BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan dalam *deep generative models* seperti *Variational Autoencoders* (VAEs), *Generative Adversarial Networks* (GANs), dan model difusi telah memungkinkan penciptaan gambar sintetis yang sangat realistis [1, 2]. Salah satu pemanfaatan teknologi ini yang paling menonjol adalah *deepfake*, sebuah teknik *deep learning* untuk mensintesis dan memanipulasi citra digital dengan cara menggabungkan, mengganti, atau menumpangkan wajah subjek pada tingkat gambar [3, 4]. Teknik ini mampu menghasilkan visual yang sangat fotorealistik dan meniru detail halus manusia, sehingga sulit dibedakan oleh persepsi visual manusia rata-rata [5, 6]. Akibatnya, manipulasi wajah berbasis *deepfake* kini menjadi ancaman serius terhadap integritas informasi digital dan privasi individu. Pihak dengan niat buruk mengeksploitasi teknik tersebut untuk memalsukan konten yang menargetkan figur publik maupun masyarakat umum [5, 7]. Ancaman ini telah terwujud secara nyata, sebagaimana terjadi pada perusahaan rekayasa Arup yang mengalami kerugian sekitar US$25,6 juta ketika seorang karyawannya ditipu melalui panggilan video yang seluruh pesertanya merupakan *deepfake* wajah eksekutif perusahaan [8]. Skala ancaman ini diperkuat oleh laporan iProov yang menunjukkan bahwa pada awal tahun 2025 *deepfake* diperkirakan menyumbang sekitar 40% dari seluruh kasus *biometric fraud* [9], serta proyeksi Deloitte yang memperkirakan kerugian *fraud* berbasis *generative AI* di Amerika Serikat meningkat dari US$12,3 miliar pada 2023 menjadi US$40 miliar pada 2027 [10]. Dengan demikian, deteksi manipulasi wajah secara andal menjadi kebutuhan yang mendesak untuk menjaga kepercayaan publik serta mencegah konsekuensi sosial dan hukum yang signifikan [1, 11].

Seiring meningkatnya kualitas sintesis, jejak manipulasi pada *deepfake* menjadi semakin halus sehingga sulit dikenali hanya melalui analisis pada domain spasial. Pendekatan yang semata-mata mengandalkan fitur spasial cenderung menangkap artefak visual yang spesifik terhadap teknik manipulasi tertentu, sehingga rentan kehilangan keandalannya ketika dihadapkan pada metode pemalsuan yang baru [12]. Di sisi lain, proses pembangkitan citra sintetis, khususnya operasi *upsampling* pada arsitektur generatif, meninggalkan artefak

spektral yang khas dan tidak kasat mata pada domain spasial, namun dapat terdeteksi pada domain frekuensi [13]. Karakteristik ini menjadikan domain frekuensi sebagai sumber informasi pelengkap yang penting, sehingga integrasi fitur spasial dan frekuensi secara bersamaan berpotensi menghasilkan representasi yang lebih komprehensif dibandingkan penggunaan salah satu domain saja.

Meskipun demikian, tantangan utama yang masih belum terselesaikan adalah penurunan performa yang signifikan ketika model diuji pada domain atau jenis manipulasi yang tidak dikenal selama pelatihan [14]. Model yang mencapai akurasi *in-domain* di atas 99% dapat mengalami penurunan drastis pada skenario *cross-domain*, sebagaimana ditunjukkan oleh penurunan akurasi hingga 42,30% ketika model yang dilatih pada satu jenis manipulasi diuji terhadap jenis manipulasi yang berbeda [14]. Kecenderungan model untuk *overfitting* terhadap karakteristik dataset pelatihannya menjadi penyebab utama lemahnya kemampuan generalisasi pada kondisi dunia nyata [15, 16]. Hal ini menegaskan bahwa keberhasilan deteksi tidak cukup hanya diukur dari performa *in-domain*, melainkan harus diuji melalui *cross-dataset testing* untuk memastikan model mampu menggeneralisasi terhadap artefak manipulasi yang beragam di luar data pelatihannya.

Guna mengatasi keterbatasan tersebut, penelitian ini mengusulkan arsitektur *dual-branch* yang mengintegrasikan domain spasial dan frekuensi secara simultan. *Branch* spasial mengekstraksi fitur struktural citra wajah menggunakan EfficientNet-B0 yang telah di-*pretrain*, sementara *branch* frekuensi terlebih dahulu mentransformasikan citra melalui *Frequency-Aware Decomposition* (FAD) sebelum diproses oleh backbone yang sama agar artefak spektral hasil proses generatif dapat ditangkap secara eksplisit. EfficientNet-B0 dipilih sebagai backbone karena menyeimbangkan kapasitas representasi dan efisiensi komputasi, sehingga cocok untuk deteksi pada citra wajah berukuran tetap tanpa beban model yang berlebihan.

Fitur yang diekstraksi oleh backbone belum tentu seluruhnya relevan bagi deteksi manipulasi. Oleh karena itu, modul *Convolutional Block Attention Module* (CBAM) disisipkan setelah setiap *stage* EfficientNet-B0 pada kedua *branch* untuk menyeleksi secara adaptif kanal dan wilayah spasial yang lebih diskriminatif. Keluaran kedua *branch* kemudian digabungkan melalui *concatenation* agar representasi spasial dan frekuensi saling melengkapi sebelum tahap klasifikasi.

Secara keseluruhan, alur penelitian ini meliputi praproses citra wajah, ekstraksi fitur melalui arsitektur *dual-branch* spasial–frekuensi yang dilengkapi modul CBAM, fusi representasi, serta klasifikasi citra sebagai asli atau palsu. Model dilatih pada FaceForensics++ dan diuji secara *in-domain* maupun *cross-dataset*

pada Celeb-DF v2 dan DFD untuk menilai kemampuan generalisasi terhadap jenis manipulasi yang tidak dikenal selama pelatihan. Rincian teoritis dan implementasi teknis diuraikan pada bab-bab berikutnya.

1.2 Rumusan Masalah

Sejalan dengan latar belakang penelitian, permasalahan yang dikaji dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana arsitektur *dual-branch* yang mengintegrasikan fitur domain spasial dan frekuensi dapat diimplementasikan untuk mendeteksi manipulasi wajah berbasis *deepfake*?
2. Bagaimana kemampuan generalisasi arsitektur yang diusulkan ketika dievaluasi secara *cross-dataset* terhadap dataset yang tidak dikenal selama pelatihan?

1.3 Batasan Permasalahan

Penelitian ini memiliki ruang lingkup yang dibatasi agar pembahasan tetap terarah dan sesuai dengan rumusan masalah yang telah ditetapkan. Adapun batasan permasalahan dalam penelitian ini adalah sebagai berikut.

- Penelitian difokuskan pada deteksi *deepfake* berbasis citra wajah statis (*frame-level*), tidak mencakup deteksi berbasis video, manipulasi audio, maupun deteksi *full-body*.
- Arsitektur dibatasi pada pendekatan *dual-branch* spasial-frekuensi dengan EfficientNet-B0 sebagai *backbone*.
- Dataset pelatihan dan validasi dibatasi pada FaceForensics++, sedangkan Celeb-DF v2 dan DeepFake Detection (DFD) hanya digunakan sebagai *cross-dataset test set* dan tidak dilibatkan dalam pelatihan.
- Deteksi dilakukan pada region wajah hasil *crop* berukuran 224×224 piksel, baik dari komposisi *setengah badan* maupun *full body*, selama wajah masih tampak jelas dan deteksi tidak mencakup keseluruhan tubuh.
- Penelitian dibatasi pada citra wajah yang *clean* atau tidak terhalang (*unoccluded*), sehingga wajah yang tertutup oleh masker, kacamata hitam, tangan, atau objek lain tidak termasuk dalam cakupan deteksi.

- Evaluasi kinerja dibatasi pada metrik ROC-AUC, *F1-Score*, *accuracy*, *precision*, dan *recall*.

1.4 Tujuan Penelitian

Mengacu pada rumusan masalah yang telah dirumuskan, penelitian ini memiliki tujuan umum dan tujuan khusus sebagai berikut:

- mengimplementasikan arsitektur *dual-branch* yang mengintegrasikan fitur domain spasial dan frekuensi untuk mendeteksi manipulasi wajah berbasis *deepfake*.
- mengevaluasi kemampuan generalisasi arsitektur yang diusulkan melalui pengujian *cross-dataset* terhadap dataset yang tidak dikenal selama pelatihan.

1.5 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan manfaat dalam berbagai aspek sebagai berikut:

- memberikan kontribusi terhadap pengembangan sistem deteksi *deepfake* melalui integrasi fitur domain spasial dan frekuensi dalam satu arsitektur.
- menawarkan analisis kemampuan generalisasi model melalui pengujian *cross-dataset* sebagai referensi bagi penelitian dan implementasi deteksi *deepfake* selanjutnya.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun menjadi lima bab utama, yaitu sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini mencakup latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah dan sistematika penulisan. Bagian ini memberikan gambaran umum mengenai alasan dan arah penelitian yang dilakukan.

- Bab 2 LANDASAN TEORI

  Bab ini membahas teori-teori yang relevan dengan penelitian, termasuk konsep *deepfake*, deteksi manipulasi citra wajah, transformasi domain frekuensi (DCT), arsitektur EfficientNet dan CBAM, serta penelitian terkait yang menjadi dasar konseptual dalam perancangan penelitian.

- Bab 3 METODOLOGI PENELITIAN

  Bab ini menjelaskan metode yang digunakan dalam penelitian, mulai dari pengumpulan data, praproses data, perancangan arsitektur *dual-branch* spasial–frekuensi, implementasi modul CBAM, serta protokol evaluasi performa model.

- Bab 4 HASIL DAN DISKUSI

  Bab ini menyajikan hasil eksperimen model dalam bentuk tabel dan grafik, meliputi performa *in-domain* pada FaceForensics++ dan pengujian *cross-dataset* pada Celeb-DF v2 dan DFD beserta pembahasannya.

- Bab 5 SIMPULAN DAN SARAN

  Bab ini berisi kesimpulan dari penelitian yang telah dilakukan, keterbatasan penelitian, serta saran untuk pengembangan penelitian selanjutnya.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA