (Madry et al. 2018) untuk menghasilkan gangguan yang berlawanan, yang kemudian melalui mekanisme penggabungan gangguan dua tingkat untuk digabungkan ke dalam *CMUA-Watermark* gabungan yang berfungsi sebagai gangguan awal untuk model berikutnya.




Gambar 3: Keseluruhan *pipeline dari Cross-Model Universal Adversarial Attack* pada beberapa jaringan modifikasi wajah.


Gambar 4: Proses terperinci dari menyerang satu model *deepfake* tertentu.

***Combating One Face Modification Model***

Pada bagian ini, kami menjelaskan pendekatan untuk mengganggu model modifikasi wajah tunggal dalam *pipeline* kami (kotak biru pada Gambar 3), yang diilustrasikan secara rinci pada Gambar 4. Untuk memulai,

- kami memasukkan sekumpulan gambar bersih $I_1$... $I_n$ ke model *deepfake* $G$ dan mendapatkan output asli $G(I_1)$... $G(I_n)$.
- Kemudian, kami memasukkan $I_1$... $I_n$ dengan gangguan awal $W$ ke $G$,
- mendapatkan output terdistorsi awal $G(I_1 + W)$... $G(I_n + W)$.
- Selanjutnya, kami menggunakan Mean Square Error (MSE) untuk mengukur perbedaan antara $G(I_1)$... $G(I_n)$ dan $G(I_1 + W)$... $G(I_n + W)$,

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

di mana E adalah nilai batas atas dari *perturbation watermark* W. Terakhir, kami menggunakan PGD (Madry et al. 2018) sebagai metode serangan dasar untuk memperbarui *adversarial perturbation* pada setiap iterasi serangan,

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

- di mana $I$ adalah citra wajah yang bersih,
- $I^r_{adv}$ adalah citra wajah lawan pada iterasi ke-$r$,
- $a$ adalah ukuran langkah dari serangan dasar,
- $L$ adalah fungsi kerugian (kami memilih MSE seperti yang dirumuskan pada Persamaan (2)),
- $G$ adalah jaringan modifikasi wajah yang kami serang,
- dan clip operasi membatasi $I_{adv}$ pada rentang [$I$ - $E$, $I$ + $E$].

Melalui proses ini, kita dapat memperoleh *Watermark Single-ImageAdversarial* (SIA-*Watermark*) yang melindungi satu gambar wajah dari satu model *deepfake* tertentu. Namun, SIA-Watermark yang dibuat tidak memadai di bawah pengaturan lintas model; mereka kurang dalam hal transferabilitas tingkat gambar dan model. Dalam dua bagian berikut ini, kami memperkenalkan solusi kami untuk mengatasi masalah ini.

***Adversarial Perturbation Fusion***

Konflik di antara watermark yang berlawanan yang dihasilkan dari gambar dan model yang berbeda akan mengurangi kemampuan transferabilitas *CMUA-Watermark* yang diusulkan. Untuk melemahkan konflik ini, kami mengusulkan strategi fusi gangguan dua tingkat selama proses serangan. Secara khusus, ketika kami menyerang satu model *deepfake* tertentu, kami melakukan **fusi tingkat gambar** untuk merata-rata gradien yang di *sign* dari sekumpulan gambar wajah,

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

- di mana *bs* adalah ukuran kumpulan gambar wajah,
- dan $I^{adv}_j$ adalah gambar lawan ke-j dari sebuah kumpulan.

Operasi ini akan menyebabkan $G_{avg}$ lebih berkonsentrasi pada atribut umum wajah manusia daripada atribut wajah tertentu.

Kemudian, kita menggunakan PGD untuk menghasilkan *adversarial* perturbation $P_{avg}$ melalui $G_{avg}$ seperti persamaan (2).

Setelah mendapatkan $P_{avg}$ dari satu model,

kami melakukan **fusi tingkat model**, yang secara iteratif menggabungkan $P_{avg}$ yang dihasilkan dari model tertentu ke $W_{CMUA}$ dalam pelatihan, dan $W_{CMUA}$ awal hanyalah $P_{avg}$ yang dihitung dari model *deepfake* pertama,

$$W^0_{CMUA} = P^0_{avg},$$
$$W^{t+1}_{CMUA} = \alpha \cdot W^t_{CMUA} + (1-\alpha) \cdot P^t_{avg}, \quad (4)$$

- di mana $\alpha$ adalah faktor peluruhan,
- $P^t_{avg}$ adalah rata-rata gangguan yang dihasilkan dari model *deepfake* yang diserang ke-*t*,
- dan $W^t_{CMUA}$ adalah *CMUA-Watermark* pelatihan setelah model *deepfake* yang diserang ke-*t*.

***Automatic Step Size Tuning based on TPE***

Selain fusi dua tingkat yang disebutkan di atas, kami menemukan bahwa ukuran langkah serangan untuk model yang berbeda juga penting untuk transferabilitas *CMUA-Watermark* yang dihasilkan. Oleh karena itu, kami mengeksploitasi pendekatan heuristik untuk secara otomatis menemukan ukuran langkah serangan yang sesuai.

Metode serangan dasar yang kami pilih (PGD) termasuk ke dalam keluarga FGSM (Goodfellow, Shlens, dan Szegedy 2015), dan gradien $\nabla_x L$ dinormalisasi oleh fungsi *sign*:

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

Dalam perhitungan nyata, elemen-elemen dalam $\nabla_x L$ hampir tidak pernah mencapai 0, sehingga $\|\text{sign}(\nabla_x L)\|_2 \approx 1$ adalah tetap untuk setiap gradien. Perturbasi $\Delta P$ yang diperbarui dalam iterasi Attack method berbasis *sign* dirumuskan sebagai:

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

Dengan kata lain, hanya ukuran langkah *a* yang menentukan tingkat pembaruan selama serangan, sehingga pemilihan *a* memiliki pengaruh yang besar terhadap performa serangan. Kesimpulan ini juga berlaku untuk serangan universal lintas model; perturbasi yang diperbarui $\Delta P^u$ dalam sebuah iterasi serangan universal lintas model dibentuk dengan menggabungkan $\Delta P^i$ dari beberapa model $G_1$,..., $G_m$:

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

Dalam rumus di atas,

- *m* adalah jumlah model,
- faktor peluruhan α adalah sebuah konstanta, dan
- *sign* ($\nabla_X L_i$) memberikan arah optimasi untuk $G_i$ .

Oleh karena itu, arah optimasi secara keseluruhan sangat dipengaruhi oleh $a_1$, ..., $a_m$, dan memilih $a_1$, ..., $a_m$ yang sesuai di berbagai model untuk menemukan arah keseluruhan yang ideal adalah masalah utama untuk serangan lintas model.

Kami memperkenalkan algoritma TPE (Bergstra et al. 2011) untuk memecahkan masalah ini, yang secara otomatis mencari $a_1$, ..., $a_m$ yang sesuai untuk menyeimbangkan arah yang berbeda yang dihitung dari berbagai model. TPE adalah metode optimasi hiper-parameter berdasarkan *Sequential Model-Based Optimization* (SMBO), yang secara berurutan membangun model untuk memperkirakan kinerja hiper-parameter berdasarkan pengukuran historis, dan kemudian memilih hiperparameter baru untuk diuji berdasarkan model ini. Dalam tugas kami, kami menganggap ukuran langkah $a_1$, ..., $a_m$ sebagai hyperparameter input *x* dan tingkat keberhasilan serangan sebagai nilai kualitas y dari TPE. TPE menggunakan $P(x|y)$ dan $P(y)$ untuk memodelkan $P(y|x)$, dan $p(x|y)$ diberikan oleh:

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

di mana $y^*$ ditentukan oleh pengamatan terbaik secara historis, e($x$) adalah densitas yang dibentuk dengan pengamatan {$x(i)$} sedemikian rupa sehingga kerugian yang sesuai lebih rendah dari $y^*$ , dan $g(x)$ adalah densitas yang dibentuk dengan pengamatan yang tersisa. Setelah memodelkan $P(y|x)$, kami terus mencari ukuran langkah yang lebih baik dengan mengoptimalkan kriteria *Expected Improvement* (EI) di setiap iterasi pencarian, yang diberikan oleh,

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

di mana $\gamma = p(\, y < y^* \,)$. Dibandingkan dengan kriteria lainnya, EI bersifat intuitif dan telah terbukti memiliki kinerja yang sangat baik. Untuk detail lebih lanjut mengenai TPE, lihat (Bergstra et al. 2011).

Table 2: The quantitative results of CMUA-Watermark.

| Dataset | Model | $L^2_{mask}\uparrow$ | $SR_{mask}\uparrow$ | FID↑ | ACS↓ | TFHC↓ |
|---|---|---|---|---|---|---|
| CelebA | StarGAN | 0.20 | 100.00% | 201.003 | 0.286 | 66.26%→20.61% |
| | AGGAN | 0.13 | 99.88% | 50.959 | 0.863 | 65.88%→55.52% |
| | AttGAN | 0.05 | 87.08% | 65.063 | 0.638 | 55.13%→28.05% |
| | HiSD | 0.11 | 99.87% | 92.734 | 0.153 | 63.30%→3.94% |
| LFW | StarGAN | 0.20 | 100.00% | 169.329 | 0.207 | 43.88%→8.20% |
| | AGGAN | 0.13 | 99.99% | 37.746 | 0.806 | 54.90%→46.32% |
| | AttGAN | 0.06 | 94.07% | 70.640 | 0.496 | 25.86%→16.73% |
| | HiSD | 0.10 | 98.13% | 88.145 | 0.314 | 50.68%→16.03% |
| Film100 | StarGAN | 0.20 | 100.00% | 259.716 | 0.425 | 61.01%→29.09% |
| | AGGAN | 0.13 | 99.88% | 129.099 | 0.832 | 60.98%→55.69% |
| | AttGAN | 0.07 | 95.82% | 177.499 | 0.627 | 34.56%→25.83% |
| | HiSD | 0.11 | 100.00% | 220.689 | 0.207 | 67.00%→14.00% |

***Experiment***

Pada bagian ini, pertama-tama kami akan menjelaskan dataset dan detail implementasi kami. Setelah itu, kami akan memperkenalkan metrik evaluasi kami. Kemudian, kami menunjukkan hasil eksperimen dari *CMUA-Watermark* yang diusulkan. Selain itu, kami secara sistemik melakukan studi ablasi. Terakhir, kami menunjukkan aplikasi model watermark yang diusulkan dalam adegan yang realistis.

***Datasets and Implementation Details***

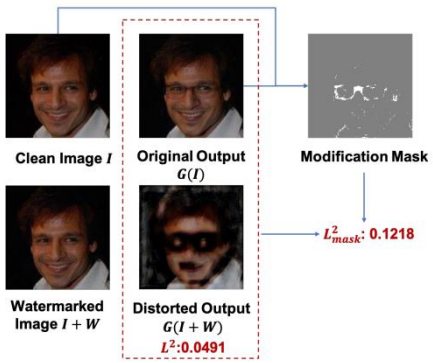
Dalam percobaan kami, kami menggunakan set uji CelebA (Liu et al. 2015) sebagai set data utama, yang berisi 19962 gambar wajah. Kami menggunakan 128 gambar pertama dalam set tersebut sebagai gambar pelatihan dan mengevaluasi metode kami pada semua gambar wajah dari set uji CelebA dan dataset LFW (Huang et al. 2007) untuk memastikan kredibilitasnya. Selain itu, kami juga secara acak memilih 100 gambar wajah dari film sebagai data tambahan (Films100) untuk memverifikasi keefektifan *CMUA-Watermark* dalam skenario nyata. Penting untuk dicatat bahwa kami tidak menggunakan data tambahan apa pun untuk melatih *CMUA-Watermark*.

Jaringan modifikasi wajah yang kami pilih dalam eksperimen kami adalah StarGAN (Choi et al. 2018), AGGAN (Tang et al. 2019), AttGAN (He et al. 2019), dan HiSD (Li et al. 2021). StarGAN dan AGGAN dilatih pada dataset CelebA untuk lima atribut: rambut hitam, rambut pirang, rambut cokelat, jenis kelamin, dan usia. AttGAN dilatih pada dataset CelebA hingga empat belas atribut, yang lebih rumit dibandingkan dengan dua jaringan di atas. Kami juga menyerang salah satu jaringan modifikasi wajah terbaru, HiSD, yang juga dilatih pada dataset CelebA dan dapat menambahkan sepasang kacamata pada orang yang ditargetkan.

Selama proses pencarian ukuran langkah, jumlah maksimum iterasi adalah 1k, dan ruang pencarian ukuran langkah untuk setiap model adalah [0, 10]. Pertama-tama kita mencari ukuran langkah dengan batchsize = 16 dan kemudian menggunakan ukuran langkah yang telah dicari untuk melakukan serangan model silang dengan batchsize = 64.

***Evaluation Metrics***

Mempertimbangkan keterbatasan metode evaluasi yang ada dan setelah memikirkan kembali tujuan memerangi model *deepfake*, kami merancang metode evaluasi yang lebih masuk akal dan komprehensif, yang berkonsentrasi pada metrik tiga aspek.




Gambar 5: Kasus bermasalah untuk metode evaluasi yang ada dan topeng modifikasi yang diusulkan untuk mengatasi masalah ini.

Sebagai permulaan, kami menganalisis keberhasilan proses modifikasi. (Ruiz, Bargal, dan Sclaroff 2020) menghitung skor $L^2$ antara keluaran asli G(I) dan keluaran terdistorsi G(I

+ W) dan menyatakan bahwa citra wajah berhasil dilindungi oleh watermark lawan ketika ||G(I + W) - G(I) ||$_2$> 0,05. Namun, seperti yang ditunjukkan pada Gambar 5, metode evaluasi ini bermasalah untuk beberapa kasus. Sebagai contoh, output terdistorsi G(I + W) sangat berbeda dengan output asli G(I) di area modifikasi, terutama di sekitar mata, sehingga berhasil mencegah model *deepfake* untuk menambahkan kacamata ke gambar wajah. Namun, evaluasi yang ada menganggap output tersebut sebagai kasus yang gagal. Untuk mengatasi masalah ini, kami memperkenalkan matriks topeng, yang lebih berkonsentrasi pada area yang dimodifikasi,

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

di mana (i, j) adalah koordinat piksel dalam gambar. Dengan cara ini, ketika menghitung $L^2_{mask}$, hanya piksel dengan perubahan besar yang akan dihitung dan area lainnya akan ditinggalkan,

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

Dalam eksperimen kami, jika $L^2_{mask}$ > 0,05, kami menentukan bahwa gambar berhasil dilindungi, dan menggunakan $SR_{mask}$ untuk merepresentasikan tingkat keberhasilan melindungi gambar wajah.

Kedua, kami menggunakan FID (Heusel et al. 2017) untuk mengukur kualitas pembuatan gambar wajah palsu. FID secara komprehensif merepresentasikan jarak vektor fitur dari Inception v3 (Szegedy et al. 2016) antara gambar asli dan gambar palsu, dan nilai FID yang lebih tinggi mengindikasikan kualitas yang lebih rendah dari gambar yang dihasilkan.

Terakhir, kami menggunakan sistem pendeteksi kehidupan sumber terbuka HyperFAS[2] untuk menguji karakteristik biologis dari gambar-gambar palsu tersebut. HyperFAS didasarkan pada Mobilenet (Howard et al. 2017) dan dilatih dengan 360 ribu gambar wajah. Dalam percobaan kami, jika skor kepercayaan lebih besar dari 0,99, kami menyimpulkan bahwa wajah tersebut adalah wajah asli dengan kepercayaan tinggi (TFHC). Selain itu, kami juga menghitung skor kepercayaan rata-rata ACS untuk evaluasi.

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

| 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** | **0.9994** | 0.02 | 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** | 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** | 0.9988 | **0.8708** | **0.9987** | 2.3032 | 1.7072 | **1.8133** | **1.9672** |

***The Results of CMUA-Watermark***

Kami melakukan eksperimen ekstensif untuk menunjukkan keefektifan *CMUA-Watermark* yang diusulkan. Pertama-tama kami menunjukkan hasil kuantitatif dan kualitatif dari *CMUA-Watermark* yang diusulkan dan kemudian membandingkan metode kami dengan metode serangan canggih.

Hasil kuantitatif dan kualitatif dari *CMUA-Watermark* yang diusulkan dilaporkan pada Tabel 2 dan Gambar 2. *CMUA-Watermark* yang diusulkan memiliki kinerja keseluruhan yang serupa pada CelebA dan LFW dan berkinerja lebih baik pada StarGAN, AGGAN, dan HiSD daripada AttGAN. Secara khusus, SRmask yang memerangi StarGAN, AGGAN, dan HiSD mendekati 100% pada CelebA dan LFW, dan ACS dari output yang terdistorsi menurun secara signifikan dibandingkan dengan output asli, yang membuat TFHC StarGAN, AGGAN, AttGAN, dan HiSD masing-masing turun 45,65%, 10,36%, 27,08%, dan 59,36% pada set uji

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

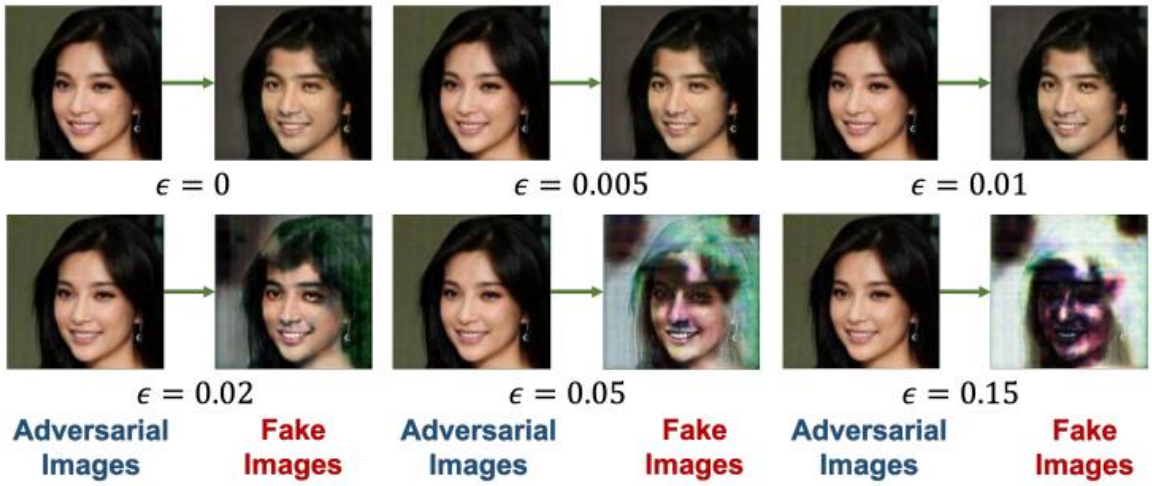
CelebA dan 35,68%, 8,58%, 9,13%, dan 34,65% pada LFW. Selain itu, pada dataset Film100 yang lebih dekat dengan adegan nyata, watermark yang diusulkan berkinerja lebih baik dibandingkan dengan dua dataset di atas. Selain itu, semua output yang terdistorsi memiliki FID yang besar, yang menunjukkan kualitas pembangkitan yang buruk. Secara keseluruhan, hasil kualitatif dan kuantitatif di atas menunjukkan bahwa *CMUA-Watermark* berhasil melindungi gambar wajah dari beberapa model *deepfake*.

Kami selanjutnya membandingkan *CMUA-Watermark* dengan metode serangan canggih pada CelebA, termasuk BIM (Kurakin, Goodfellow, dan Bengio 2017), MIM (Dong dkk. 2018), PGD (Madry dkk. 2018), DI2-FGSM (Xie dkk. 2019), MDI2-FGSM (Xie dkk. 2019), AutoPGD (Croce dan Hein 2020). Kami mengacu pada (Moosavi-Dezfooli et al. 2017) dan menyesuaikan metode-metode ini dengan pengaturan universal (penjelasan rinci pada Lampiran F). Hasil perbandingan SRmask dan FID dilaporkan pada Tabel 3, dan kami dapat mengamati bahwa tanda air lawan yang dibuat oleh metode yang dibandingkan (misalnya MIM) terlalu mengoptimalkan satu atau dua model sehingga berkinerja sangat buruk pada model lainnya. Sebaliknya, metode kami mencapai kinerja yang sangat baik pada semua model. Hasil ini menunjukkan transferabilitas tingkat gambar dan tingkat model yang lebih baik dari metode yang diusulkan dibandingkan dengan metode yang sudah ada.

***Ablation Study***

Pada bagian ini, pertama-tama kami menyelidiki keefektifan fusi gangguan dan penyetelan ukuran langkah otomatis. Kemudian, kami menyelidiki pengaruh hiperparameter lain pada CMUA-Watermark yang diusulkan. Seperti yang dilaporkan pada Tabel 4, algoritma PGD dasar memiliki kinerja yang sangat buruk pada AttGAN dan HiSD, yang mengindikasikan bahwa transferabilitas tingkat modelnya lemah. Ketika secara terpisah menggunakan strategi fusi gangguan, kinerja keseluruhan meningkat, tetapi hasil untuk StarGAN turun secara signifikan. Di sisi lain, jika kita hanya melakukan penyetelan ukuran

langkah otomatis, hasil untuk semua model telah ditingkatkan, tetapi hasil untuk AttGan dan HiSD masih belum cukup baik. Setelah menggabungkan keduanya, performa CMUA-Watermark kami meningkat secara signifikan untuk semua model deepfake. Hasil ini menunjukkan bahwa penggabungan gangguan dan penyetelan ukuran langkah otomatis sangat penting untuk metode yang diusulkan dan harus digunakan bersama-sama.




Gambar 6: Contoh CMUA-Watermark dengan pengaturan berbeda dari .

Kami juga menyelidiki pengaruh pada watermark yang diusulkan dengan perubahan algoritma base attack dan batas atas. Seperti yang ditunjukkan pada Tabel 5, dengan metode yang diusulkan, mengubah metode base attack memiliki pengaruh yang kecil terhadap CMUA-Watermark. Selain itu, seperti yang diilustrasikan pada Gambar 6, gambar wajah palsu yang dihasilkan lebih terdistorsi ketika parameter menjadi lebih besar, yang berarti bahwa kinerja proteksi menjadi lebih baik. Namun, ketika menjadi terlalu besar, watermark lawan yang dihasilkan lebih mungkin terlihat. Kami secara empiris menemukan bahwa pengaturan sekitar 0.05 dapat membuat trade-off yang baik antara kinerja perlindungan dan ketidaktampakan tanda air lawan yang dihasilkan.

***Conclusion***

Dalam makalah ini, kami telah mengusulkan sebuah pipeline serangan universal lintas model untuk menghasilkan sebuah watermark yang dapat melindungi sejumlah besar gambar

wajah dari beberapa model deepfake. Secara khusus, kami mengusulkan strategi fusi gangguan untuk mengurangi konflik tanda air yang berlawanan yang dihasilkan dari gambar dan model yang berbeda dalam proses serangan. Lebih lanjut, kami menganalisis masalah utama dari optimasi lintas model dan memperkenalkan algoritma penyetelan ukuran langkah otomatis berdasarkan TPE untuk menentukan arah optimasi secara keseluruhan. Selain itu, kami merancang metode evaluasi yang lebih masuk akal dan komprehensif untuk mengevaluasi CMUA-Watermark yang diusulkan. Hasil eksperimen menunjukkan bahwa CMUA-Watermark kami dapat secara efektif mengganggu modifikasi oleh model deepfake, menurunkan kualitas gambar yang dihasilkan, dan mencegah gambar wajah palsu melewati verifikasi sistem deteksi kehidupan.

Step

1. Input data
2. Preposesing
    a. Data di proteksi awal dengan *Cmua* acak.
    b. Data ditambahkan inisiasi *step size* $a_1 \sim a_m$.
3. Proses mencari Auto Step Size Tuning
    a. Proses perulangan sampai batch yang ditentukan
        i. Prose mencari perturbation untuk Satu model deepfake
            1. Data gambar bersih diserang model deepfake,
            2. data gambar yang diproteksi serang model deepfake
            3. Melakukan proses perbandingan dengan metode MSE
            4. Mencari model perturbation baru dengan PGD
            5. Kembali lagi melakukan iterasi ke awal dengan model Perturbation baru sampai dengan mendapatkan semua model perturbation.
            6. Menentukan Adversarial Perturbation Fusion tingkat gambar
        ii. Proses Menentukan Adversarial Perturbation Fusion tingkat Model
        iii. Menghasilkan $W_{cmua}$
        iv. Evaluasi
4. Proses mencari $W_{cmua}$ final dengan auto Step Size tuning.
    a. Proses perulangan sampai batch yang ditentukan
        i. Prose mencari perturbation untuk Satu model deepfake
            1. Data gambar bersih diserang model deepfake,
            2. data gambar yang diproteksi serang model deepfake

3. Melakukan proses perbandingan dengan metode MSE
4. Mencari model perturbation baru dengan PGD
5. PGD meghasilkan model Perturbation baru
6. Kembali lagi melakukan iterasi ke awal sampai dengan mendapatkan semua model perturbation.
7. Menentukan Adversarial Perturbation Fusion tingkat gambar

ii. Proses Menentukan Adversarial Perturbation Fusion tingkat Model

iii. Menghasilkan $W_{cmua}$