HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 2, Desember 2026, p417-429
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no2

IMPLEMENTASI ARSITEKTUR HIBRIDA XCEPTION-TRANSFORMER DENGAN PENDEKATAN MULTI-DOMAIN PADA DETEKSI VIDEO DEEPFAKE

Tubagus Muhammad Rayhan[1], Odi Nurdiawan[2], Nisa Dienwati Nuris[3].

[123]Program Studi Teknik Informatika, STMIK IKMI Cirebon.
Jl. Perjuangan No 10B Kesambi Kota Cirebon – Jawa Barat, Indonesia
Email: [1]rayhan0706@gmail.com, [2]odinurdiawan2020@gmail.com, [3]nisadienwatinuris@gmail.com

ABSTRAK

Penelitian ini bertujuan merancang dan mengevaluasi keandalan arsitektur hibrida Xception-Transformer berbasis fusi multi-domain untuk mendeteksi manipulasi video deepfake. Subjek penelitian menggunakan dataset sekunder Celeb-DF v2, yang memiliki tingkat ketimpangan kelas tinggi (5.639 sampel manipulasi dan 890 asli). Pemilihan sampel dilakukan melalui pembagian data pada tingkat video (*video-level split*) untuk mencegah kebocoran data, dengan mengekstraksi 8 bingkai gambar secara sekuensial. Pendekatan metodologis ini menggabungkan fitur domain spasial (RGB), frekuensi (FFT), dan temporal (Optical Flow) ke dalam tensor fusi awal. Teknik analisis data mengukur performa prediksi melalui metrik akurasi, kurva ROC, dan uji ablasi, didukung pengaplikasian fungsi *Binary Focal Crossentropy* serta pembobotan kelas dinamis. Hasil pengujian menunjukkan bahwa pada 38 siklus pelatihan, model mencapai tingkat akurasi 98% dan skor AUC 0,985. Model terbukti sukses menangani ketimpangan kelas dengan mereduksi tingkat False Positive menjadi hanya 13 kasus. Uji ablasi membuktikan bahwa penghapusan Transformer menurunkan akurasi drastis ke tingkat 81%, mengonfirmasi krusialnya mekanisme atensi temporal. Analisis *Explainable* AI melalui proyeksi t-SNE dan Grad-CAM mengonfirmasi pemisahan fitur yang optimal dan lokalisasi visual pada area transisi wajah. Kesimpulannya, fusi Xception-Transformer sangat andal dalam forensik digital. Penelitian selanjutnya merekomendasikan evaluasi silang dataset dan penerapan distilasi pengetahuan guna efisiensi komputasi real-time.

Kata kunci: deepfake, Xception-Transformer, multi-domain, explainable AI, celeb-DF

ABSTRACT

This study aims to design and evaluate the reliability of a multi-domain fusion-based Xception-Transformer hybrid architecture for deepfake video detection. The research subject utilizes the Celeb-DF v2 secondary dataset, which exhibits high class imbalance (5,639 manipulated and 890 authentic samples). Sample selection was conducted via video-level data splitting to prevent data leakage, sequentially extracting 8 frames per video. This methodological approach concatenates spatial (RGB), frequency (FFT), and temporal (Optical Flow) domain features into an early fusion tensor. The data analysis technique measures predictive performance through accuracy metrics, ROC curves, and an ablation study, supported by Binary Focal Crossentropy and dynamic class weighting. Experimental results demonstrate that after 38 training epochs, the model achieved an accuracy of 98% and an AUC score of 0.985. The model successfully handled class imbalance, significantly reducing False Positives to 13 cases. The ablation study proves that excluding the Transformer drastically drops accuracy to 81%, confirming the critical role of temporal attention mechanisms. Explainable AI analysis via t-SNE projections and Grad-CAM confirmed optimal feature separation and visual localization on facial transition areas. In conclusion, the Xception-Transformer fusion is highly reliable in digital forensics. Future research recommends cross-dataset evaluation and knowledge distillation for real-time computational efficiency.

Keywords: deepfake, Xception-Transformer, multi-domain, explainable AI, celeb-DF

1. PENDAHULUAN

Perkembangan pesat teknologi *deep learning* dan *Generative Adversarial Networks* (GANs) telah memicu peningkatan kualitas sintesis media digital, salah satunya adalah teknologi deepfake [1], [2]. Teknologi ini memungkinkan manipulasi wajah dan ekspresi seseorang dalam rekaman video dengan tingkat kehalusan visual yang sangat tinggi [3], [4]. Ketersediaan algoritma *open-source* mempermudah pembuatan konten *deepfake*, yang berdampak pada munculnya ancaman serius terhadap integritas informasi digital, seperti kampanye disinformasi, penipuan finansial, dan pencemaran nama baik [5], [6]. Hal tersebut memunculkan berbagai dampak negatif di masyarakat modern [7], [8]. Dampak nyata dari penyalahgunaan ini bermanifestasi dalam bentuk kampanye disinformasi politik [9], ancaman pornografi non-konsensual dan pencemaran nama baik [10], serta memicu

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

berbagai tren ancaman keamanan siber lainnya. Kondisi ini mendesak kebutuhan akan sistem deteksi *deepfake* otomatis yang andal dan akurat [11], [12], [13].

Dalam merespons ancaman tersebut, penelitian deteksi *deepfake* umumnya mendayagunakan arsitektur *Convolutional Neural Network* (CNN) [14]. Arsitektur seperti ResNet atau EfficientNet terbukti efektif mengenali anomali visual statis (spasial), seperti artefak penyatuan piksel (*blending boundaries*) atau ketidaksejajaran geometri wajah pada bingkai tunggal [15], [16]. Namun, manipulasi video modern sering kali meninggalkan jejak pada inkonsistensi sekuensial yang luput dari pandangan CNN murni. Hal ini mendorong tren penggunaan arsitektur *Transformer* yang mengandalkan mekanisme *self-attention* untuk memodelkan ketergantungan jarak jauh (*long-range dependencies*) antarbingkai guna menangkap anomali pergerakan mikro [17], [18], [19], [20], [21]. Sayangnya, penggunaan *Transformer* secara mandiri membutuhkan beban komputasi kuadratik yang besar dan kurang optimal dalam mengekstraksi tekstur mikroskopis karena tidak memiliki bias induktif (*inductive bias*) layaknya operasi konvolusi [2], [22].

Beberapa penelitian terdahulu (State-of-the-Art / SOTA) telah mencoba menggabungkan CNN dan *Transformer* untuk deteksi *deepfake* [23], [24], [25], [26]. Akan tetapi, mayoritas model hibrida SOTA saat ini masih sangat bertumpu pada masukan domain tunggal, yaitu citra spasial RGB. Pendekatan ini rentan terkecoh oleh dataset generasi terbaru seperti Celeb-DF v2, yang memiliki kualitas rekayasa resolusi tinggi sehingga mampu menyembunyikan artefak kasatmata pada domain spasial murni [18], [27]. Selain itu, sebagian besar model deteksi mutakhir masih bersifat "kotak hitam" (*black-box*), di mana sistem mampu memberikan akurasi tinggi namun sering gagal menjelaskan dasar pengambilan keputusannya secara visual [28], [29].

Untuk mengatasi batasan dari penelitian-penelitian sebelumnya, penelitian ini mengusulkan arsitektur hibrida spasio-temporal yang mengintegrasikan *backbone* Xception dan *Transformer Encoder*. Berbeda dengan pendekatan SOTA terdahulu, kebaruan utama (*novelty*) dari penelitian ini terletak pada pemanfaatan fusi awal ekstraksi fitur multi-domain yang menggabungkan domain spasial (RGB), domain frekuensi (*Fast Fourier Transform*), dan domain temporal (*Optical Flow*) sebelum diproses oleh mekanisme atensi. Pemilihan Xception sebagai *backbone*, alih-alih model *pre-trained* lain seperti ResNet atau VGG, didasari pada keunggulan efisiensi arsitektur *Depthwise Separable Convolution*. Xception secara empiris lebih sensitif dan komputasional lebih ringan dalam menangkap anomali tekstur resolusi rendah yang sering ditinggalkan oleh algoritma generator GAN [30], [31].

Tujuan dari penelitian ini adalah merancang dan mengevaluasi keandalan model hibrida Xception-Transformer pada dataset Celeb-DF v2, dengan menerapkan metode pembobotan kelas (*class weight*) dan fungsi kerugian *Binary Focal Crossentropy* untuk menangani ketimpangan data secara dinamis. Kontribusi saintifik yang ditargetkan adalah menghasilkan arsitektur deteksi yang tangguh dengan akurasi tinggi, seraya menyediakan transparansi inferensi. Keputusan klasifikasi dari model ini akan dievaluasi dan diinterpretasikan melalui pendekatan *Explainable AI* (XAI) menggunakan pemetaan Grad-CAM dan reduksi klastering t-SNE sebagai referensi visual pendukung bagi analisis forensik digital [28].

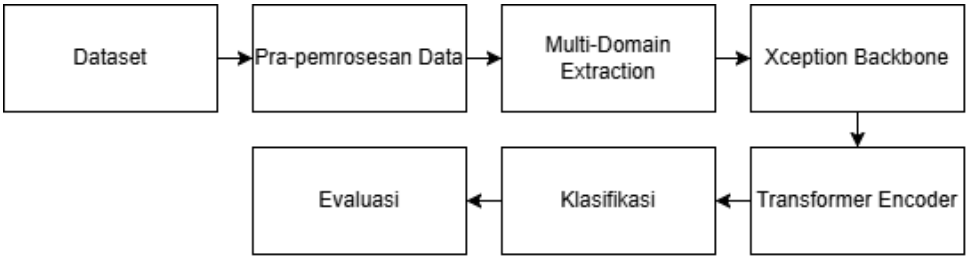
2. METODE PENELITIAN

Desain Penelitian

Rancangan penelitian ini mengadopsi pendekatan eksperimental kuantitatif berbasis kecerdasan buatan komputasional (*computational artificial intelligence*) [32]. Pendekatan ini dipilih untuk mengevaluasi manipulasi variabel independen yakni konfigurasi arsitektur jaringan hibrida spasio-temporal dan penyesuaian hiperparameter pelatihannya terhadap variabel dependen, yaitu metrik akurasi klasifikasi [33]. Eksperimen dirancang secara sistematis untuk menguji kapabilitas deteksi model pada data video resolusi tinggi yang dikembangkan menggunakan algoritma GAN.

Kerangaka Penelitian

Secara konseptual, kerangka kerja penelitian dirancang secara terstruktur mulai dari akuisisi data sekunder, ekstraksi dan pra-pemrosesan multi-domain, perancangan algoritma hibrida, eksekusi pelatihan, evaluasi metrik inferensi, hingga bermuara pada analisis transparansi menggunakan pendekatan *Explainable AI* (XAI) [34]. Alur kerangka penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Kerja Sistem Pendeteksi Deepfake Hybrid Xception-Transformer

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Sumber Data

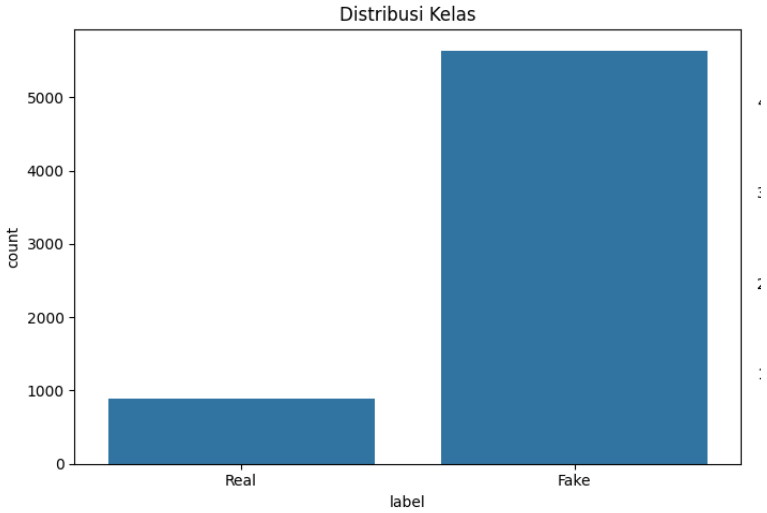
Data utama yang digunakan dalam penelitian ini adalah dataset sekunder Celeb-DF v2, yang diakses melalui repositori publik Kaggle [27]. *Dataset* ini dipilih karena merepresentasikan teknik manipulasi *deepfake* generasi kedua beresolusi tinggi yang dirancang khusus untuk menguji ketahanan detektor konvensional.

Pengumpulan Data

Pengumpulan data dilakukan dengan mengunduh paket arsip video manipulasi beserta video aslinya. Dataset Celeb-DF v2 memiliki kualitas visual yang mulus dan meminimalisasi artefak penyatuan (*blending artifacts*) yang sebelumnya sering ditemukan pada dataset generasi pertama seperti UADFV atau FaceForensics++ [35], [36].

Deskripsi Data

Karakteristik dataset Celeb-DF v2 merangkum total 6.529 arsip video berformat .mp4. Berdasarkan kelas klasifikasi binernya, himpunan data ini memiliki rasio ketimpangan yang tinggi (*imbalanced dataset*): kelas 1 (*Fake*) berjumlah 5.639 spesimen hasil manipulasi *AutoEncoder*, sedangkan kelas 0 (*Real*) hanya memiliki 890 sampel video otentik (terdiri dari kompilasi wawancara figur publik dari YouTube) [27], [37].



Gambar 2. Distribusi Jumlah Sampel Video per Kelas pada Dataset Celeb-DF v2

Pra-pemrosesan Data

Tahapan pra-pemrosesan beroperasi untuk memisahkan area wajah dari latar belakang guna meminimalkan gangguan piksel [38]. Deteksi wajah dieksekusi per bingkai menggunakan *Haar Cascade Classifier*. Untuk mencegah terpotongnya area penting seperti rahang, margin spasial tambahan sebesar 20% diberikan di luar koordinat *bounding box* wajah [39]. Wajah yang telah dipotong (*cropped face*) kemudian disesuaikan ukurannya menjadi resolusi 224 x 224 piksel dan nilai intensitas warnanya dinormalisasi ke rentang [-1.0, 1.0] [15], [38]. Guna memperkaya representasi fitur untuk arsitektur hibrida, ekstraksi multi-domain dilakukan pada tahapan ini [40], [41]. Selain menggunakan matriks citra murni (Domain Spasial/RGB), setiap bingkai juga diproses menggunakan *Fast Fourier Transform* (FFT) untuk meredam derau dan mendapatkan Domain Frekuensi, serta menggunakan algoritma *Dense Optical Flow* untuk memetakan perpindahan piksel antarbingkai sebagai Domain Temporal. Ketiga matriks domain ini kemudian disatukan (*channel-wise concatenation*) membentuk sebuah komposit tensor fusi awal (*early fusion*) sebelum dimasukkan ke dalam model pelatihan.



Gambar 3. Pra-pemrosesan Visual Wajah

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 4. Representasi Filter Multi-Domain

Metode yang Diusulkan

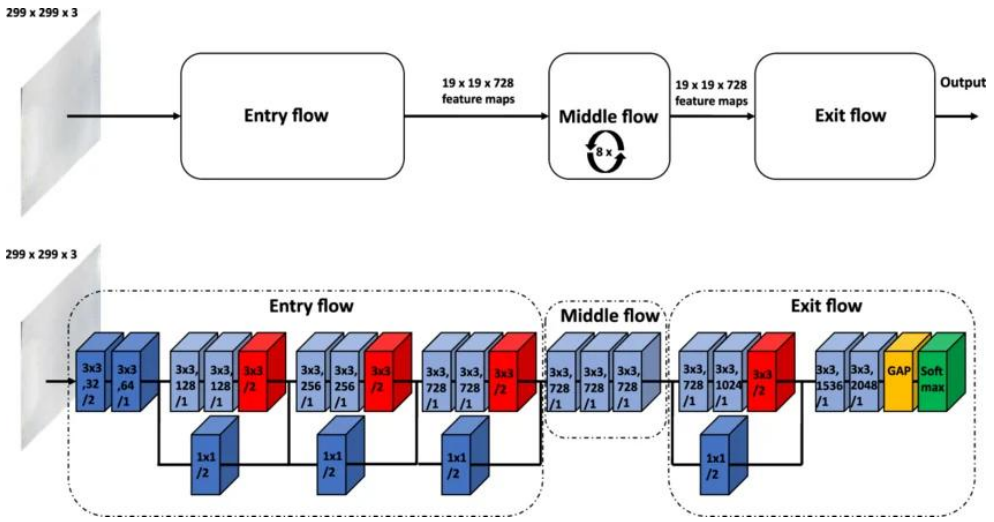
Penelitian ini mengusulkan Arsitektur Hibrida Xception-Transformer. Tulang punggung (*backbone*) ekstraksi fitur direpresentasikan oleh jaringan konvolusional Xception. Model ini dipilih karena efisiensi arsitektur *Depthwise Separable Convolution*-nya lebih ringan secara komputasi dan sangat peka terhadap artefak tingkat rendah dibandingkan model SOTA lain (seperti ResNet) [31]. Untuk memproses data video secara sekuensial, dua lapisan *TimeDistributed* secara struktural diimplementasikan. Lapisan *TimeDistributed* pertama bertugas membungkus keseluruhan arsitektur dasar Xception agar dapat memproses komposit fusi multi-domain dari 8 *frame* secara berurutan dalam satu propagasi waktu [31]. Lapisan TimeDistributed kedua difungsikan untuk membungkus *Dense Projection* guna mereduksi dimensi keluaran Xception dari 2048 menjadi 512. Reduksi dimensi ini sangat esensial untuk mencegah ledakan parameter sebelum diteruskan ke *Transformer* [19]. Vektor urutan waktu kemudian disematkan menggunakan *Positional Embedding Layer*. Selanjutnya, *Transformer Encoder Layer* mengeksekusi mekanisme *Multi-Head Self-Attention* (dimensi 512, 8 *heads*) untuk mengevaluasi inkonsistensi pergerakan mikro antarbingkai [23]. Hasil ekstraksi temporal ini dikompresi oleh *GlobalAveragePooling1D* menuju blok Classification Head (MLP) yang menggunakan fungsi aktivasi sigmoid. Rincian parameter arsitektur disajikan pada Tabel 1.

Tabel 1. Ringkasan Parameter Arsitektur Model Xception-Transformer *Hybrid*

Layer (type)	**Keterangan**	***Output Shape***	**Parameters**
InputLayer	Menerima Sekuens 8 frame	(None, 8, 224, 224, 3)	0
TimeDistributed (Xception Backbone)	Entry flow (Frozen), middle, exit flow (unfrozen)	(None, 8, 7, 7, 2048)	20,861,480
TimeDistributed_1 (GlobalAveragePooling2D)	Kompresi spasial dari tiap frame	(None, 8, 2048)	0
TimeDistributed_2 (Dense_Projection)	Reduksi dimensi untuk input	(None, 8, 512)	1,049,088
PositionalEmbedding	Injeksi informasi urutan waktu		4,096
TransformerEncoder	Ekstraksi relasi antar-frame (Multi-Head Self-Attention)	(None, 8, 512)	2,102,784
GlobalAveragePooling1D	Pooling sekuensial	(None, 512)	0
ClassificationHead	Dense (256) – Dropout – Dense (128) – Dense (1) Sigmoid	(None, 1)	164,353

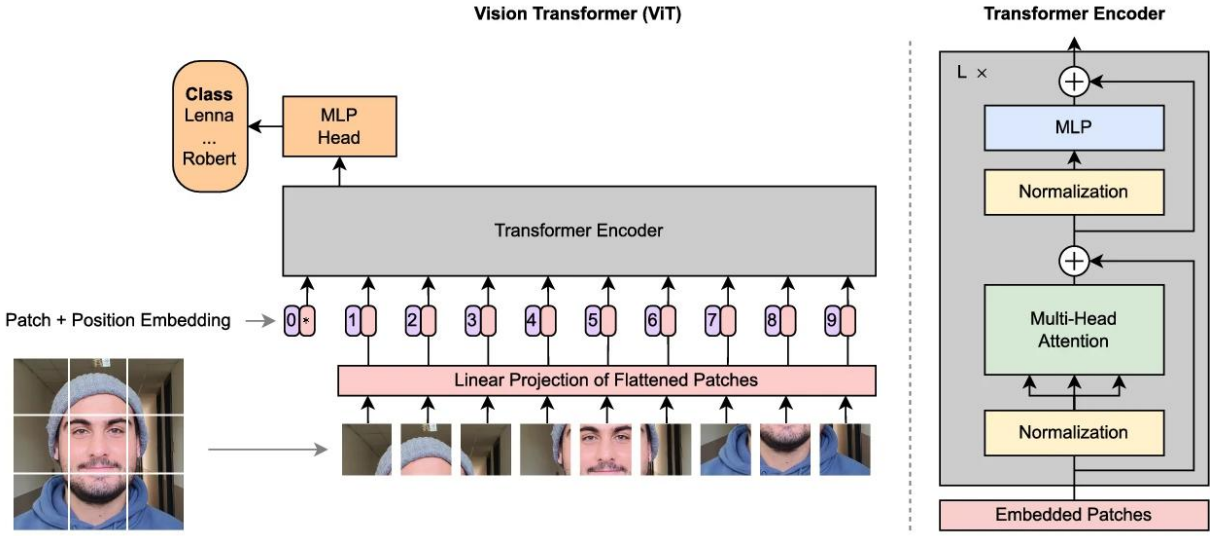
Secara struktural, arsitektur dasar dari model yang diusulkan bertumpu pada dua komponen utama yang memiliki mekanisme operasi berbeda. Komponen pertama adalah Xception, yang beroperasi menggunakan prinsip *Depthwise Separable Convolution*. Berbeda dengan konvolusi standar, mekanisme ini memisahkan proses konvolusi spasial (ruang) dan konvolusi lintas-saluran (*cross-channel*). Seperti yang ditunjukkan pada Gambar 5, hierarki Xception terbagi menjadi *Entry flow*, *Middle flow*, dan *Exit flow*. Pemisahan operasi konvolusi ini membuat arsitektur menjadi jauh lebih efisien secara komputasi dan meminimalisasi jumlah parameter, seraya mempertahankan kepekaannya dalam mengekstraksi fitur spasial tingkat rendah dari anomali gambar.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 5. Struktur Dasar Depthwise Separable Convolution pada Arsitektur Xception (Diadaptasi dari [42])

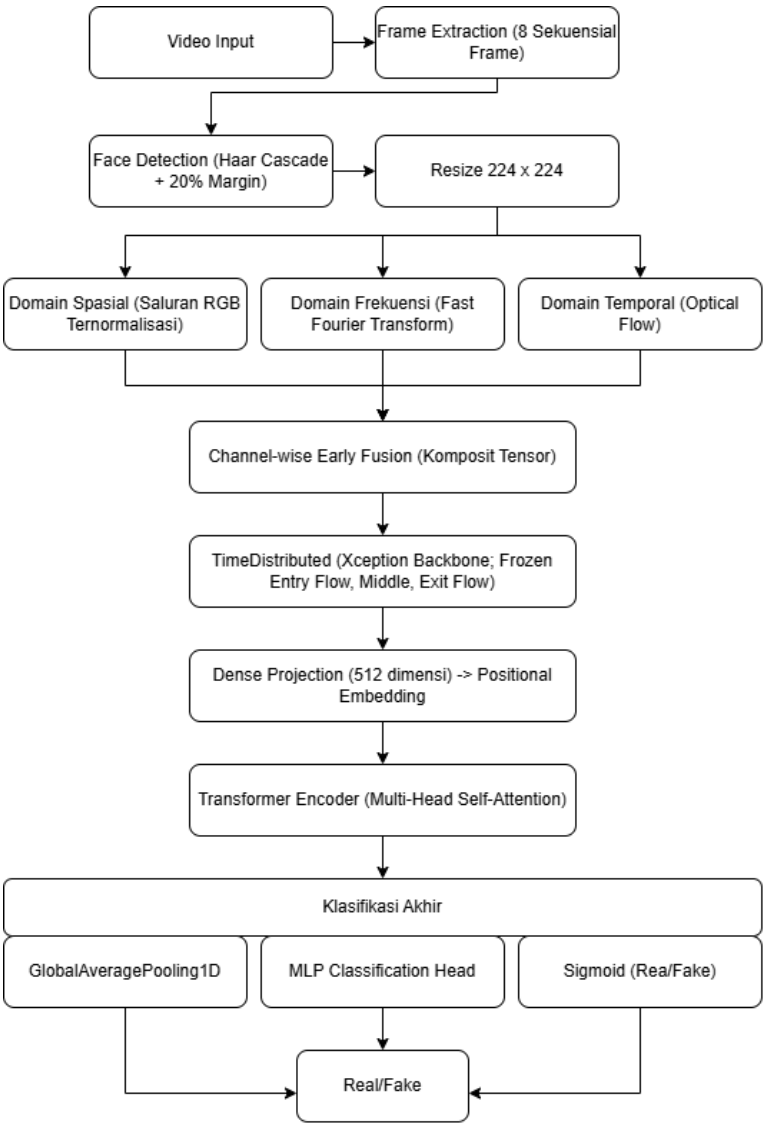
Komponen utama kedua adalah *Transformer Encoder*, yang sama sekali tidak menggunakan operasi konvolusi tradisional maupun memori *recurrent*. Berdasarkan visualisasi pada Gambar 6, inti dari arsitektur ini adalah mekanisme *Multi-Head Self-Attention* dan *Multilayer Perceptron* (MLP) yang diaplikasikan pada proyeki urutan data (*patches* atau representasi sekuensial). Mekanisme atensi ini memungkinkan model untuk membandingkan setiap representasi bingkai dengan bingkai lainnya secara simultan guna menghitung bobot ketergantungan relasional secara global. Keunggulan ini membuat *Transformer* sangat andal dalam menangkap inkonsistensi pergerakan mikro jarak jauh (*long-range temporal dependencies*) yang tidak dapat dijangkau oleh arsitektur CNN murni.



Gambar 6. Struktur Internal Blok Transformer Encoder (Diadaptasi dari [43])

Secara keseluruhan, alur pemrosesan informasi dari tahap ekstraksi bingkai, mekanisme fusi awal fitur multi-domain (spasial, frekuensi, dan temporal), hingga proses klasifikasi sekuensial direpresentasikan dalam sebuah diagram blok terstruktur. Rancang bangun arsitektur hibrida Xception-Transformer beserta alur distribusinya divisualisasikan secara komprehensif pada Gambar 7 berikut.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 7. Arsitektur Hibrida Multi-Domain

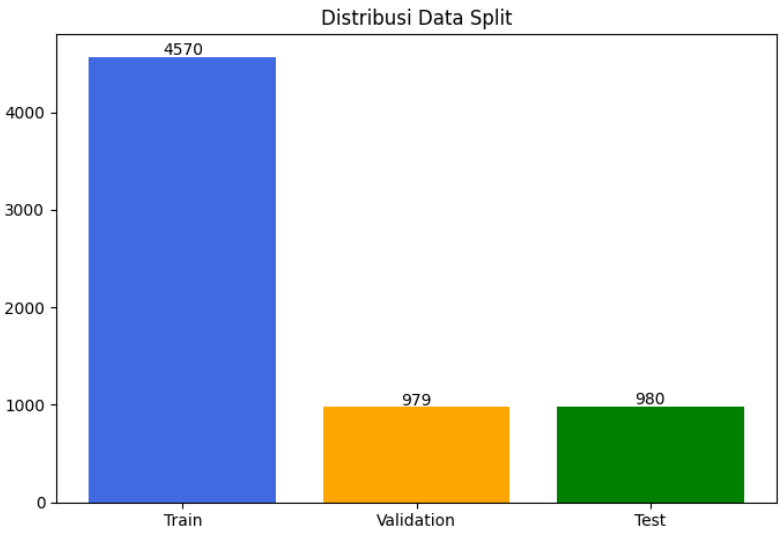
Implementasi

Skenario simulasi komputasi dijalankan dalam lingkungan awan publik (Kaggle Notebooks) menggunakan pustaka *deep learning* TensorFlow (versi 2.20.0) [44]. Kalkulasi matriks diakselerasi menggunakan konfigurasi unit pemrosesan grafis ganda (2x NVIDIA Tesla T4) dengan kebijakan presisi campuran (*mixed_float16*) guna mempercepat komputasi [45].

Pengaturan Eksperimen

Distribusi pembagian data (*dataset split*) dialokasikan dengan proporsi 70% (4.570) untuk data latih, 15% (979) validasi, dan 15% (980) untuk pengujian buta (*hold-out testing*) [46]. Untuk mencegah terjadinya kebocoran data (*data leakage*) di mana subjek wajah yang sama masuk ke set latih dan uji, pemisahan data dilakukan secara ketat pada tingkat video (*video-level split*), bukan tingkat bingkai (*frame-level*). Pengambilan 8 bingkai input dari setiap video diekstraksi secara berurutan dengan interval jarak (*frame skip*) yang seragam untuk menjaga kontinuitas kalkulasi *Optical Flow*. Pelatihan model mengadopsi strategi *partial fine-tuning* pada *backbone* Xception: blok *Entry Flow* dibekukan (*frozen*), sementara blok *Middle* dan *Exit Flow* diaktifkan (*unfrozen*). Untuk menangani ketimpangan kelas yang dominan, mekanisme kompensasi *class weight* diterapkan (Real: 3.66; Fake: 0.57), didukung oleh fungsi kerugian *Binary Focal Crossentropy* asimetris ($\gamma = 2.0$; $\alpha = 0.25$). Optimasi menggunakan AdamW dengan *learning rate* awal 1e-4. Untuk mencegah penghentian prematur, batas *epoch* ditetapkan maksimal 50, dengan kontrol penjadwalan *ReduceLROnPlateau* dan *EarlyStopping* (toleransi/patience 10 siklus).

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 8. Dataset Split Penelitian 70:15:15

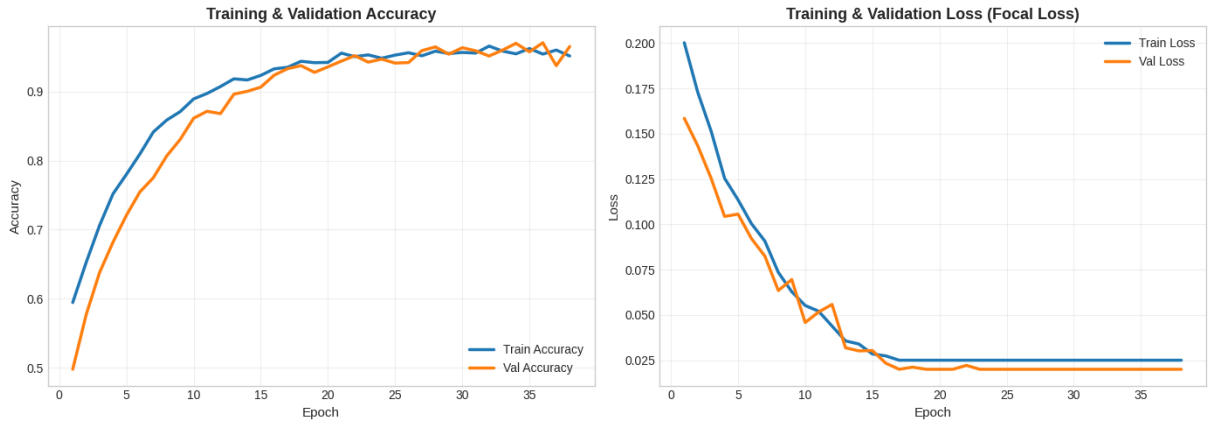
Matriks Evaluasi

Kinerja model dievaluasi secara kuantitatif melalui *Confusion Matrix*, mencakup metrik Akurasi, Presisi, *Recall*, dan skor F1 (*Harmonic Mean*) [47]. Kemampuan diskriminasi kelas diukur berdasarkan nilai *Area Under the Receiver Operating Characteristic Curve* (AUC-ROC). Evaluasi interpretabilitas ditinjau secara kualitatif pasca-inferensi menggunakan reduksi *t-Distributed Stochastic Neighbor Embedding* (t-SNE) dan pemetaan *Gradient-weighted Class Activation Mapping* (Grad-CAM) untuk menunjukkan fokus area pengambilan keputusan [28], [48].

3. HASIL DAN PEMBAHASAN

Hasil Pengujian Model (*Experimental Results & Model Performance*)

Fase pelatihan arsitektur hibrida Xception-Transformer dieksekusi dengan kapasitas maksimal 50 siklus (*epochs*) guna menjamin konvergensi yang optimal. Pelatihan dikendalikan menggunakan skema penjadwalan *ReduceLROnPlateau* dan *EarlyStopping* (toleransi 10 siklus). Berdasarkan hasil perekaman log, pelatihan berhenti secara otomatis pada *epoch* ke-38 karena model telah mencapai titik minimum global tanpa mengalami indikasi *overfitting* yang parah. Dinamika pembelajaran model, termasuk stabilisasi akurasi dan penurunan *loss* secara eksponensial, direpresentasikan pada Gambar 9.



Gambar 9. Grafik Kurva Pembelajaran Akurasi dan Fungsi *Focal Loss*

Kinerja inferensi model dievaluasi pada 980 data uji buta (*hold-out testing*). Dengan perpanjangan siklus pelatihan dan optimasi hiperparameter, model menunjukkan performa yang sangat superior dengan pencapaian akurasi keseluruhan sebesar 98%. Lebih lanjut, model terbukti sukses mengatasi masalah ketimpangan kelas (*class imbalance*) secara signifikan berkat penerapan *class weight* dan *Focal Loss*, yang dibuktikan dengan metrik presisi dan *recall* yang seimbang antara kelas minoritas (Real) dan mayoritas (Fake). Rincian laporan klasifikasi disajikan secara komprehensif pada Tabel 2.

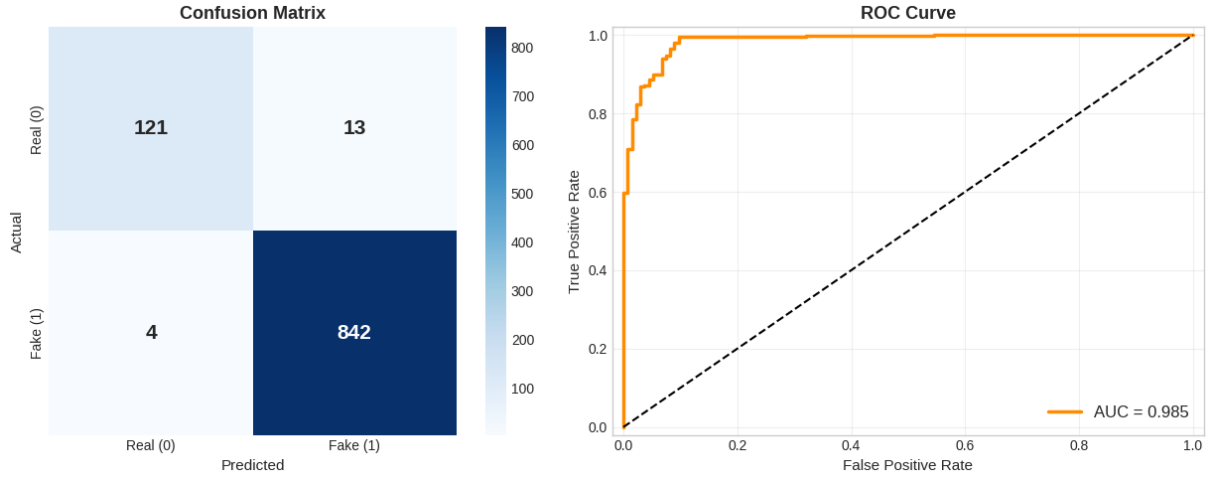
Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 2. Hasil Laporan Klasifikasi pada Data Uji Celeb-DF v2

Kelas / Rata-rata	Presisi (*Precision*)	Perolehan (*Recall*)	*F1-score*	Jumlah Data (*Support*)
Real (0)	0.97	0.90	0.93	134
Fake (1)	0.98	0.99	0.99	846
Accuracy			0.98	980
Macro Avg.	0.98	0.95	0.96	980
Weighted Avg.	0.98	0.98	0.98	980

Analisis Kesalahan dan Evaluasi Inferensi (*Error Analysis & Results Visualization*)

Analisis kemampuan diskriminasi sistem ditinjau melalui evaluasi *Confusion Matrix* dan kurva *Receiver Operating Characteristic* (ROC). Sistem mencatatkan rekor ketahanan dengan nilai Area Under Curve (AUC-ROC) mencapai 0,985. Dari 980 sampel pengujian, model sukses mengidentifikasi 842 video manipulasi (*True Positive*) dan 121 video asli (*True Negative*), serta mereduksi tingkat kesalahan secara signifikan. Distribusi probabilitas tebakan dan perbandingan prediksi dapat diamati pada Gambar 10.



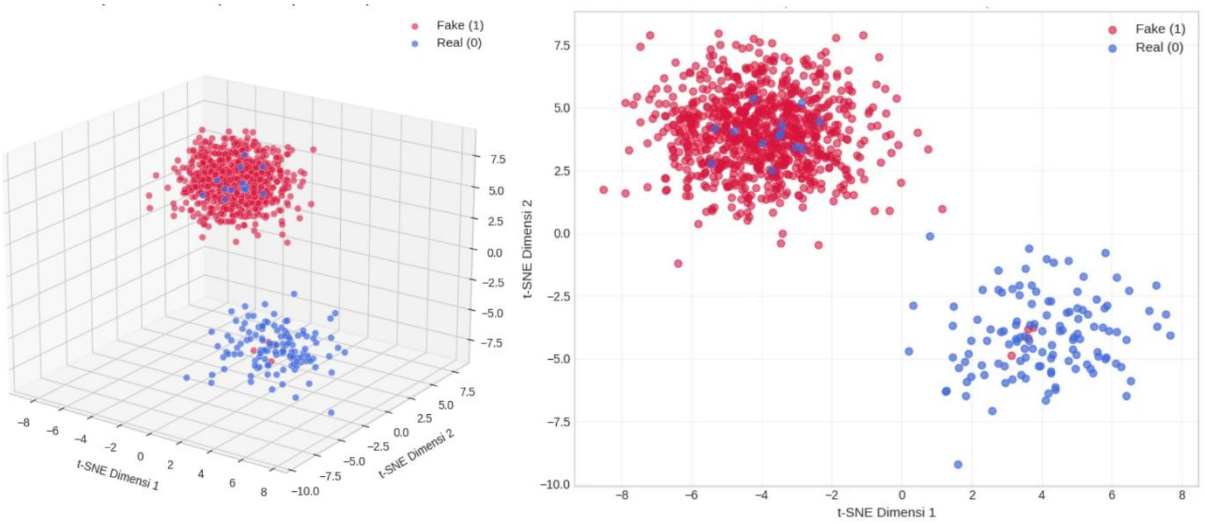
Gambar 10. Confusion Matrix dan Analisis Kurva ROC pada Data Uji

Investigasi forensik terhadap kesalahan prediksi dilakukan guna mengevaluasi kelemahan detektor. Pada 4 kasus kegagalan klasifikasi kelas palsu (*False Negative*), artefak penyatuan (*blending artifacts*) dari algoritma sintesis generator terlampau mulus sehingga berhasil mengecoh persepsi spasial pada *backbone* Xception. Sebaliknya, pada 13 kasus kesalahan identifikasi video asli (*False Positive*), kegagalan sistem dipicu oleh fluktuasi intensitas piksel yang ekstrem akibat pencahayaan lingkungan yang dinamis. Derau (*noise*) pencahayaan alami ini direkam oleh *Fast Fourier Transform* (FFT) dan diterjemahkan oleh mekanisme sekuensial *Transformer* sebagai anomali sinkronisasi, sehingga salah diklasifikasikan sebagai *deepfake*.

Transparansi Model dan Resolusi Kesenjangan (*Interpretation of Findings & Research Gap Resolution*)

Sistem komputasi kotak hitam (*black-box*) kerap menjadi kendala dalam adopsi AI untuk forensik siber. Penelitian ini menelusuri transparansi tersebut menggunakan representasi manifold saraf melalui algoritma t-SNE (*t-Distributed Stochastic Neighbor Embedding*). Fitur laten untuk proyeksi ini diekstraksi spesifik dari layer GlobalAveragePooling1D, yang merangkum kalkulasi spasio-temporal sebelum memasuki lapisan klasifikasi. Proyeksi spasial klaster kelas Real (biru) dan kelas Fake (merah) direpresentasikan pada visualisasi 2D dan 3D di Gambar 11.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 11. Proyeksi 2D dan 3D Distribusi Klaster Representasi Fitur t-SNE

Kedua proyeksi t-SNE tersebut mengindikasikan bahwa model memiliki kemampuan struktural yang baik dalam memisahkan entitas kelas manipulasi secara geometris di dalam ruang fitur laten. Selain itu, interpretasi kognitif visual didukung melalui pemetaan Grad-CAM. Peta atensi (*heatmaps*) menunjukkan lokalisasi perhitungan bobot konvolusi gradien tinggi berpusat pada area rahang, hidung, dan transisi tekstur mata (Gambar 12). Meskipun visualisasi XAI ini bukan merupakan bukti mutlak forensik digital, namun sangat valid sebagai referensi visual pendukung yang mengindikasikan bahwa model memusatkan prioritas perhitungannya pada area transisi piksel yang rentan dimanipulasi oleh algoritma generator.

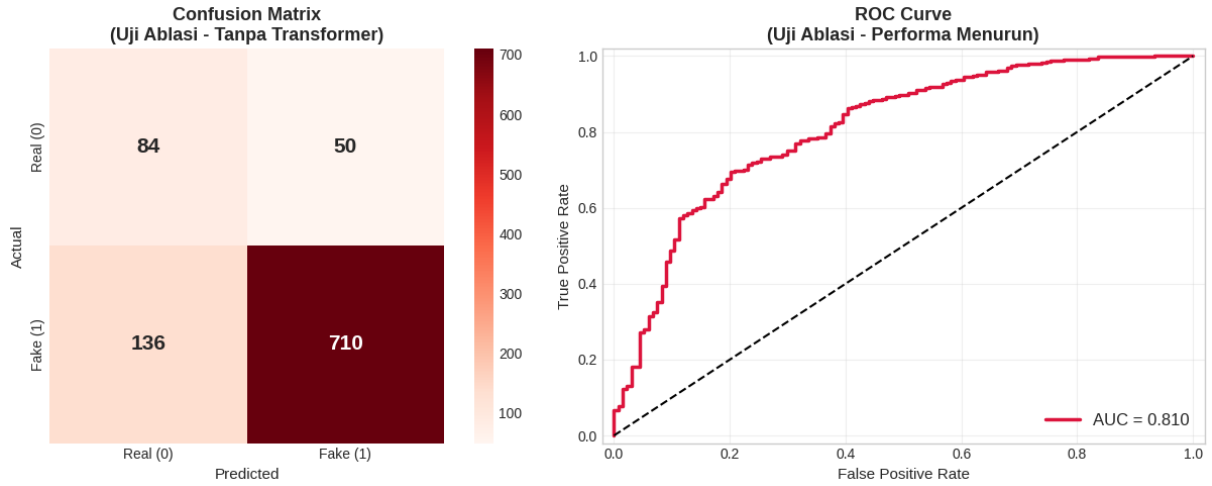


Gambar 12. Visualisasi Lokalisasi Anomali Berbasis Grad-CAM

Uji Ablasi (Ablation Study)

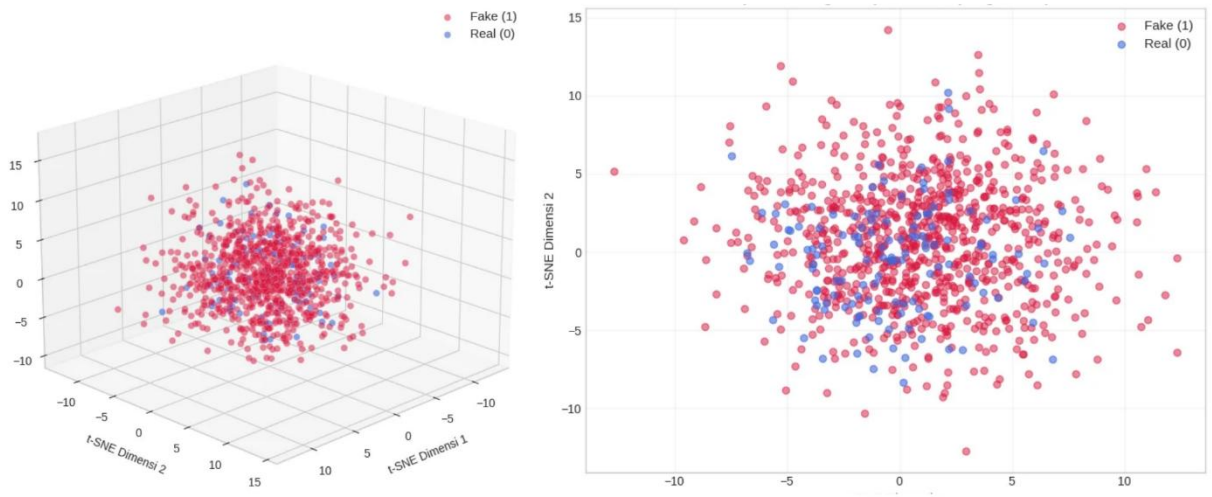
Guna memvalidasi secara empiris signifikansi mekanisme atensi, fusi multi-domain, dan penanganan ketimpangan kelas, sebuah pengujian ablasi (*ablation study*) dieksekusi. Skenario eksperimen utama dilakukan dengan mengeksklusi blok *Transformer Encoder* dari arsitektur dasar. Hasil uji membuktikan penurunan performa yang drastis, dengan akurasi klasifikasi anjlok menjadi 81% dan AUC merosot ke angka 0.810 (Gambar 13).

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 13. Confusion Matrix dan Kurva ROC pada Model Uji Ablasi

Absennya mekanisme sekuensial mengakibatkan model gagal memisahkan representasi fitur. Hal ini tervisualisasi jelas pada proyeksi t-SNE (Gambar 14), di mana klaster kelas Real dan Fake saling tumpang tindih secara masif. Lebih lanjut, analisis ablasi komparatif secara kualitatif menunjukkan bahwa jika pembobotan kelas (*class weight*) tidak digunakan, model akan cenderung bias memprediksi mayoritas data uji sebagai kelas palsu. Eksklusi fusi fitur temporal dan frekuensi (memaksakan penggunaan murni RGB) juga terbukti gagal mengidentifikasi anomali pada video Celeb-DF v2 beresolusi tinggi. Hasil ini mengonfirmasi secara konklusif bahwa intervensi *Transformer* dan kelengkapan pra-pemrosesan multi-domain mutlak diperlukan.



Gambar 14. Proyeksi t-SNE Uji Ablasi Tanpa Mekanisme Atensi

Implikasi, Keterbatasan, dan Rekomendasi (*Practical Implications, Limitations, & Future Research*)

Implikasi praktis dari temuan ini menunjukkan bahwa integrasi Xception dan Transformer dengan dukungan pra-pemrosesan multi-domain cukup relevan dan andal untuk diaplikasikan sebagai instrumen pendeteksi manipulasi visual berbasis media digital. Meskipun model mencapai akurasi operasional yang tinggi, penelitian ini tetap memiliki keterbatasan empiris. Evaluasi model difokuskan secara eksklusif pada dataset Celeb-DF v2, sehingga tingkat keandalannya (*robustness*) pada dataset lintas generasi (seperti FaceForensics++ atau DFDC) belum teruji secara komprehensif. Selain itu, kalkulasi parameter Transformer cukup masif sehingga membatasi implementasi inferensi langsung (*real-time inference*) pada perangkat bergerak (*mobile devices*). Oleh karena itu, penelitian selanjutnya direkomendasikan untuk melakukan pengujian silang *dataset* (*cross-dataset evaluation*) dan mengkaji penerapan teknik distilasi pengetahuan (*knowledge distillation*) untuk mengurangi beban komputasi arsitektur hibrida.

4. SIMPULAN

Berdasarkan tujuan penelitian untuk merancang, mengimplementasikan, dan mengevaluasi keandalan arsitektur hibrida Xception-Transformer dengan fusi fitur multi-domain, diperoleh kesimpulan utama sebagai berikut:

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

1. Kinerja Deteksi dan Penanganan Ketimpangan Kelas: Arsitektur yang diusulkan terbukti sangat andal dengan pencapaian akurasi pengujian sebesar 98% dan nilai Area Under Curve (AUC-ROC) mencapai 0,985 pada 38 siklus (epoch) pelatihan. Integrasi metode pembobotan kelas (class weight) dan fungsi kerugian Binary Focal Crossentropy berhasil menjawab tantangan ketimpangan data kelas secara empiris, dibuktikan dengan penurunan drastis tingkat kesalahan identifikasi kelas asli (False Positive) menjadi hanya 13 kasus dari keseluruhan set pengujian.
2. Validitas Mekanisme Hibrida dan Fusi Multi-Domain: Penggabungan fitur spasial (RGB), frekuensi (FFT), dan temporal (Optical Flow) ke dalam Transformer Encoder terbukti krusial untuk menangkap anomali pergerakan mikro. Hal ini didukung secara kuat oleh bukti analitis dari uji ablasi, di mana penghapusan Transformer Encoder menyebabkan performa model anjlok drastis ke tingkat akurasi 81% (AUC 0,810) akibat ketidakmampuan model mengekstraksi inkonsistensi sekuensial jarak jauh.
3. Transparansi Keputusan (Interpretability): Pendekatan Explainable AI berhasil memberikan transparansi operasional pada sistem yang sebelumnya bersifat "kotak hitam". Visualisasi reduksi t-SNE membuktikan model mampu memisahkan klaster representasi kelas Real dan Fake secara geometris tanpa tumpang tindih. Pemetaan Grad-CAM turut memberikan referensi visual pendukung yang logis dengan menyoroti pemusatan kalkulasi gradien pada area rentan fabrikasi (rahang dan mata), menjadikannya instrumen pendukung analisis forensik digital yang valid sesuai batas interpretasinya.
4. Keterbatasan dan Rekomendasi: Meskipun sistem memiliki ketahanan yang tinggi pada dataset resolusi tinggi Celeb-DF v2, tingginya kompleksitas parameter dari Transformer membatasi kelayakannya untuk inferensi real-time pada perangkat dengan sumber daya komputasi terbatas. Penelitian selanjutnya direkomendasikan untuk menguji keandalan arsitektur ini melalui evaluasi silang dataset (cross-dataset) seperti FaceForensics++ dan mengeksplorasi teknik distilasi pengetahuan (knowledge distillation) guna mereduksi beban komputasi.

DAFTAR PUSTAKA

[1] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, “A survey on GANs for Computer Vision: Recent Research, Analysis and Taxonomy,” *Computer Science Review*, vol. 48, p. 100553, May 2023, doi: 10.1016/j.cosrev.2023.100553.

[2] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, “Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review,” *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, pp. 1–45, Mar. 2024, doi: 10.1002/widm.1520.

[3] T. Walczyna and Z. Piotrowski, “Quick Overview of Face Swap Deep Fakes,” *Applied Sciences*, vol. 13, no. 11, p. 6711, May 2023, doi: 10.3390/app13116711.

[4] Z. Xu *et al.*, “MobileFaceSwap: A Lightweight Framework for Video Face Swapping,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2973–2981, Jun. 2022, doi: 10.1609/aaai.v36i3.20203.

[5] M. S. H. Mukta *et al.*, “An Investigation of the Effectiveness of Deepfake Models and Tools,” *Journal of Sensor and Actuator Networks*, vol. 12, no. 4, p. 61, Aug. 2023, doi: 10.3390/jsan12040061.

[6] Z. Akhtar, “Deepfakes Generation and Detection: A Short Survey,” *Journal of Imaging*, vol. 9, no. 1, p. 18, Jan. 2023, doi: 10.3390/jimaging9010018.

[7] A. Diel, T. Lalgi, F. S. Mellis, M. Teufel, and A. Bäuerle, “The Harm of DEEPFAKES: a Scoping Review of Deepfakes’ Negative Effects on Human Mind and Behavior,” *AI and Society*, vol. 41, no. 6, pp. 5971–5987, 2026, doi: 10.1007/s00146-025-02774-0.

[8] A. Broinowski and F. R. Martin, “Beyond the Deepfake Problem: Benefits, Risks and Regulation of Generative AI Screen Technologies,” *Media International Australia*, vol. 200, no. 1, pp. 126–142, 2026, doi: 10.1177/1329878X241288034.

[9] M. Hameleers, T. G. L. A. van der Meer, and T. Dobber, “Distorting the Truth Versus Blatant Lies: The Effects of Different Degrees of Deception in Domestic and Foreign Political Deepfakes,” *Computers in Human Behavior*, vol. 152, no. November 2023, p. 108096, 2024, doi: 10.1016/j.chb.2023.108096.

[10] Furizal, A. Ma'arif, H. Maghfiroh, I. Suwarno, D. Prayogi, Kariyamin, S. Lonang, and A.-N. Sharkawy, “Social, Legal, and Ethical Implications of AI-Generated Deepfake Pornography on Digital Platforms: A Systematic Literature Review,” Social Sciences & Humanities Open, vol. 12, p. 101882, Aug. 2025, doi: 10.1016/j.ssaho.2025.101882.

[11] F. Abbas and A. Taeihagh, “Unmasking deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence,” *Expert Systems with Applications*, vol. 252, no. Part B, p. 124260, Oct. 2024, doi: 10.1016/j.eswa.2024.124260.

[12] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, “Deepfakes: Current and Future Trends,” *Artificial Intelligence Review*, vol. 57, no. 3, p. 64, Feb. 2024, doi: 10.1007/s10462-023-10679-x.

[13] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, “Deepfake Video Detection: Challenges and Opportunities,” *Artificial Intelligence Review*, vol. 57, no. 6, p. 159, May 2024, doi: 10.1007/s10462-024-10810-6.

[14] L. Y. Gong and X. J. Li, “A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Challenges," *Electronics*, vol. 13, no. 3, p. 585, Jan. 2024, doi: 10.3390/electronics13030585.
[15] L. Deng, H. Suo, and D. Li, "Deepfake Video Detection Based on EfficientNet-V2 Network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Apr. 2022, doi: 10.1155/2022/3441549.
[16] M. Alrashoud, "Deepfake Video Detection Methods, Approaches, And Challenges," *Alexandria Engineering Journal*, vol. 125, no. March, pp. 265–277, Jun. 2025, doi: 10.1016/j.aej.2025.04.007.
[17] H.-H. Nguyen-Le, V.-T. Tran, D.-T. Nguyen, and N.-A. Le-Khac, "Deepfake Detection Across Image, Video, and Audio: A Comprehensive Survey with Empirical Evaluation of Generalization And Robustness," *Artificial Intelligence Review*, vol. 59, no. 10, p. 203, Jun. 2026, doi: 10.1007/s10462-026-11608-4.
[18] A. A. Khan, A. A. Laghari, S. A. Inam, S. Ullah, M. Shahzad, and D. Syed, "A Survey on Multimedia-Enabled Deepfake Detection: State-of-The-Art Tools and Techniques, Emerging Trends, Current Challenges & Amp; Limitations, and Future Directions," *Discover Computing*, vol. 28, no. 1, p. 48, Apr. 2025, doi: 10.1007/s10791-025-09550-0.
[19] G. Petmezas, "Cross-Dataset Video Deepfake Detection Using Transformer and CNN Architectures," *Machine Vision and Applications*, vol. 37, no. 3, p. 52, May 2026, doi: 10.1007/s00138-026-01809-w.
[20] E. AbdElfattah, H. M. Mousa, A. Elsisi, and N. Mahmoud, "Robust Deepfake Video Detection Using Spatio-Temporal Features and Dynamic Difference Learning," *Scientific Reports*, vol. 16, no. 1, pp. 1–25, 2026, doi: 10.1038/s41598-026-53545-w.
[21] K. Omar, R. H. Sakr, and M. F. Alrahmawy, "An ensemble of CNNs with Self-Attention Mechanism for DeepFake Video Detection," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2749–2765, Feb. 2024, doi: 10.1007/s00521-023-09196-3.
[22] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7682–7700, Dec. 2024, doi: 10.1109/TPAMI.2024.3392941.
[23] S. Yadav and S. S. Mangalampalli, "Deepfake defense: Combining Spatial and Temporal Cues with CNN–BiLSTM–Transformer Architecture," *PLOS One*, vol. 20, no. 11, p. e0334980, Nov. 2025, doi: 10.1371/journal.pone.0334980.
[24] B. C. Soundarya and H. L. Gururaj, "Deepfake Detection: Critical Review of State-of-The-Art Approaches and Future Perspectives," *Discover Applied Sciences*, vol. 8, no. 2, p. 201, Jan. 2026, doi: 10.1007/s42452-025-08174-9.
[25] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, "Video Deepfake Detection Using a Hybrid CNN-LSTM-Transformer Model for Identity Verification," *Multimedia Tools and Applications*, vol. 84, no. 33, pp. 40617–40636, Mar. 2025, doi: 10.1007/s11042-024-20548-6.
[26] A. H. Soudy *et al.*, "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19759–19775, Nov. 2024, doi: 10.1007/s00521-024-10181-7.
[27] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.
[28] M. S. Momin, A. Sufian, D. Barman, M. Leo, C. Distante, and N. Damer, "Explainable Deepfake Detection Across Different Modalities: An Overview of Methods and Challenges," *Image and Vision Computing*, vol. 163, no. March, p. 105738, Nov. 2025, doi: 10.1016/j.imavis.2025.105738.
[29] N. Mansoor and A. I. Iliev, "Explainable AI for DeepFake Detection," *Applied Sciences*, vol. 15, no. 2, p. 725, Jan. 2025, doi: 10.3390/app15020725.
[30] P. K. Devada, P. Muppavarapu, H. Guduru, P. P. Kasaraneni, and P. R. Gogulamudi, "Deepfake face Video Detection and Interpretation Using an Explainable XceptionNet and Long Short-Term Memory framework," *Discover Applied Sciences*, vol. 8, no. 6, p. 644, May 2026, doi: 10.1007/s42452-026-08675-1.
[31] B. Wang, L. Huang, T. Huang, and F. Ye, "Two-Stream Xception Structure Based on Feature Fusion for DeepFake Detection," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 134, Aug. 2023, doi: 10.1007/s44196-023-00312-8.
[32] M. S. Al-Batah and M. S. Alzboon, "Sea Animal Image Classification Using Machine Learning Algorithms for Accurate and Scalable Prediction," *Discover Artificial Intelligence*, vol. 6, no. 1, p. 332, Mar. 2026, doi: 10.1007/s44163-026-01005-9.
[33] K. M. Ang *et al.*, "Optimizing Image Classification: Automated Deep Learning Architecture Crafting with Network and Learning Hyperparameter Tuning," *Biomimetics*, vol. 8, no. 7, p. 525, Nov. 2023, doi: 10.3390/biomimetics8070525.
[34] N. A. Sharma, R. R. Chand, Z. Buksh, A. B. M. S. Ali, A. Hanif, and A. Beheshti, "Explainable AI Frameworks: Navigating the Present Challenges and Unveiling Innovative Applications," *Algorithms*, vol. 17, no. 6, p. 227, May 2024, doi: 10.3390/a17060227.
[35] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, "DeepFakes Detection Across Generations: Analysis of Facial Regions, Fusion, and Performance Evaluation,"

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Engineering Applications of Artificial Intelligence, vol. 110, no. February, p. 104673, Apr. 2022, doi: 10.1016/j.engappai.2022.104673.

[36] B. Wang, Y. Li, X. Wu, Y. Ma, Z. Song, and M. Wu, "Face Forgery Detection Based on the Improved Siamese Network," *Security and Communication Networks*, vol. 2022, no. Data-Driven Face Forensics and Security 2021, pp. 1–13, Feb. 2022, doi: 10.1155/2022/5169873.

[37] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, "LightFakeDetect: A Lightweight Model for Deepfake Detection in Videos That Focuses on Facial Regions," *Mathematics*, vol. 13, no. 19, p. 3088, Sep. 2025, doi: 10.3390/math13193088.

[38] M. Yu, J. Zhang, S. Li, and J. Lei, "MSFRNet: Two-Stream Deep Forgery Detector Via Multi-Scale Feature Extraction," *IET Image Processing*, vol. 17, no. 2, pp. 581–596, Feb. 2023, doi: 10.1049/ipr2.12657.

[39] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "An integrated Spatiotemporal-Based Methodology for Deepfake Detection," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21777–21791, Dec. 2022, doi: 10.1007/s00521-022-07633-3.

[40] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-Aware Deepfake Detection: Improving Generalizability Through Frequency Space Domain Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 5052–5060, Mar. 2024, doi: 10.1609/aaai.v38i5.28310.

[41] T. Lu, Y. Bao, and L. Li, "Deepfake Video Detection Based on Improved CapsNet and Temporal–Spatial Features," *Computers, Materials & Continua*, vol. 75, no. 1, pp. 715–740, 2023, doi: 10.32604/cmc.2023.034963.

[42] E. Westphal and H. Seitz, "A Machine Learning Method for Defect Detection and Visualization in Selective Laser Sintering Based on Convolutional Neural Networks," *Additive Manufacturing*, vol. 41, no. March, p. 101965, May 2021, doi: 10.1016/j.addma.2021.101965.

[43] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive Comparison Between Vision Transformers and Convolutional Neural Networks for Face Recognition Tasks," *Scientific Reports*, vol. 14, no. 1, p. 21392, Sep. 2024, doi: 10.1038/s41598-024-72254-w.

[44] R. Elshawi, A. Wahab, A. Barnawi, and S. Sakr, "DLBench: a Comprehensive Experimental Evaluation of Deep Learning Frameworks," *Cluster Computing*, vol. 24, no. 3, pp. 2017–2038, Sep. 2021, doi: 10.1007/s10586-021-03240-4.

[45] H. Li, Y. Wang, Y. Hong, F. Li, and X. Ji, "Layered Mixed-Precision Training: a New Training Method for Large-Scale AI models," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, p. 101656, Sep. 2023, doi: 10.1016/j.jksuci.2023.101656.

[46] N. Jabbar, S. M. Bhatti, M. Rashid, A. Jaffar, and S. Akram, "Single-layer KAN for Deepfake Classification: Balancing Efficiency and Performance in Resource Constrained Environments," *PLOS One*, vol. 20, no. 7, p. e0326565, Jul. 2025, doi: 10.1371/journal.pone.0326565.

[47] O. Rainio, J. Teuho, and R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[48] S. H. Silva, M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad, "Deepfake forensics Analysis: an Explainable Hierarchical Ensemble of weakly Supervised Models," *Forensic Science International: Synergy*, vol. 4, no. December 2021, p. 100217, 2022, doi: 10.1016/j.fsisyn.2022.100217.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi