BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan *Generative Adversarial Networks* (GAN) telah memungkinkan terciptanya citra dan video sintetis yang sulit dibedakan dari rekaman asli, dan teknologi inilah yang melahirkan istilah *deepfake*, yaitu konten visual atau audio yang dimanipulasi sedemikian rupa sehingga menampilkan adegan atau ucapan yang tidak pernah terjadi. Meskipun awalnya dikembangkan untuk hiburan dan riset sintesis media, *deepfake* kini menjadi ancaman serius pada beberapa tingkatan. Pada tingkat geopolitik, teknologi ini dipakai sebagai instrumen disinformasi dan propaganda [1]. Pada tingkat individual, video *deepfake* yang beredar global berbentuk pornografi non-konsensual yang sebagian besar menargetkan perempuan, menimbulkan kerugian reputasi, psikologis, dan hukum yang nyata [2]. Skala dan kecepatan distribusi konten ini menjadikan deteksi otomatis *deepfake* bukan sekadar tantangan riset, melainkan kebutuhan forensik digital yang mendesak.

Penelitian ini merupakan sebuah studi komparatif yang membandingkan tiga pendekatan deteksi *deepfake*, yaitu pendekatan berbasis domain spasial, pendekatan berbasis domain frekuensi, dan pendekatan *hybrid* yang menggabungkan keduanya. Tujuannya adalah mengukur secara langsung dan terkontrol seberapa besar kontribusi analisis domain frekuensi dalam meningkatkan kemampuan deteksi, terutama ketahanan model ketika diuji lintas *dataset*. Uraian selanjutnya memaparkan perkembangan pendekatan deteksi *deepfake* beserta celah penelitian yang mendasari pemilihan rancangan komparatif tersebut.

Berbagai metode telah dikembangkan untuk mendeteksi *deepfake*, salah satunya berbasis domain spasial yang menganalisis informasi visual langsung dari citra atau *frame* video. Beberapa arsitektur yang terbukti efektif di domain ini antara lain *Mesoscopic Network* (MesoNet), *Residual Network* (ResNet), dan *Extreme Inception Network* (XceptionNet) [3, 4]. MesoNet merupakan *Convolutional Neural Network* (CNN) ringan yang mendeteksi manipulasi pada tingkat *mesoscopic*, yaitu diantara analisis *microscopic* berbasis *noise* (yang rusak oleh kompresi video) dan tingkat semantik (yang sulit dibedakan mata manusia). Model ini memanfaatkan kenyataan bahwa wajah hasil *deepfake* cenderung tampak buram dan kehilangan detail halus akibat keterbatasan ruang penyandian (*encoding space*) *autoencoder* [4]. ResNet-50 memperkenalkan *residual learning* untuk merepresentasikan fitur spasial yang lebih dalam [5], namun performanya pada deteksi

deepfake dilaporkan lebih rendah dibanding XceptionNet dan menurun pada pengujian lintas *dataset* [3]. XceptionNet menjadi *baseline* domain spasial yang paling banyak dipakai karena *Depthwise Separable Convolution* (DSC) memberikan efisiensi komputasi tanpa mengorbankan akurasi, dan dilaporkan mencapai akurasi deteksi 99,26% pada *FaceForensics*++ (FFPP) tanpa kompresi [6, 7]. Namun, akurasi tinggi ini tidak konsisten pada pengujian lintas *dataset*, yang mengindikasikan ketergantungan model terhadap artefak visual khas *dataset* pelatihan.

Penyebab utama dari kegagalan generalisasi model *deepfake* biasanya terletak pada sifat fundamental artefak, yang bersifat algoritmik, bukan visual [8]. Artefak biasanya muncul dari kelemahan dalam operasi *upsampling*, yang digunakan oleh GAN untuk meningkatkan resolusi citra, yang cenderung menciptakan pola *noise* atau *spectral distortion* yang tidak alami, dan biasanya dikenal sebagai GAN *fingerprints*, yang tidak terlihat jelas dalam domain spasial [9].

Untuk mengatasi keterbatasan tersebut, riset deteksi *deepfake* bergeser ke domain frekuensi. Pendekatan berbasis *Fast Fourier Transform* (FFT) memanfaatkan asimetri spektral pada citra GAN dan terbukti dapat membedakan citra sintetis dari citra natural meski manipulasi visualnya tidak terlihat di domain spasial [8]. Pendekatan berbasis *Discrete Cosine Transform* (DCT) bekerja pada koefisien blok dan menyingkapkan anomali pada *band* frekuensi menengah yang konsisten antar-arsitektur GAN [10]. Sementara itu, pembelajaran adaptif pada representasi frekuensi seperti *Frequency-aware Clues* [11] dan *Frequency-aware Deepfake Detection* berbasis spektrum [12] meningkatkan kemampuan generalisasi terhadap teknik sintesis baru dan citra terkompresi. Konvergensi ketiga pendekatan ini menunjukkan bahwa jejak generatif paling stabil bukan pada piksel, melainkan pada distribusi spektralnya.

Sebagian besar generator *deepfake* gagal mereplikasi distribusi spektral citra natural pada frekuensi tinggi. Pengukuran *azimuthal frequency profile* memperlihatkan kelebihan energi konsisten pada *frame* sintetis di empat arsitektur GAN yang diuji [8], dan *Frequency Domain Analysis* (FDA) dapat mendeteksi perbedaan tersebut bahkan ketika manipulasi visual tampak realistis di domain spasial [8]. Studi terkini mulai mengombinasikan deteksi spasial dengan pendekatan frekuensi. *Alam et al.* menunjukkan bahwa penggabungan domain spasial dan domain frekuensi dapat meningkatkan kemampuan generalisasi dan ketahanan terhadap kompresi [13]. Temuan-temuan ini memperkuat pentingnya integrasi dua domain analisis untuk menciptakan detektor *deepfake* yang lebih *robust*.

Pada penerapan nyata, detektor *deepfake* menghadapi distribusi data yang berbeda dari data pelatihan, seperti generator baru yang bermunculan setiap tahun, tingkat kompresi platform sosial yang bervariasi, dan teknik pasca-pemrosesan *adversarial* yang dipakai untuk menyamarkan jejak. Studi sistematis melaporkan penurunan *Area Under the Curve* (AUC) sebesar 10 hingga 20 poin ketika detektor berbasis CNN spasial dipindahkan dari satu *dataset* ke *dataset* lain [14, 15]. Karena itu, kemampuan generalisasi lintas *dataset*, bukan akurasi *in-dataset*, menjadi metrik yang paling relevan untuk menilai kelayakan deteksi *deepfake* pada kondisi nyata, sekaligus menjadi fokus utama penelitian ini.

Studi awal hibridisasi domain spasial dan frekuensi telah menunjukkan potensi peningkatan generalisasi. *Spectral Cross-Attentional Network* (SpecXNet) menggabungkan cabang spasial XceptionNet dengan cabang frekuensi untuk meningkatkan ketahanan deteksi terhadap berbagai manipulasi [13]. *Frequency Enhanced Self-Blended Images* (FSBI) memanfaatkan *self-blended image* yang ditingkatkan frekuensi untuk memperkuat generalisasi lintas *dataset* [16]. *Frequency-Domain Masking* mengintegrasikan informasi spasial dan frekuensi melalui mekanisme adaptif [17]. Studi-studi ini melaporkan bahwa fusi dua domain bermanfaat, tetapi sebagian besar masih dioptimalkan untuk performa *in-dataset* pada FFPP, dan belum mengevaluasi secara sistematis kontribusi fusi *late* dan *gating* terhadap *robustness* lintas *dataset*.

Penelitian ini mengisi celah tersebut melalui sebuah studi komparatif yang mengukur secara terkontrol seberapa besar kontribusi domain frekuensi terhadap kemampuan generalisasi lintas *dataset*. Alih-alih hanya membangun satu detektor, penelitian ini merancang, melatih, dan mengevaluasi tiga konfigurasi model pada kondisi eksperimen yang identik, yaitu model domain spasial murni berbasis XceptionNet sebagai *baseline*, model domain frekuensi murni berbasis FreqCNN yang bekerja pada peta *log-magnitude* FFT, dan model *hybrid* yang menggabungkan kedua cabang melalui *late fusion* dengan *Squeeze-and-Excitation* (SE) *gating*. Dengan menyetarakan data, protokol pelatihan, dan metrik pada ketiga model, perbedaan kinerja yang teramati dapat diatribusikan langsung pada domain fitur yang digunakan, sehingga kontribusi domain frekuensi dapat diisolasi dan diukur, bukan sekadar diasumsikan.

Ketiga model dilatih dan diuji pada dua *dataset benchmark*, yaitu FaceForensics++ (FFPP) dan Celeb-DF (CDF), pada beberapa ukuran sampel untuk mengamati pengaruh volume data terhadap kinerja. Evaluasi dilakukan dalam dua skenario, yaitu *in-dataset* ketika model dilatih dan diuji pada *dataset* yang sama, serta *cross-dataset* ketika model dilatih pada

FFPP lalu diuji pada CDF dan sebaliknya CDF ke FFPP. AUC digunakan sebagai metrik utama karena paling relevan untuk menilai kelayakan deteksi di dunia nyata. Melalui perbandingan langsung ketiga konfigurasi tersebut, penelitian ini menilai apakah penambahan fitur frekuensi mampu menahan penurunan kinerja lintas *dataset* yang lazim dialami detektor domain spasial murni, sekaligus menetapkan apakah dan seberapa besar pendekatan *hybrid* memberikan keunggulan dibandingkan model domain tunggal.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang menggarisbawahi kegagalan detektor spasial murni dan potensi yang belum dimanfaatkan dari FDA, permasalahan utama dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Sejauh mana detektor *deepfake* berbasis domain spasial murni (XceptionNet) mengalami penurunan performa ketika diuji pada **video sintetis** dari *dataset* yang berbeda dengan data pelatihannya?
2. Sejauh mana penambahan analisis domain frekuensi (FFT) ke dalam arsitektur XceptionNet dapat memperkecil penurunan tersebut?
3. Seberapa besar kontribusi masing-masing komponen (fitur spasial vs fitur frekuensi) terhadap akurasi, presisi, *recall*, dan AUC pada pengujian *in-dataset* maupun *cross-dataset*?

1.3 Tujuan

Tujuan yang hendak dicapai pada penelitian ini adalah:

1. Mengimplementasikan arsitektur *hybrid* XceptionNet–FFT dengan *late fusion* dan SE *gating* untuk deteksi *deepfake* tingkat *frame* video.
2. Melakukan *ablation study* (spasial-saja vs frekuensi-saja vs *hybrid*) untuk memisahkan kontribusi masing-masing domain.
3. Mengevaluasi generalisasi model pada skenario *in-dataset* (FFPP ke FFPP, CDF ke CDF) dan *cross-dataset* (FFPP ke CDF, CDF ke FFPP) menggunakan akurasi, presisi, *recall*, dan AUC.

1.4 Manfaat

Penelitian ini diharapkan memberikan kontribusi terhadap komunitas riset, serta diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Akademik: Memberikan kontribusi ilmiah berupa evaluasi sistematis terhadap kontribusi fitur frekuensi dalam arsitektur *hybrid* deteksi *deepfake*.
2. Manfaat Teknologi: Menyediakan rancangan arsitektur dan pipeline reproduktif yang dapat dijadikan acuan oleh peneliti lain yang mengembangkan detektor *hybrid* pada domain spasial–frekuensi.
3. Manfaat Sosial: Mendukung upaya pengurangan penyalahgunaan *deepfake* (misinformasi, *deepfake porn*, propaganda politik) dengan menyediakan dasar metodologis yang dapat dievaluasi oleh lembaga forensik digital.
4. Manfaat Praktis: Memberikan *baseline cross-dataset* yang dapat dipakai pengembang sistem verifikasi konten sebagai pembanding untuk arsitektur deteksi lain.

1.5 Ruang Lingkup

Ruang lingkup penelitian ini difokuskan pada pengembangan dan evaluasi model deteksi *deepfake* berbasis *hybrid* yang menggabungkan analisis domain spasial dan domain frekuensi. Penelitian ini hanya mencakup tahap perancangan, implementasi, dan evaluasi model deteksi menggunakan *dataset* publik, tanpa membahas aspek produksi video *deepfake* atau sistem penyebarannya.

Penelitian ini difokuskan pada:

1. Cakupan Data: Penelitian menggunakan **video** dari dua *dataset* publik yang kemudian diekstraksi menjadi ***frame* citra** untuk diproses model:
   a. **FaceForensics++** [7], repositori *https://github.com/ondyari/FaceForensics*, kompresi *high-quality* (c23), dengan empat metode manipulasi (*Deepfakes*, *Face2Face*, *FaceSwap*, *NeuralTextures*).
   b. **Celeb-DF v2** [18], repositori *https://github.com/yuezunli/celeb-deepfakeforensics*, kategori *celebrity face-swap*.

   Pengambilan sampel dilakukan dengan rasio 50:50 antara kelas asli dan kelas manipulasi yang diterapkan pada level video, sementara pemisahan *train, val, test* juga dilakukan pada level video untuk menghindari kebocoran *frame*. Penelitian tidak mencakup deteksi manipulasi audio atau *text-based deepfake*.
2. Cakupan Metode: Analisis frekuensi dilakukan menggunakan metode FFT untuk mengekstraksi fitur spektral yang merepresentasikan artefak sintesis. Model spasial yang digunakan terbatas pada arsitektur XceptionNet.

3. Cakupan Evaluasi: Evaluasi kinerja model dilakukan berdasarkan metrik *accuracy*, *precision*, *recall*, dan AUC, serta pengujian *cross-dataset generalization* untuk mengukur ketahanan model terhadap variasi sumber data.
4. Cakupan Sistem: Implementasi dilakukan dalam lingkungan eksperimental, dengan keterbatasan komputasi dan waktu pelatihan tertentu. Penelitian tidak terfokus pada optimasi waktu inferensi secara mendalam.

Tabel 1.1 Komposisi total *dataset* yang digunakan

Dataset	**Real videos**	**Fake videos**	**Total frame (~)**	**Sumber**
FaceForensics++ (n=750, c23)	375	375	~61.000	[7]
Celeb-DF v2 (n=750)	375	375	~61.000	[18]

Dengan batasan ini, penelitian diarahkan untuk menjaga fokus pada evaluasi kontribusi integrasi FDA ke dalam arsitektur XceptionNet terhadap kemampuan generalisasi, ketahanan, dan akurasi dalam mendeteksi *deepfake*.