USULAN PENELITIAN

Deteksi Lip-Syncing Deepfake dengan Analisis Inkoherensi Gerakan Mulut

Oleh:

Ikhlasul Amal	**24/541027/PPA/06821**
Magnolia Gina R. S.	**24/551548/PPA/06978**
Nilam Mufidah	**24/551986/PPA/06994**
Sisi Florensia	**24/552780/PPA/07018**

PROGRAM MAGISTER KECERDASAN ARTIFISIAL
DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA
2025

A. Latar Belakang

Perkembangan pesat pada grafik komputer dan teknologi *generative* AI membawa pada terciptanya *deepfake*. *Deepfake* merupakan media yang dimanipulasi secara digital yang meliputi audio, video, dan gambar palsu dari suatu objek seperti kendaraan, hewan, maupun manusia. Selain menimbulkan kekhawatiran etikal, video *deepfake* juga menawarkan manfaat yang positif jika digunakan secara bertanggung jawab. Beberapa kegunaan positif dari video *deepfake*, antara lain meningkatkan *dubbing* dan teknologi manipulasi visual pada film, membantu pelatihan medis dengan menciptakan simulasi pasien yang realistis, dan membantu melestarikan warisan budaya dengan menciptakan kembali tokoh-tokoh sejarah dan menghidupkan kembali bahasa-bahasa yang terancam punah. Di sisi lain, perkembangan *deepfake* memiliki resiko yang besar, terutama dalam hal disinformasi, pencurian identitas, dan penipuan. *Deepfake* digunakan untuk menciptakan video palsu tentang seseorang yang mengatakan sesuatu yang tidak sebenarnya mereka katakan. Hal ini dapat berujung pada penyebaran informasi palsu, manipulasi opini publik, hingga pencemaran nama baik. Maraknya penyalahgunaan *deepfake* menjadi ancaman pada integrasi media digital sehingga diperlukan suatu alat yang mampu untuk mendeteksi video *deepfake*.

Sinkronisasi bibir (*lip-syncing*) merupakan jenis video *deepfake* di mana gerakan bibir seseorang dimanipulasi secara digital agar sinkron dengan audio tertentu. Modul Mouth Spatial-Temporal Inconsistency Extractor (MSTIE) (Datta, Jia and Lyu, 2025) menggunakan pendekatan identifikasi pola inkonsistensi spasial-temporal guna membedakan video *lipsync* dari video asli. Modul ini memproses informasi spasial dan temporal dari area mulut, mengekstraksi fitur-fitur halus yang penting untuk membedakan video asli dari *deepfake*. Dengan mengintegrasikan frame lokal dan global, model ini mampu menangkap variasi jangka pendek dan jangka panjang dalam gerakan mulut, meningkatkan kemampuannya untuk mendeteksi inkonsistensi secara efektif. Analisis ganda yang konsisten ini memungkinkan model untuk melakukan pemeriksaan komprehensif, memanfaatkan konteks lokal dan temporal. Menghasilkan kinerja yang efektif dalam mengidentifikasi *deepfake* sinkronisasi bibir.

B. Rumusan Masalah

Permasalah yang dijumpai adalah pendekatan-pendekatan yang telah dilakukan sebelumnya yang berfokus pada gerak dan sinkronisasi masih mengalami kesulitan terkait *false positive* dan performanya berkurang saat dihadapkan pada video *deepfake* sinkronisasi bibir yang kompleks maupun yang dimanipulasi sebagian.

C. Batasan Masalah

Batasan masalah pada penelitian ini, antara lain:

1. Jenis video *deepfake* yang digunakan adalah sinkronisasi bibir (*lipsync*).
2. Modul yang digunakan berfokus pada informasi di area mulut.

3. Metrik evaluasi yang digunakan adalah Average Precision (AP), Area Under ROC Curve (AUC), dan Intersection over Union (IoU).

D. Tujuan

Penelitian ini bertujuan untuk mengembangkan sistem deteksi video yang di generasi oleh kecerdasan artifisial yang lazimnya dikenal sebagai video *deepfake*. Fokus penelitian ini adalah pada aspek analisis inkonsistensi spasial-temporal area mulut dengan tujuan spesifik sebagai berikut:

1. Mengembangkan metode ekstraksi frame mulut yang optimal melalui kombinasi local frames dan global frames berdasarkan tingkat keterbukaan mulut (mouth openness) untuk menangkap konsistensi jangka pendek dan jangka panjang pada pergerakan bibir.
2. Merancang mekanisme untuk mendeteksi inkonsistensi halus pada pergerakan mulut dalam video *deepfake* yang menggunakan konsep *lip-sync* dan mengimplementasikan fungsi loss untuk meningkatkan sensitivitas model terhadap manipulasi *lip-sync*.
3. Mengevaluasi performa sistem menggunakan metrik *Average Precision (AP)*, *Area Under ROC Curve (AUC)*, dan *Intersection over Union (IoU)* untuk mengukur akurasi klasifikasi dan kemampuan lokalisasi segmen video yang dimanipulasi.

E. Manfaat

Manfaat dari penelitian ini akan dapat digunakan secara praktis, sosial maupun secara akademis.

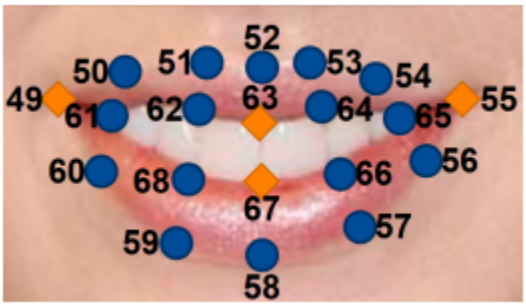
1. Manfaat praktis dari penelitian ini adalah teknologi deteksi video *deepfake* yang lebih akurat untuk verifikasi konten digital, terutama pada aplikasi-aplikasi platform media sosial, forensik digital, dan peningkatan keamanan autentikasi video.
2. Secara sosial, penelitian ini dapat bermanfaat untuk melindungi masyarakat dari disinformasi video melalui peningkatan keamanan siber autentikasi video
3. Di bidang akademis, penelitian ini akan berkontribusi sebagai metodologi baru dalam deteksi *deepfake* berbasis spasial-temporal, pengembangan arsitektur untuk analisis inkonsistensi video, dan inovasi pemilihan frame berbasis mouth openness.

F. Metodologi

Pada penelitian ini, tujuan utama adalah mendeteksi serta mengklasifikasikan video menjadi asli (real) atau *deepfake*. Metodologi yang digunakan terdiri dari tiga tahapan utama, yaitu ekstraksi frame mulut lokal dan global, ekstraksi inkonsistensi *spasial-temporal* mulut, dan fungsi loss untuk pelatihan model. Setelah proses pelatihan, model dievaluasi menggunakan metrik *Average Precision (AP)*, *Area Under ROC Curve (AUC)*, dan *Intersection over Union (IoU)*.

1. *Local and Global Mouth Frame Extractor*

Tahap ini bertujuan untuk mengisolasi area mulut dari setiap frame video yang diuji. Proses dilakukan dengan menggunakan detektor wajah seperti *Dlib*, yang kemudian digunakan untuk melakukan cropping serta alignment wajah pada tiap frame, sehingga area mulut berada di posisi yang konsisten. Proses ekstraksi area mulut diawali dengan penggunaan *Dlib face detector* untuk mendeteksi wajah pada setiap frame video, kemudian dilakukan alignment wajah agar posisi fitur wajah, khususnya mulut, konsisten antar-frame. Setelah wajah terdeteksi, digunakan *Dlib 68-point facial landmark predictor* untuk mengidentifikasi titik-titik penting pada wajah, di mana bagian mulut direpresentasikan oleh *landmark 49–68*. Dari titik-titik ini, khususnya titik 63 (bagian atas bibir dalam) dan 67 (bagian bawah bibir dalam), dihitung jarak vertikal sebagai ukuran keterbukaan mulut, serta jarak horizontal antar-sudut mulut sebagai pembanding. Hasil landmark tersebut digunakan untuk membentuk bounding box area mulut dengan tambahan padding, kemudian dipotong (*cropping*) dan diubah ukurannya menjadi 64×144 piksel sesuai input model. Proses ini memastikan bahwa hanya area mulut yang relevan, dengan informasi keterbukaan dan pergerakan bibir, gigi, serta lidah, yang digunakan sebagai *RGB mouth frames* untuk tahap pemrosesan lebih lanjut dalam mendeteksi inkonsistensi pada video deepfake.



Perhitungan mouth openness dilakukan dengan memanfaatkan landmark wajah yang dihasilkan oleh Dlib 68-point facial landmark predictor, khususnya titik 63 (bagian atas bibir dalam) dan 67 (bagian bawah bibir dalam). Jarak vertikal antar kedua titik tersebut dihitung dengan rumus Euclidean distance sebagai ukuran utama keterbukaan mulut:

$$H_m(t) = ||p_{63}(t) - p_{67}(t)||_2$$

Selain itu, jarak horizontal antar sudut mulut (landmark kiri dan kanan bibir) dihitung sebagai lebar mulut:

$$W_m(t) = ||p_{left\ corner}(t) - P_{right\ corner}(t)||_2$$

Untuk menghindari pengaruh skala wajah akibat jarak kamera atau pose kepala, nilai keterbukaan mulut kemudian dinormalisasi dalam bentuk rasio:

$$S(t) = \frac{H_m(t)}{W_m(t) + \epsilon}$$

dengan ϵ sebagai nilai kecil untuk mencegah pembagian nol. Nilai $S(t)$ yang tinggi menandakan mulut lebih terbuka, sedangkan nilai rendah menandakan mulut lebih tertutup. Hasil pengukuran ini digunakan sebagai dasar dalam pemilihan local frames (L) yang berurutan untuk menangkap konsistensi jangka pendek, serta global frames (G) yang tidak berurutan dengan tingkat keterbukaan mulut serupa untuk menangkap konsistensi jangka panjang pada video deepfake.

Pada tahap pemilihan frame, algoritma difokuskan untuk memperoleh dua jenis frame, yaitu local frames (L) dan global frames (G), yang keduanya ditentukan berdasarkan nilai *mouth openness* hasil perhitungan landmark Dlib.

a. Local frames (L): dipilih dalam bentuk sekumpulan frame yang berurutan di sekitar frame pusat t^* dengan mulut terbuka, misalnya mengambil $[L/2]$ frame sebelum dan sesudah t^*. Pemilihan ini bertujuan untuk menangkap hubungan spasial-temporal jangka pendek antar-frame berdekatan.
b. Global frames (G): dipilih dari frame yang tidak berurutan namun memiliki tingkat keterbukaan mulut $S(t)$ yang serupa dengan frame pusat, dengan syarat terdapat jarak waktu minimal sebesar 0,09 detik atau $[0.09 \times fps]$ frame agar tidak terlalu berdekatan. Frame global ini berfungsi untuk mengukur konsistensi jangka panjang, sehingga kombinasi local dan global frames memungkinkan model mendeteksi inkonsistensi pergerakan mulut baik dalam rentang pendek maupun panjang

2. Delta Frames

Selain pemilihan local frames (L) dan global frames (G), algoritma juga menambahkan proses perhitungan delta frames (D) untuk menyoroti perubahan antar-frame secara berurutan. Setelah didapatkan urutan RGB mouth frames $R = \{R_1, R_2, R_3, \ldots, R_N\}$ dengan $N = L + G$, maka delta frames dihitung sebagai selisih intensitas piksel antar-frame:

$$D_t = R_{t+1} - R_t, \qquad 1 \leq t \leq N$$

dengan hasil berupa urutan $D \in R^{(N-1)\times H\times W\times 3}$, di mana $H \times W$ adalah dimensi citra (misalnya 64×144) dan 3 adalah jumlah kanal RGB. Delta frames ini secara khusus menyoroti perubahan bentuk, tepi, serta kontur mulut dan gigi antar-frame yang sering kali lebih jelas menunjukkan inkonsistensi lip-sync deepfake dibanding hanya menggunakan frame RGB biasa. Dengan demikian, kombinasi local frames,

global frames, dan delta frames memberikan model representasi yang lebih kaya untuk mendeteksi inkonsistensi jangka pendek maupun panjang pada pergerakan mulut.

3. Mouth Spatial-Temporal Inconsistency Extractor

Mouth Spatial-Temporal Inconsistency Extractor (MSTIE) merupakan modul utama untuk menangkap inkonsistensi halus pada pergerakan mulut dalam video lip-sync deepfake. Modul ini menerima dua masukan, yaitu RGB mouth frames $R = \{R_1, R_2, R_3, \ldots, R_N\}$ yang merupakan hasil cropping area mulut, dan delta frames $D = \{D_1, D_2, D_3, \ldots, D_{N-1}\}$ yang dihitung dari selisih antar-frame dengan rumus:

$$D_t = R_{t+1} - R_t, \quad 1 \leq t \leq N$$

Delta frames ini menonjolkan perubahan struktural seperti kontur bibir, gigi, dan lidah yang sering kali menjadi indikator manipulasi. MSTIE kemudian memanfaatkan Vision Temporal Transformer (VTT) untuk mengekstraksi fitur spasial dan temporal secara bersamaan. Pertama, setiap frame dibagi ke dalam patch dan diproyeksikan ke dalam embedding, kemudian diproses oleh spatial transformer encoder untuk memodelkan hubungan spasial dalam satu frame, serta temporal transformer encoder untuk memodelkan hubungan antar-frame. Proses ini dapat diformulasikan sebagai:

$$Z^{(l)} = MSA(LN(Z^{(l-1)})) + Z^{(l-1)}, \quad Z^{(l+1)} = MLP(LN(Z^{(l)})) + Z^{(l)}$$

di mana MSA adalah multi-head self-attention, MLP adalah multilayer perceptron, dan LN adalah layer normalization. Setelah fitur dari jalur RGB (v_R) dan delta (v_D) diperoleh, keduanya digabung dengan mekanisme multi-head cross-attention agar model dapat secara adaptif menekankan aspek komplementer dari informasi visual statis dan perubahan temporal. Mekanisme cross-attention ini didefinisikan sebagai:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

dengan query Q berasal dari salah satu jalur dan key-value (K, V) dari jalur lainnya (misalnya delta). Hasil akhirnya adalah representasi fitur gabungan yang kemudian digunakan untuk klasifikasi, apakah video termasuk asli atau deepfake. Dengan memanfaatkan kombinasi RGB frames, delta frames, transformer encoder, dan cross-attention, MSTIE mampu mendeteksi inkonsistensi spasial-temporal jangka pendek maupun panjang yang tidak terlihat jelas pada frame.

4. Loss Function

Fungsi loss yang digunakan tidak hanya bertujuan mengklasifikasikan video sebagai asli atau deepfake, tetapi juga menekankan pada deteksi inkonsistensi spasial-temporal pada area mulut. Terdapat dua komponen utama dalam loss function ini, yaitu Classification Loss (LCL) dan Inconsistency Loss (LIL).

a. Classification Loss (LCL)
Komponen ini menggunakan Binary Cross-Entropy (BCE) untuk melatih model dalam membedakan kelas real (1) dan fake (0). Rumus BCE didefinisikan sebagai:

$$L_{CL} = -\frac{1}{M}\sum_{i=1}^{M}[y_i \, log(\widehat{y}_i) + (1 - y_i) \, log(1 - \widehat{y}_i)]$$

dengan M jumlah sampel, $y_i \in \{0, 1\}$ label ground truth, dan $\widehat{y}_i$ probabilitas prediksi dari model. Fungsi ini memastikan model mampu melakukan klasifikasi biner secara langsung.

b. Inconsistency Loss (LIL)

Untuk mendeteksi inkonsistensi halus antar-frame, digunakan Structural Similarity Index (SSIM) yang mengukur kesamaan struktural antara dua frame. SSIM bernilai mendekati 1 pada video asli (konsisten antar-frame), dan lebih rendah pada video deepfake. Nilai loss didefinisikan sebagai:

$$L_{IL} = 1 - SSIM(R_t, R_{t+1})$$

di mana R_t dan R_{t+1} adalah dua frame RGB mulut yang berurutan. Dengan cara ini, semakin besar perbedaan struktural antar-frame, semakin tinggi nilai loss, sehingga model terdorong lebih sensitif terhadap manipulasi.

c. ***Total Loss***

Kombinasi kedua komponen tersebut menghasilkan total loss yang digunakan dalam proses pelatihan:

$$L_{total} = \lambda_1 L_{CL} + \lambda_2 L_{IL}$$

Pada penelitian, digunakan $\lambda_1 = 1$ dan $\lambda_1 = 5$, sehingga inkonsistensi antar-frame lebih ditekankan dibanding sekadar klasifikasi.

Dengan desain ini, fungsi loss tidak hanya melatih model untuk mengenali pola global real dengan fake, tetapi juga mendeteksi ketidaksesuaian halus dalam pergerakan mulut yang menjadi ciri khas lip-sync deepfake.

Dalam penelitian ini, metode evaluasi yang digunakan mencakup tiga metrik utama,

1. ***Average Precision (AP)***

Average Precision mengukur keseimbangan antara presisi dan recall pada berbagai ambang batas klasifikasi. Presisi didefinisikan sebagai:

$$Precision = \frac{TP}{TP+FP}$$

dan recall sebagai:

$$Recall = \frac{TP}{TP+FN}$$

dengan TP (True Positive), FP (False Positive), dan FN (False Negative). Nilai AP diperoleh dari rata-rata presisi pada seluruh tingkat recall, sehingga semakin tinggi nilai AP berarti model mampu mempertahankan presisi meski recall meningkat.

2. ***Area under ROC Curve (AUC)***

AUC mengukur kemampuan model dalam membedakan kelas asli dan deepfake dengan menghitung area di bawah kurva ROC (Receiver Operating Characteristic). Kurva ROC menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR):

$$TPR = \frac{TP}{TP+FN}, \qquad FPR = \frac{FP}{FP+TN}$$

Nilai AUC mendekati 1 menunjukkan performa model sangat baik dalam memisahkan dua kelas, sementara nilai mendekati 0,5 menunjukkan performa setara dengan tebak acak.

3. ***Intersection Over Union (IoU)***

IoU digunakan untuk mengevaluasi segment-wise localization, yaitu akurasi model dalam mendeteksi segmen video yang dimanipulasi. IoU dihitung berdasarkan

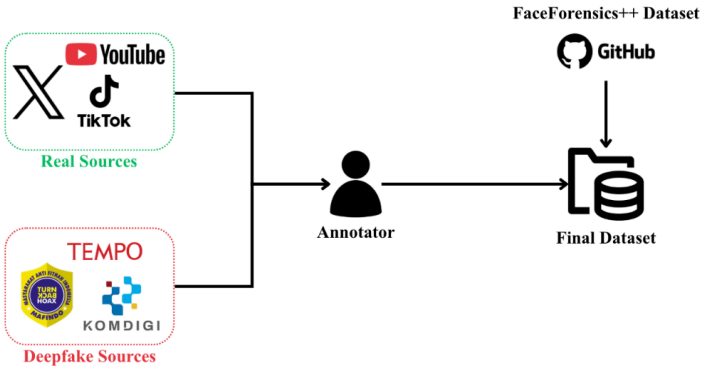
perbandingan luas irisan (intersection) antara segmen prediksi dan ground truth terhadap luas gabungan (union) keduanya:

$$IoU = \frac{|Segmen_{pred} \cap Segmen_{gt}|}{|Segmen_{pred} \cup Segmen_{gt}|}$$

Semakin tinggi nilai IoU, semakin akurat model dalam menemukan lokasi dan durasi manipulasi pada video deepfake.

G. Dataset

Penelitian ini menggunakan FaceForensics++ dataset yang tersedia di GitHub (https://github.com/ondyari/FaceForensics/tree/master/dataset). Dataset ini digunakan sebagai benchmark dalam penelitian deteksi *deepfake* karena mencakup video asli maupun hasil manipulasi dengan berbagai metode, seperti DeepFakes, Face2Face, FaceSwap, dan NeuralTextures. Dalam kasus *lip-sync manipulation* pada penelitian ini, dataset tersebut disesuaikan dengan cara melakukan ekstraksi area mulut dari setiap frame video menggunakan Dlib face detector dan 68-point facial landmark predictor, sebagaimana dijelaskan pada bagian metodologi.



Proses pengumpulan dan anotasi dataset

Selain itu, penelitian ini juga menyusun dataset tambahan untuk pengujian yang lebih kontekstual dengan kasus di Indonesia. Dataset ini dibuat dengan mengumpulkan video *deepfake* yang teridentifikasi sebagai *lip-sync manipulation* dari beberapa portal cek fakta resmi, antara lain TurnBackHoax (turnbackhoax.id), Tempo Cek Fakta (tempo.co/cekfakta), dan Komdigi (komdigi.go.id/berita/berita-hoaks). Proses pengumpulan dilakukan dengan menggunakan kata kunci tertentu seperti *"deepfake"* untuk menelusuri artikel klarifikasi hoaks video *deepfake*. Sementara itu, video yang dikategorikan sebagai *real* diperoleh dari sejumlah platform media sosial antara lain TikTok, YouTube Shorts, dan X/Twitter. Video-video yang diperoleh berupa potongan wawancara, pernyataan publik tokoh tertentu,

ataupun konten percakapan singkat yang menampilkan ekspresi wajah dan gerakan mulut yang jelas dalam konteks yang sesuai dengan kasus di Indonesia.

Untuk menjaga kualitas, seluruh data dalam dataset tambahan ini melalui proses anotasi manual oleh *annotator*. Seorang *annotator* dari penulis meninjau setiap video guna memverifikasi kesesuaian label, apakah termasuk kategori *real* atau *deepfake.* Pada tahap ini, *annotator* juga bertugas memastikan tidak terdapat duplikasi, baik berupa video yang sama persis maupun versi serupa dengan sedikit perbedaan.

H. Kesimpulan

Penelitian ini mengusulkan metode deteksi video *deepfake* yang fokus pada inkonsistensi spasial-temporal area mulut melalui tiga komponen utama: ekstraksi frame mulut berbasis mouth openness, arsitektur MSTIE dengan *Vision Temporal Transformer*, dan fungsi *loss* gabungan. Keunggulan metode ini adalah kemampuan mendeteksi inkonsistensi jangka pendek dan panjang melalui kombinasi lokal-global *frames* serta delta *frames*, dengan mekanisme *cross-attention* yang menggabungkan informasi RGB dan temporal secara adaptif. Evaluasi menggunakan metrik AP, AUC, dan IoU diharapkan menunjukkan superioritas metode yang diusulkan. Penelitian ini berkontribusi pada pengembangan teknologi deteksi *deepfake* yang lebih akurat dan memberikan solusi praktis untuk memerangi disinformasi dalam konten video digital.

I. Daftar Pustaka

Datta, S.K., Jia, S. and Lyu, S. (2025) “Detecting Lip-Syncing Deepfakes: Vision Temporal Transformer for Analyzing Mouth Inconsistencies.” arXiv. Available at: https://doi.org/10.48550/arXiv.2504.01470.

D E King, “Dlib-ml: A machine learning toolkit,” The Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann, “Video transformer network,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3163–3172.

Z Liu, J Ning, Y Cao, Y Wei, Z Zhang, S Lin, and H Hu, “Video swin transformer,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

J Huang and C X Ling, “Using auc and accuracy in evaluating learning algorithms,” IEEE Transactions on knowledge and Data Engineering, vol. 17, no. 3, pp. 299–310, 2005.

Z Yan, Y Luo, S Lyu, Q Liu, and B Wu, “Transcending forgery specificity with latent space augmentation for generalizable deepfake detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8984–8994.

P Kwon, J You, G Nam, S Park, and G Chae, “Kodf: A large-scale korean deepfake detection dataset,” in Proceedings of the ICCV, 2021, pp. 10744–10753.