

BAB II

KAJIAN LITERATUR

2.1 *Deepfake*

Istilah *deepfake* berasal dari gabungan dua kata, yaitu *deep learning* dan *fake*. Istilah ini merujuk pada konten multimedia, baik berupa gambar, video maupun audio, yang telah dimodifikasi secara visual, atau auditori dengan menggunakan teknik *deep learning* hingga tampak sangat realistis dan menyerupai aslinya [14]. Secara teknis, *deepfake* memanfaatkan model generatif seperti GAN atau arsitektur *encoder-decoder* untuk memodifikasi konten sintesis seperti *face swapping* atau *face reenactment*.

Secara umum, teknologi *deepfake* dikembangkan berdasarkan jaringan saraf tiruan (*neural network*). Dalam prosesnya, model dilatih untuk mempelajari pola, ekspresi, serta pergerakan wajah manusia, sehingga wajah seseorang pada suatu video dapat digantikan dengan wajah orang lain, sehingga menghasilkan video yang tampak realistis dan meyakinkan [1]. Dengan kemampuan tersebut, *deepfake* menjadi salah satu penerapan *deep learning* yang paling menonjol dalam bidang sintesis media [19].

Tujuan utama dari pengembangan *deepfake* pada awalnya adalah untuk penelitian dan pengembangan teknologi sintesis media yang realistis, bukan untuk penipuan atau penyebaran informasi palsu. Teknologi ini dirancang untuk mengeksplorasi kemampuan dari *neural network* dalam melakukan rekonstruksi ekspresi wajah, pergerakan bibir, serta sinkronisasi audio-visual dalam konteks hiburan, industri film, *lip-sync translation*, dan *virtual reality* [19]. Selain itu, *deepfake* juga dimanfaatkan dalam ranah ilmiah seperti pelatihan model *face recognition*, penciptaan data sintetis untuk meningkatkan generalisasi model, hingga simulasi terhadap perilaku manusia dalam bidang psikologi, dan forensik digital [14, 19].

Dalam konteks penelitian ini, istilah *deepfake* merujuk secara khusus pada manipulasi wajah pada citra dan *frame* video yang dihasilkan menggunakan model generatif berbasis *deep learning*, seperti GAN. Fokus penelitian ditujukan pada pendeteksian artefak visual dan spektral pada gambar serta *frame* video, bukan pada manipulasi audio atau bentuk media sintetis lainnya.

2.1.1 Ancaman *Deepfake*

Deepfake membawa berbagai potensi penyalahgunaan yang serius, mulai dari penyebaran informasi palsu, dan propaganda politik, hingga penipuan, dan kerusakan reputasi individu melalui konten palsu [14]. Salah satu bentuk penyalahgunaan yang paling meresahkan adalah *deepfake porn*, yakni video pornografi hasil rekayasa yang menampilkan wajah korban, seolah-olah korban terlibat dalam adegan tersebut. Studi global menunjukkan bahwa 96% dari seluruh video *deepfake* yang beredar di internet, berjenis pornografi, dan menargetkan perempuan [2]. Kondisi ini jelas mengancam privasi, dan kehormatan korban, serta dapat menimbulkan trauma dan kerugian sosial.

Selain ancaman terhadap individu, *deepfake* juga berdampak luas pada aspek sosial, politik, dan keamanan digital. Dalam konteks sosial, *deepfake* dapat menurunkan tingkat kepercayaan publik terhadap media visual, dan memperburuk fenomena *information disorder* di masyarakat [14, 15]. Pada bidang politik, teknologi ini dapat dimanfaatkan untuk menciptakan konten manipulatif yang dapat mengubah persepsi publik terhadap tokoh, atau peristiwa, dan berpotensi digunakan sebagai alat propaganda dan disinformasi terarah [1, 14].

2.1.2 Tantangan dalam Deteksi *Deepfake*

Perkembangan teknologi generatif seperti GAN membuat *deepfake* semakin sulit dibedakan dari konten asli. Peningkatan kualitas manipulasi menyebabkan konten palsu sulit dikenali, baik oleh pengamat manusia maupun oleh sistem pendeteksi konvensional yang hanya mengandalkan fitur spasial [14]. Kondisi ini menjadikan kemampuan mendeteksi dan memverifikasi keaslian konten digital sebagai kebutuhan penting untuk menjaga keandalan informasi.

Ancaman tersebut menunjukkan bahwa teknik deteksi konvensional yang hanya mengandalkan informasi spasial tidak lagi memadai untuk menghadapi kualitas *deepfake* modern. Banyak manipulasi visual kini dibuat sangat halus sehingga artefaknya tidak tampak jelas pada domain spasial. Oleh karena itu, diperlukan pendekatan pendeteksian yang lebih komprehensif dengan memanfaatkan informasi frekuensi untuk mengidentifikasi jejak matematis proses generatif, sehingga sistem deteksi menjadi lebih *robust* terhadap berbagai jenis teknik *deepfake* yang terus berkembang.

2.2 Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GAN) adalah model *deep learning* yang diperkenalkan pada tahun 2014 yang terdiri dari dua jaringan saraf, generator dan diskriminator. Generator berfungsi untuk menghasilkan data palsu yang menyerupai data asli (misalnya, gambar wajah sintesis), sedangkan diskriminator berfungsi membedakan apakah *input* yang diberikan merupakan data asli atau hasil dari generator. Keduanya dilatih secara bersamaan, di mana generator berusaha untuk “menipu” diskriminator dengan menghasilkan *output* yang semakin realistis, sementara diskriminator berusaha semakin mahir dalam mendeteksi mana data yang palsu.

Dalam proses pelatihan *adversarial* ini, generator secara bertahap belajar pola-pola dari data latih dan meningkatkan kualitas keluarannya hingga sedemikian rupa sehingga diskriminator tidak lagi mampu membedakan data sintesis dengan data asli. Pada saat generator dan diskriminator mencapai keseimbangan, *output* yang dihasilkan generator memiliki kemiripan statistik dan visual yang tinggi dengan data asli. Biasanya setelah proses *training* selesai, komponen diskriminator dapat diabaikan, dan generator digunakan sendiri untuk menghasilkan data baru yang realistis. Arsitektur ini yang membuat GAN sangat efektif dalam memodelkan distribusi data dan menciptakan sampel-sampel baru yang autentik [15, 20].

2.2.1 Peran GAN dalam Pembuatan Deepfake

Teknologi *deepfake* umumnya mengacu pada pembuatan media sintesis terutama *swap* wajah pada gambar atau video dengan tingkat kemiripan yang tinggi terhadap aslinya. GAN telah menjadi alat utama dalam banyak metode *deepfake* modern. Dalam pembuatan *deepfake*, GAN dimanfaatkan untuk menghasilkan wajah hasil pertukaran (*face swap*) atau wajah sintesis lainnya yang tampak *photo-realistic* dan sulit dibedakan dari wajah asli oleh mata manusia. Pada teknik *face swapping* berbasis GAN, model generator mempelajari ciri-ciri wajah target dan mencoba menggabungkan wajah sumber ke dalam wajah target sehingga sulit terdeteksi perbedaan, sementara diskriminator berperan mengidentifikasi bagian tidak asli pada hasil gabungan tersebut. Proses *adversarial* ini memastikan hasil *deepfake* memiliki detail visual yang konsisten (misalnya pencahayaan, pose, ekspresi) dan seminimal mungkin artefak yang mencurigakan [15].

Metode *deepfake* umumnya menggunakan arsitektur *autoencoder* pada dua wajah (sumber dan target) untuk menukar wajah sumber dan target, lalu ditingkatkan dengan

komponen *adversarial* dengan GAN agar hasilnya lebih halus dan nyata. Penerapan GAN dalam *deepfake* memungkinkan generator belajar memadukan wajah ke tubuh target secara rapi dan sulit terdeteksi, sementara diskriminator memastikan video atau gambar asli hasil manipulasi tampak alami. Pendekatan GAN ini digunakan dalam berbagai perangkat lunak *deepfake* modern untuk menghasilkan video *swap* wajah yang sulit dibedakan dari video asli [1, 15].

Pada proses manipulasi citra, GAN menggunakan operasi *upsampling* untuk meningkatkan resolusi dari citra berukuran kecil menjadi citra beresolusi tinggi. Proses ini sering kali dilakukan dengan metode interpolasi, yang secara matematis menyisipkan nilai nol dalam sinyal. Langkah ini menyebabkan replikasi spektrum frekuensi dari citra sumber ke area frekuensi tinggi yang baru, menghasilkan pola distorsi khas atau *spectral artifacts* [8, 9]. Artefak ini tidak mudah dikenali secara visual di domain spasial, namun dapat diidentifikasi pada domain frekuensi menggunakan transformasi seperti FFT. Artefak yang muncul akibat proses pembangkitan GAN inilah yang menjadi dasar metode deteksi *deepfake* berbasis analisis frekuensi [9].

2.2.2 *Upsampling dan Interpolasi dalam Proses Generatif Deepfake*

Upsampling merupakan proses meningkatkan resolusi citra atau *feature map* pada generator GAN dengan menambah jumlah piksel baru. Teknik umum yang digunakan meliputi *nearest-neighbor upsampling*, *bilinear interpolation*, dan *transposed convolution*, yang masing-masing memiliki karakteristik dan kelemahan tersendiri [21, 22]. Pada arsitektur GAN, khususnya saat menggunakan *transposed convolution*, proses *upsampling* dapat menghasilkan pola artefak periodik seperti *checkerboard patterns* karena distribusi *kernel* yang tidak merata [21]. Selain itu, berbagai studi menunjukkan bahwa citra hasil GAN cenderung memiliki bias frekuensi tinggi yang tidak muncul pada citra natural, sehingga memunculkan pola anomali pada domain frekuensi [8].

Interpolasi adalah teknik estimasi nilai piksel baru berdasarkan piksel di sekitarnya, dan digunakan dalam berbagai proses manipulasi citra pada *deepfake*, termasuk *face warping*, *resizing*, dan tahap *upsampling* pada GAN. Metode seperti *nearest*, *bilinear*, dan *bicubic interpolation* menghasilkan karakteristik frekuensi yang berbeda, yang pada citra *deepfake* sering meninggalkan distribusi frekuensi tinggi.

Berbagai studi, menunjukkan bahwa proses *upsampling* pada GAN menjadi sumber utama pola artefak *high-frequency* yang tidak muncul pada citra natural. Artefak inilah yang dapat dimanfaatkan oleh metode deteksi berbasis domain frekuensi seperti FFT [8, 9].

2.2.3 Artefak yang Dihasilkan Proses Generatif *Deepfake*

Proses generatif pada *deepfake*, khususnya yang berbasis GAN, menghasilkan sejumlah artefak khas yang menjadi dasar bagi metode deteksi. Artefak-arterfak ini muncul karena keterbatasan model generatif dalam mereplikasi seluruh karakteristik statistik citra alami, baik pada domain spasial maupun domain frekuensi [8, 9]. Secara umum, artefak yang dihasilkan proses generatif *deepfake* dapat diklasifikasikan ke dalam dua kategori utama:

1. Artefak Domain Spasial

Menurut *Afchar et al.*, wajah hasil *deepfake* cenderung buram dan halus, kehilangan detail halus seperti tekstur kulit, sedangkan area yang tidak dimanipulasi tetap kaya detail [4]. Selain itu, *Rao et al.* mencatat bahwa proses pembangkitan *deepfake* meninggalkan sejumlah artefak visual, antara lain distorsi akibat penyesuaian (*warping*) wajah sintetis terhadap wajah sumber, tepi wajah yang buram (*blurred edges*), koreksi fitur wajah yang tidak wajar, serta inkonsistensi tekstur dan distorsi warna [15, 23]. Meskipun artefak spasial dapat dideteksi oleh model CNN seperti XceptionNet, kualitas manipulasi yang semakin tinggi pada GAN modern membuat artefak ini semakin sulit diamati [14].

2. Artefak Domain Frekuensi

Artefak frekuensi muncul akibat ketidakmampuan model generatif dalam mereproduksi distribusi spektral alami citra. Jenis artefak ini meliputi distribusi energi frekuensi tinggi yang berlebihan akibat proses *upsampling*, pola *checkerboard* yang dihasilkan oleh *transposed convolution*, *periodic noise* berupa titik-titik intensitas berulang pada spektrum FFT, serta *spectral dropoff* yang tidak konsisten dengan pola penurunan energi alami [8, 9, 21]. Artefak frekuensi bersifat algoritmik dan lebih stabil dibanding artefak spasial, sehingga menjadi dasar penting bagi metode deteksi berbasis FDA [11].

Kedua kategori artefak ini bersifat komplementer, di mana artefak spasial menangkap inkonsistensi visual yang tampak, sedangkan artefak frekuensi mengungkap jejak matematis proses generatif yang tersembunyi. Oleh karena itu, pendekatan deteksi yang menggabungkan analisis kedua domain diharapkan dapat menghasilkan sistem yang lebih *robust* dan mampu mendeteksi berbagai jenis manipulasi *deepfake* [13, 16].

2.3 Pendekatan Deteksi *Deepfake*

Pendekatan deteksi *deepfake* dapat diklasifikasikan berdasarkan domain analisis yang digunakan. Secara umum, terdapat dua pendekatan utama, yaitu deteksi berbasis domain spasial yang menganalisis fitur visual langsung dari citra, dan deteksi berbasis domain frekuensi yang menganalisis representasi spektral untuk mengidentifikasi artefak yang tidak terlihat secara kasat mata. Beberapa penelitian terkini juga mengeksplorasi pendekatan *hybrid* yang menggabungkan kedua domain untuk meningkatkan akurasi dan kemampuan generalisasi. Bagian berikut menguraikan masing-masing pendekatan beserta perbandingannya.

2.3.1 Deteksi Berbasis Domain Spasial

Pendekatan berbasis domain spasial merupakan metode deteksi *deepfake* yang berfokus pada analisis langsung terhadap piksel citra atau *frame* video untuk menemukan pola dan artefak anomali. Detektor spasial bekerja dengan mempelajari perbedaan visual antara konten asli dan konten hasil manipulasi, seperti ketidaksesuaian tekstur kulit, pencahayaan, tepi wajah, bayangan, maupun distorsi geometrik pada area yang telah dimodifikasi [14].

Beberapa penelitian mengembangkan model deteksi dengan arsitektur CNN khusus, seperti *MesoNet*, yang dirancang ringan namun efektif dalam mendeteksi artefak lokal pada wajah sintesis [4]. Selain itu, arsitektur ResNet yang diperkenalkan oleh He et al. [5], dan XceptionNet oleh Chollet [6], menjadi *backbone* populer dalam sistem *deepfake*, karena kemampuannya dalam mengekstraksi fitur spasial secara mendalam. Penelitian oleh Haq [3] menunjukkan bahwa model XceptionNet, dan ResNet-50 mampu melakukan klasifikasi video *deepfake* dengan akurasi tinggi, terutama ketika dikombinasikan dengan metode ekstraksi fitur seperti *Local Binary Pattern* (LBP).

Dataset seperti FFPP juga menjadi standar dalam pelatihan dan evaluasi model deteksi berbasis spasial. *Dataset* ini mencakup berbagai jenis manipulasi wajah, tingkat kompresi, dan kondisi pencahayaan, sehingga mendukung penelitian terhadap berbagai variasi *deepfake* [7].

Secara umum, pendekatan spasial menawarkan performa yang baik dalam kondisi data yang serupa dengan data pelatihan, di mana model dapat mendeteksi kepalsuan ketika dihadapkan dengan data yang sama, namun ketika dihadapkan dengan data yang berbeda, model cenderung bergantung pada artefak permukaan yang spesifik terhadap proses

pembuatan tertentu, sehingga akurasi dapat menurun, dan model berpotensi tidak dapat mendeteksi anomali tersebut [14]. Oleh karena itu, penelitian terbaru mulai mengombinasikan pendekatan spasial dengan FDA untuk meningkatkan *robustness* dan kemampuan generalisasi terhadap berbagai jenis *deepfake* yang semakin kompleks.

2.3.2 Deteksi Berbasis Domain Frekuensi

Pendekatan berbasis domain frekuensi atau FDA merupakan metode deteksi *deepfake* yang terfokus pada menganalisis komponen spektral dari sebuah citra atau video. Berbeda dengan domain spasial yang bekerja secara langsung terhadap intensitas piksel, pendekatan dengan FDA mengubah citra ke dalam representasi frekuensi dengan menggunakan transformasi matematis seperti *Fast Fourier Transform* (FFT). Dengan representasi ini, karakteristik yang tersembunyi dan tidak terdeteksi di domain spasial, terutama artefak yang muncul akibat proses generatif seperti *upsampling* atau *interpolation*, dapat terlihat lebih jelas [8, 9, 11].

Durall et al. [8] mengemukakan bahwa model generatif berbasis CNN, seperti GAN, kerap tidak mampu mereplikasi karakteristik distribusi spektral yang lazim ditemukan pada citra asli. Hal ini menyebabkan gambar sintetis menampilkan perbedaan pola dalam domain frekuensi, khususnya pada bagian frekuensi tinggi. *Zhang et al.* [9] menyatakan bahwa artefak dalam domain frekuensi dapat ditiru dan dimanfaatkan sebagai semacam sidik jari digital untuk mengidentifikasi gambar yang telah dimanipulasi. Studi lanjutan oleh *Qian et al.* [11] memperkenalkan pendekatan *Thinking in Frequency*, dengan mengeksplorasi petunjuk frekuensi guna memperkuat kemampuan deteksi dan meningkatkan generalisasi pada berbagai *dataset*.

Pendekatan berbasis frekuensi memiliki keunggulan utama dalam mendeteksi anomali spektral yang tidak terlihat oleh mata manusia. Sebagai contoh, citra wajah hasil *deepfake* sering memperlihatkan distribusi spektral yang tidak konsisten, atau pola frekuensi konsentris yang menyimpang dari citra asli. Metode seperti FSBI oleh *Hasanaath et al.* [16], dan pendekatan *Frequency Space Domain Learning* (FSDL) oleh *Tan et al.* [12], menunjukkan bahwa pemanfaatan informasi dalam domain frekuensi dapat memperbaiki akurasi serta mendukung generalisasi antar-*dataset* pada sistem pendeteksi *deepfake*. Dalam upaya meningkatkan ketahanan model terhadap manipulasi berbasis generatif, *Luo dan Wang* [17] merancang metode *Frequency-Domain Masking* yang secara adaptif menggabungkan elemen spasial dan frekuensi dalam proses pendeteksian.

Berbagai studi tersebut mengindikasikan bahwa informasi domain frekuensi dapat melengkapi pendekatan berbasis spasial karena mampu menangkap karakteristik yang tidak tampak langsung pada ruang gambar. Artefak spektral yang konsisten pada beragam jenis *deepfake* membuat pendekatan FDA lebih andal terhadap variasi model generatif maupun perbedaan distribusi antar *dataset*. Hal ini mendorong berkembangnya penelitian yang menggabungkan analisis spasial dan frekuensi dalam satu kerangka *deep learning* untuk menghasilkan detektor yang lebih akurat dan tahan terhadap teknik manipulasi baru.

2.3.3 Perbandingan Domain Spasial dengan Domain Frekuensi

Pendekatan deteksi *deepfake* berbasis domain spasial dan domain frekuensi memiliki karakteristik, kekuatan, dan kelemahan yang berbeda. Perbandingan kedua pendekatan ini penting untuk memahami mengapa integrasi keduanya dalam satu arsitektur *hybrid* dapat menghasilkan sistem deteksi yang lebih *robust* [8, 9, 14].

Tabel 2.1 Perbandingan Pendekatan Domain Spasial dan Domain Frekuensi .

Aspek Perbandingan	Domain Spasial	Domain Frekuensi (FDA)
Jenis informasi	Intensitas piksel, tekstur, warna, tepi objek	Distribusi energi frekuensi, pola spektral, amplitudo dan fase
Artefak yang dideteksi	Ketidaksesuaian <i>blending</i> , distorsi geometri, inkonsistensi pencahayaan	<i>Spectral distortions, periodic noise, checkerboard artifacts</i>
Sensitivitas	Tinggi terhadap artefak visual yang tampak	Tinggi terhadap artefak algoritmik yang tersembunyi
Generalisasi lintas <i>dataset</i>	Terbatas, cenderung <i>overfit</i> pada tekstur spesifik <i>dataset</i>	Lebih baik, karena artefak frekuensi bersifat universal
Ketahanan terhadap kompresi	Menurun signifikan pada kompresi tinggi R	elatif stabil, pola spektral tetap terdeteksi
Kelemahan utama	Bergantung pada pola visual yang dapat disamarkan oleh GAN modern	Kehilangan informasi konteks spasial (lokasi artefak)
Contoh arsitektur	XceptionNet, ResNet, <i>MesoNet</i>	FT-based classifier, DCT-based detector

Dari perbandingan tersebut, terlihat bahwa pendekatan spasial unggul dalam menangkap anomali visual lokal seperti tekstur dan tepi wajah, namun rentan terhadap peningkatan kualitas GAN modern yang semakin mampu menyamarkan artefak visual [14].

Sebaliknya, pendekatan frekuensi mampu menangkap pola algoritmik yang bersifat universal dan stabil, tetapi kehilangan informasi mengenai lokasi spasial dari artefak tersebut [8].

Sifat komplementer kedua domain inilah yang mendorong banyak penelitian terkini untuk menggabungkan analisis spasial dan frekuensi dalam satu kerangka kerja deteksi, guna memperoleh representasi fitur yang lebih lengkap dan *robust* terhadap berbagai variasi *deepfake* [13, 11, 16].

2.3.4 Pendekatan *Hybrid Domain Spasial-Frekuensi*

Pendekatan *hybrid domain* spasial-frekuensi merupakan strategi deteksi *deepfake* yang mengintegrasikan fitur dari kedua domain dalam satu kerangka kerja *deep learning*. Tujuan utama pendekatan ini adalah memanfaatkan sifat komplementer kedua domain: domain spasial menangkap anomali visual seperti tekstur, tepi, dan konsistensi pencahayaan, sedangkan domain frekuensi mengungkap artefak algoritmik yang tidak terlihat secara visual, seperti *spectral distortions* dan *periodic noise* [13, 16].

Secara arsitektural, terdapat dua strategi utama dalam mengintegrasikan domain spasial dan frekuensi [11, 13, 16]:

1. *Early Fusion* (Fusi Awal)

Pada pendekatan *early fusion*, representasi frekuensi digabungkan dengan citra RGB sebagai kanal tambahan sebelum diproses oleh model CNN. Sebagai contoh, spektrum magnitudo FFT ditambahkan sebagai kanal keempat sehingga *input* model menjadi *tensor* 4-kanal (RGB + FFT). Dengan cara ini, seluruh informasi spasial dan frekuensi diproses bersamaan oleh satu jaringan *konvolusional*. Pendekatan ini memungkinkan model mempelajari interaksi antara fitur spasial dan frekuensi sejak lapisan konvolusi pertama, namun memerlukan arsitektur yang mampu menangani *input* multi-kanal secara efektif [11].

2. *Late Fusion* (Fusi Akhir) / *Two-Branch Architecture*

Pada pendekatan *late fusion*, fitur spasial dan fitur frekuensi diekstraksi secara terpisah oleh dua *branch* atau cabang jaringan yang independen. Cabang spasial menggunakan model CNN (misalnya XceptionNet) untuk mengekstraksi fitur dari citra RGB, sedangkan cabang frekuensi menggunakan CNN terpisah yang lebih ringan untuk memproses representasi FFT. Fitur-fitur dari kedua cabang kemudian digabungkan (*concatenated*) dan diteruskan ke lapisan klasifikasi bersama. Pendekatan ini

memberikan fleksibilitas dalam memilih arsitektur yang optimal untuk masing-masing domain, serta memungkinkan setiap cabang untuk mengekstraksi fitur secara spesifik tanpa saling mengganggu [11, 13].

Beberapa penelitian terkini yang mengadopsi pendekatan *hybrid* melaporkan peningkatan performa. *Alam et al.* melalui SpecXNet menggabungkan fitur spasial dari XceptionNet dengan informasi frekuensi untuk meningkatkan ketahanan deteksi terhadap berbagai jenis manipulasi [13]. *Hasanaath et al.* dengan metode FSBI memanfaatkan fitur frekuensi eksplisit untuk memperkuat kemampuan generalisasi lintas *dataset* [16]. *Luo dan Wang* mengembangkan mekanisme *frequency-domain masking* yang secara adaptif mengintegrasikan informasi spasial dan frekuensi [17]. Namun, peningkatan tersebut tidak selalu konsisten karena efektivitas pendekatan *hybrid* masih bergantung pada representasi frekuensi, strategi fusi, dan karakteristik *dataset*.

Penelitian ini berfokus pada strategi *late fusion*, yang diimplementasikan melalui arsitektur *two-branch* yang menggabungkan XceptionNet untuk fitur spasial dan CNN ringan untuk fitur frekuensi, dilengkapi *SE gating*. Pemilihan *late fusion* didasari fleksibilitas pemodelan arsitektur yang berbeda pada setiap domain. Di sisi lain, strategi *early fusion* (XceptionNet 4-kanal RGB+FFT) disajikan sebagai alternatif konseptual, namun tidak menjadi fokus evaluasi eksperimental pada penelitian ini.

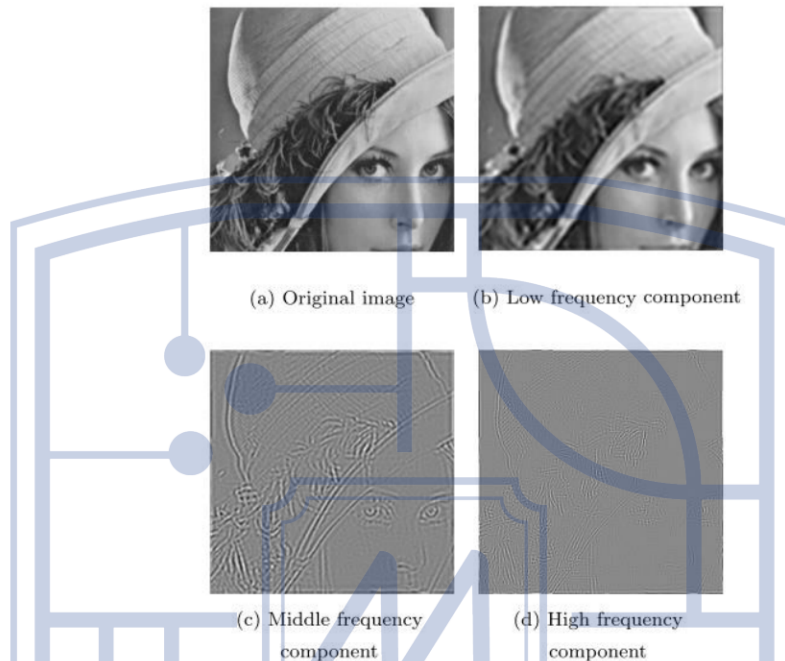
2.4 Frequency Domain Analysis (FDA)

Frequency Domain Analysis (FDA) merupakan pendekatan yang mengubah representasi citra dari domain spasial ke domain frekuensi untuk mengungkap pola-pola yang tidak terdeteksi oleh analisis visual langsung. Dalam konteks deteksi *deepfake*, FDA sangat relevan karena proses sintesis generatif menghasilkan artefak spektral yang khas, seperti distorsi distribusi frekuensi dan puncak periodik pada komponen frekuensi tinggi. Bagian berikut menjelaskan konsep dasar domain frekuensi, transformasi *Fourier*, dan relevansinya terhadap deteksi *deepfake*.

2.4.1 Konsep Domain Frekuensi

Dalam domain spasial, setiap piksel merepresentasikan intensitas atau warna di posisi tertentu dalam citra. Sebaliknya, dalam domain frekuensi, citra dianggap sebagai kombinasi dari gelombang sinusoidal dengan berbagai frekuensi dan amplitudo yang secara bersama-sama membentuk pola keseluruhan gambar [8].

Transformasi dari domain spasial ke domain frekuensi dilakukan menggunakan metode matematis seperti *Fast Fourier Transform* (FFT), *Discrete Cosine Transform* (DCT), atau *Wavelet Transform*. Melalui proses ini, citra dapat diuraikan menjadi dua komponen utama, sebagaimana dijelaskan pada Gambar 2.1 berikut:



Gambar 2.1 Distribusi Komponen *Low-Frequency* dan *High-Frequency* [24]

Perbandingan komponen *low-frequency* dan *high-frequency* dalam citra hasil transformasi FFT:

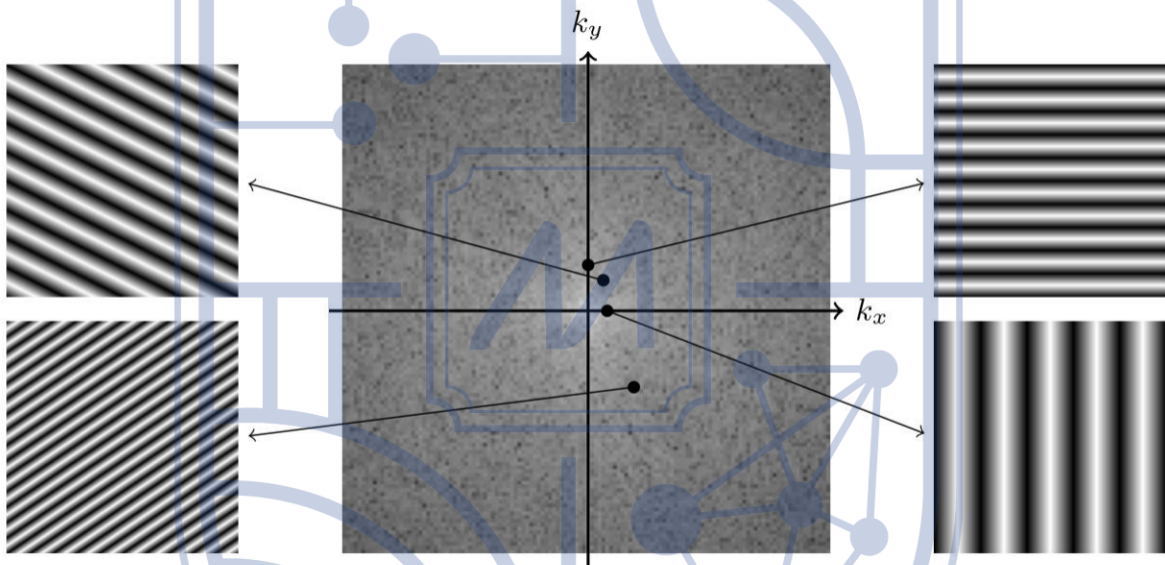
1. Citra kiri: hasil *inverse* FFT menggunakan hanya frekuensi rendah, menampilkan bentuk umum dan pencahayaan lembut.
2. Citra kanan: hasil *inverse* FFT dengan hanya frekuensi tinggi, menonjolkan tepi, detail, dan tekstur tajam.

Tabel 2.2 Komponen dalam Domain Frekuensi [8, 9]

Komponen Frekuensi	Deskripsi	Ciri Visual	Peran dalam Analisis Citra
<i>Low-Frequency Components</i>	Menyimpan informasi global seperti bentuk, pencahayaan, dan area halus citra.	Terpusat di bagian tengah spektrum frekuensi (warna terang di tengah).	Memberikan gambaran umum struktur citra dan pola besar.

Komponen Frekuensi	Deskripsi	Ciri Visual	Peran dalam Analisis Citra
<i>High-Frequency Components</i>	Mengandung detail halus seperti tepi, tekstur, dan perubahan intensitas tajam.	Muncul di bagian tepi spektrum (warna terang atau garis halus).	Mengungkap artefak manipulasi, distorsi, atau detail buatan.

Pendekatan ini menjadi penting dalam deteksi *deepfake* karena proses generatif pada *Generative Adversarial Networks* (GAN) sering kali meninggalkan artefak matematis akibat operasi *upsampling* dan *interpolation*, yang sulit diamati pada domain spasial tetapi muncul jelas pada domain frekuensi [8, 9].



Gambar 2.2 Representasi Domain Frekuensi [25]

Domain frekuensi menggambarkan bagaimana berbagai gelombang sinusoidal dengan frekuensi yang berbeda menyusun pola citra. Warna terang di pusat spektrum menandakan frekuensi rendah (bentuk global dan pencahayaan), sementara area terang di tepi menandakan frekuensi tinggi (tepi, tekstur, detail halus).

2.4.2 Transformasi *Fourier* (FFT)

Transformasi *Fourier* dua dimensi mengubah citra spasial $f(x,y)$ berukuran dua dimensi $M \times N$ menjadi representasi frekuensi $F(u,v)$, yang menggambarkan distribusi energi pada berbagai tingkat frekuensi.

Persamaan umum *Discrete Fourier Transform* (DFT) adalah:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-j \cdot 2\pi \left(\frac{u \cdot x}{M} + \frac{v \cdot y}{N} \right)} \quad (2.1)$$

di mana $f(x,y)$ menyatakan nilai piksel citra pada posisi spasial (x,y) . Adapun $F(u,v)$ merupakan koefisien frekuensi pada koordinat (u,v) , sedangkan M dan N adalah dimensi citra (jumlah kolom dan baris). Simbol j adalah satuan imajiner, yaitu $j = \sqrt{-1}$. Faktor eksponensial kompleks pada persamaan tersebut berperan sebagai fungsi basis sinusoidal sesuai rumus Euler, $e^{-j\theta} = \cos \theta - j \sin \theta$, sehingga setiap nilai $F(u,v)$ menunjukkan seberapa besar kontribusi gelombang sinusoidal berfrekuensi (u,v) dalam menyusun citra.

Persamaan *inverse*-nya (*Inverse DFT*) untuk mengembalikan ke domain spasial:

$$f(x,y) = \left(\frac{1}{M \cdot N} \right) \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u,v) \cdot e^{j \cdot 2\pi \left(\frac{u \cdot x}{M} + \frac{v \cdot y}{N} \right)} \quad (2.2)$$

Hasil transformasi *Fourier* berupa bilangan kompleks yang terdiri dari bagian real dan imajiner. Untuk mendapatkan distribusi energi, dihitung nilai *magnitude spectrum*:

$$S(u,v) = |F(u,v)| = \sqrt{\text{Re}(F(u,v))^2 + \text{Im}(F(u,v))^2} \quad (2.3)$$

Agar pola energi frekuensi lebih mudah divisualisasikan, nilai magnitudo biasanya diubah ke skala logaritmik:

$$S_{\log(u,v)} = \log(1 + S(u,v)) \quad (2.4)$$

Spektrum magnitudo, yang menampilkan intensitas energi pada berbagai tingkat frekuensi:

1. Pusat spektrum (*DC component*) mewakili frekuensi rendah (informasi global citra).
2. Area tepi menunjukkan frekuensi tinggi (detail halus dan tepi objek).

Distribusi energi yang alami pada citra nyata menurun secara bertahap dari pusat ke arah tepi, sedangkan pada citra hasil GAN, pola ini menunjukkan *spectral anomaly*, seperti pola *radial*, *spike*, atau garis periodik yang tidak wajar [8, 11].

2.4.3 Spektrum Magnitudo dan Fase

Hasil Transformasi *Fourier* sebuah citra berupa bilangan kompleks yang dapat diuraikan menjadi dua komponen, yaitu **spektrum magnitudo** ($|F(u,v)|$) dan **spektrum fase** ($\angle F(u,v)$). Spektrum magnitudo menyatakan seberapa besar energi yang terkandung pada setiap frekuensi, sedangkan spektrum fase menyatakan posisi serta penyelarasan setiap komponen frekuensi yang menentukan susunan struktur spasial citra [25]. Sebagian besar pendekatan deteksi *deepfake* berbasis frekuensi, termasuk representasi yang digunakan pada

penelitian ini, hanya memanfaatkan spektrum magnitudo, sementara informasi fase umumnya diabaikan.

Padahal, sejumlah studi menunjukkan bahwa fase membawa porsi informasi struktural yang besar. *Oppenheim* dan *Lim* [26] memperlihatkan bahwa citra yang direkonstruksi hanya dari spektrum fase masih mempertahankan struktur penting seperti tepi dan bentuk objek, sedangkan rekonstruksi hanya dari spektrum magnitudo justru kehilangan struktur tersebut. Dalam konteks deteksi pemalsuan wajah, *Liu et al.* [27] melalui metode *Spatial-Phase Shallow Learning* (SPSL) menunjukkan bahwa spektrum fase efektif menonjolkan artefak akumulatif dari proses *up-sampling* generatif yang tidak tampak pada domain spasial maupun pada spektrum magnitudo. Dengan demikian, representasi frekuensi yang hanya berbasis magnitudo memberikan gambaran spektral yang tidak lengkap, sebuah keterbatasan desain yang relevan dalam menafsirkan kinerja cabang frekuensi berbasis-magnitudo.

2.4.4 Artefak Frekuensi pada Citra Deepfake

Jaringan GAN dengan operasi *upsampling transposed-convolution* secara sistematis menghasilkan distribusi spektral yang lebih landai (*spectral fall-off*) dibanding citra natural, sehingga rasio energi *high-band* terhadap *low-band* dapat dijadikan ciri pembeda [8]. *Zhang et al.* mengonfirmasi bahwa artefak frekuensi tinggi ini konsisten di berbagai jenis GAN dan dapat berfungsi sebagai “GAN fingerprints” [8, 9].

Selanjutnya, *Qian et al.* memperkenalkan konsep *Frequency-aware Clues*, yang membuktikan bahwa detektor yang memanfaatkan sinyal frekuensi dapat lebih akurat dalam mendeteksi citra sintesis [11]. *Hasanaath et al.* melalui metode FSBI menegaskan bahwa fitur frekuensi membantu meningkatkan *robustness* dan kemampuan generalisasi antar-dataset [16]. *Tan et al.* dan *Luo dan Wang* juga menunjukkan bahwa pembelajaran adaptif pada domain frekuensi menghasilkan peningkatan akurasi pada deteksi lintas arsitektur generatif dan *cross-dataset* [17, 28].

Tabel 2.3 Karakteristik Artefak Frekuensi pada Deepfake [8, 9, 11, 16]

Jenis Artefak	Deskripsi	Penyebab Utama
Distribusi energi tinggi pada <i>high-frequency</i>	Intensitas tinggi di area tepi spektrum.	Proses <i>upsampling</i> GAN.
Pola radial/simetri berulang	Pola garis melingkar di sekitar pusat spektrum.	Interpolasi atau <i>kernel transposed convolution</i> .

Jenis Artefak	Deskripsi	Penyebab Utama
Ketidakseimbangan arah energi	Energi condong pada orientasi tertentu.	Struktur jaringan tidak isotropik.
Noise periodik	Pola titik-titik berulang.	Zero-padding atau <i>aliasing</i> .

2.4.5 Bias Frekuensi dan Tekstur pada CNN

Efektivitas sebuah arsitektur dalam memanfaatkan petunjuk frekuensi tidak hanya ditentukan oleh ketersediaan sinyal, tetapi juga oleh kecenderungan induktif (*inductive bias*) jaringan itu sendiri. Geirhos et al. [29] menunjukkan bahwa CNN yang dilatih pada citra natural cenderung mengambil keputusan berdasarkan tekstur lokal alih-alih bentuk global. Kecenderungan ini menjelaskan mengapa *backbone* spasial seperti XceptionNet efektif menangkap anomali tekstur kulit dan ketidaksesuaian *blending* pada wajah *deepfake*, karena ranah kerjanya selaras dengan bias bawaan CNN.

Di sisi lain, jaringan saraf memiliki *spectral bias* yang cenderung mempelajari komponen frekuensi rendah terlebih dahulu, kemudian jika memungkinkan, komponen frekuensi tinggi [30]. Wang et al. [31] lebih jauh menunjukkan bahwa kemampuan generalisasi CNN berkaitan erat dengan cara jaringan memanfaatkan komponen frekuensi tinggi citra. Implikasinya bagi penelitian ini, sebuah CNN frekuensi yang relatif dangkal dan menerima peta magnitudo FFT mentah harus mempelajari selektivitas frekuensi yang diskriminatif tanpa dukungan bias arsitektural yang memadai. Ketika sinyal artefak lemah atau terdistorsi, jaringan semacam ini rentan gagal menemukan pola yang berguna. Pemahaman ini menjadi landasan teoretis dalam membahas kinerja cabang frekuensi pada arsitektur *hybrid*.

2.4.6 Peran Frequency Domain Analysis dalam Deteksi Deepfake

FDA berperan penting dalam mendeteksi konten sintesis karena artefak yang muncul di domain frekuensi bersifat universal dan algoritmik, bukan semata-mata visual. Pendekatan ini terbukti efektif dalam:

1. Deteksi lintas arsitektur generatif, karena setiap GAN meninggalkan GAN *fingerprints* yang unik [9].
2. *Robustness* terhadap kompresi dan *noise*, sebab pola frekuensi relatif stabil meski citra terkompresi [16].

3. Efisiensi representasi fitur, karena hanya memerlukan peta magnitudo atau fase spektrum [12].

Namun, FDA juga memiliki keterbatasan:

1. Hilangnya informasi konteks spasial (lokasi artefak).
2. Sensitivitas terhadap rotasi atau translasi citra.
3. Memerlukan integrasi dengan *spatial feature extractor* agar hasil deteksi lebih komprehensif.

Untuk itu, banyak penelitian terkini menggabungkan FDA dengan jaringan *konvolusional* seperti XceptionNet dalam model *hybrid spatial-frequency*, guna memanfaatkan kekuatan analisis struktural spasial dan deteksi artefak frekuensi secara bersamaan [13, 16, 17, 28].

2.5 *Spectral Distortions* dalam Deteksi *Deepfake*

Pendekatan deteksi berbasis domain frekuensi semakin banyak digunakan karena ditemukannya pola *spectral distortions*, yaitu penyimpangan distribusi spektral yang muncul secara sistematis pada citra dan video manipulasi. Distorsi ini muncul akibat keterbatasan model generatif seperti GAN dan *encoder-decoder* CNN dalam mereproduksi distribusi frekuensi yang alami [8, 9]. Berbeda dengan artefak spasial yang dapat disamarkan oleh model generatif generasi baru, *spectral distortions* bersifat lebih stabil karena berakar pada proses matematis dari *up-convolution*, *interpolation*, dan *feature reconstruction* [8, 21].

2.5.1 Akar Kemunculan *Spectral Distortions* pada Model Generatif

Model generatif berbasis CNN memiliki keterbatasan dalam mempelajari struktur statistik frekuensi alami. Operasi *up-convolution* dengan *stride* lebih dari 1 menghasilkan pola periodik di domain frekuensi karena *kernel overlap* yang tidak seragam. Fenomena ini dikenal dengan *checkerboard artifacts* [8, 21].

Zhang *et al.* [9] juga menemukan bahwa citra hasil sintesis GAN memiliki anomali spektral pada frekuensi menengah hingga tinggi. Pola ini muncul sebagai “sidik jari” khas citra manipulasi (GAN *fingerprints*), yang dapat dikenali tanpa bergantung pada ciri spasial. Kesulitan model generatif mereproduksi distribusi frekuensi ini merupakan penyebab utama mengapa *spectral distortions* tetap muncul bahkan pada GAN canggih seperti *Progressive Growing of GANs* (PGGAN) [28].

2.5.2 Jenis-Jenis *Spectral Distortions* pada Gambar dan Video *Deepfake*

1. Distorsi Frekuensi Tinggi

Model generatif sering gagal menghasilkan detail mikro yang alami, menyebabkan ketidakseimbangan energi pada komponen frekuensi tinggi. *Qian et al.* [11] serta *Mejri et al.* [32] menunjukkan bahwa *deepfake* cenderung memiliki tekstur kulit yang terlalu halus dan pola frekuensi tinggi yang tidak konsisten. Distorsi ini menjadi salah satu indikator paling kuat dalam deteksi berbasis domain frekuensi.

2. *Checkerboard Patterns*

Pola *checkerboard* merupakan artefak klasik yang muncul dari *transposed convolution* dan menjadi penyebab dominan *spectral distortions* pada banyak model GAN [8, 21]. Artefak ini tampak sebagai pola *grid* berulang pada domain spasial dan sebagai titik-titik intensitas periodik pada domain frekuensi.

3. Anomali Koefisien DCT

Giudice et al. [10] menemukan bahwa citra GAN menampilkan anomali pada koefisien DCT, terutama pada *band* frekuensi menengah. Ketidakseimbangan ini disebabkan oleh ketidakmampuan GAN untuk mempelajari korelasi alami antar blok DCT yang lazim pada citra asli.

4. Distorsi Frekuensi Menengah

Penelitian FSBI [16] dan SpecXNet [13] melaporkan bahwa *deepfake* menunjukkan perbedaan signifikan pada *band* frekuensi menengah. *Band* ini sangat penting karena berhubungan dengan pola tekstur wajah seperti kontur, lipatan, dan mikro-ekspresi.

5. Distorsi Spektral Temporal pada Video

Anomali frekuensi pada video *deepfake* tidak hanya terjadi antarpiksel dalam satu *frame*, tetapi juga antarframe. *Kim et al.* [33] menemukan bahwa manipulasi wajah menimbulkan inkonsistensi frekuensi *temporal*. Hasil ini sejalan dengan *Nguyen et al.* serta *Güera dan Delp* yang melaporkan bahwa video *deepfake* cenderung memiliki pola frekuensi *temporal* yang tidak stabil selama pergerakan wajah [34, 35].

2.5.3 Relevansi *Spectral Distortions* untuk Deteksi *Deepfake*

Analisis domain frekuensi memiliki keunggulan penting dibanding pendekatan berbasis spasial. *Rao dan Uehara* [15] serta *Rana et al.* [14] menegaskan bahwa peningkatan kualitas *deepfake* modern membuat artefak spasial semakin sulit dideteksi, sedangkan

spectral distortions tetap muncul karena berasal dari keterbatasan struktural model generatif. Keunggulan utama penggunaan *spectral distortions* dalam deteksi *deepfake* meliputi:

1. Generalisasi yang Lebih Baik

Metode berbasis frekuensi terbukti lebih *generalizable* untuk menghadapi *dataset* dan teknik manipulasi baru [11, 17, 28]. Artefak frekuensi tidak bergantung pada pola piksel spesifik *dataset* sehingga lebih stabil dalam pengujian *cross-dataset*.

2. Ketahanan terhadap Kompresi

Citra dan video yang mengalami kompresi sering kehilangan ciri spasial, namun pola spektral tertentu tetap dapat dipertahankan. *Tan et al.* menunjukkan bahwa detektor berbasis frekuensi memiliki performa lebih baik pada citra terkompresi [12].

3. *Robustness* terhadap GAN Generasi Baru

Walaupun GAN modern seperti PGGAN atau *Style Generative Adversarial Network* (StyleGAN) meningkatkan kualitas visual, struktur matematis *convolutional upsampling* tetap menghasilkan distorsi spektral. Hal ini membuat *spectral distortions* tetap relevan pada deteksi *deepfake* generasi terbaru.

4. Dukungan untuk Deteksi Berbasis Video

Pada video *deepfake*, distorsi spektral temporal merupakan ciri yang sulit dihilangkan oleh model generatif. Penelitian *Kim et al.* dan *Guera et al.* menunjukkan bahwa analisis frekuensi pada dimensi waktu efektif untuk mengungkap ciri tersebut [33, 35].

2.5.4 Pendekatan Modern yang Memanfaatkan *Spectral Distortions*

1. *Frequency-aware CNN*

Penelitian oleh *Qian et al.* serta *Tan et al.* mengusulkan pembelajaran langsung dari representasi frekuensi, yang dapat meningkatkan sensitivitas model terhadap distorsi spektral [11, 12].

2. Pendekatan Dual-Domain

SpecXNet dan FSBI memadukan domain spasial dan frekuensi untuk memperkaya representasi fitur. Pendekatan ini menunjukkan peningkatan signifikan dalam *robustness* detektor terhadap berbagai jenis serangan *deepfake* [13, 16].

3. *Masking* dan Interaksi Frekuensi–Spasial

Luo dan *Wang* mengembangkan mekanisme *frequency-domain masking* yang meningkatkan kemampuan model dalam menangkap interaksi antara domain spasial dan spektral [17].

4. Analisis Temporal *Frequency* untuk Video

Metode berbasis frekuensi temporal, seperti yang diperkenalkan *Kim et al.*, memberikan kemampuan untuk mendeteksi anomali antar *frame* yang tidak dapat ditangkap oleh CNN spasial [33].

2.5.5 Relevansi *Spectral Distortions* terhadap Penelitian Ini

Integrasi FFT dalam arsitektur deteksi yang memanfaatkan XceptionNet memberikan akses langsung pada pola spektral yang tidak dapat ditangkap oleh CNN spasial murni. Dengan memanfaatkan *frequency-aware features*, model tidak hanya mempelajari tekstur dan pola visual, tetapi juga distribusi energi frekuensi yang menjadi ciri khas *deepfake* [6].

Hal ini sejalan dengan tren penelitian modern seperti SpecXNet, FSBI, dan *Frequency-aware Deepfake Detection* yang menekankan pentingnya penggabungan domain spasial dan frekuensi untuk meningkatkan performa serta generalisasi [13, 16, 12].

2.6 Deep Learning

Deep learning merupakan cabang dari *machine learning* yang menggunakan jaringan saraf tiruan (*artificial neural networks*) banyak lapisan untuk mengekstraksi representasi fitur secara bertingkat (*feature hierarchy*) dari data mentah seperti citra, suara, atau teks. Pendekatan ini memungkinkan sistem belajar mengenali pola kompleks tanpa memerlukan rekayasa fitur manual (*manual feature engineering*) [36]. Teknologi *deep learning* inilah yang kemudian menjadi dasar dari *deepfake*. Istilah *deepfake* terbentuk dari gabungan *deep learning* dan *fake*, merujuk pada konten visual atau audio yang dimodifikasi dengan jaringan saraf dalam hingga menyerupai rekaman asli [19].

Salah satu bentuk implementasi paling umum dari *deep learning* adalah CNN, yang dirancang khusus untuk mengenali pola spasial pada data visual [6]. CNN menggunakan operasi konvolusi untuk mengekstraksi fitur spasial dari citra dan telah menjadi tulang punggung dalam banyak penelitian deteksi *deepfake*, seperti ResNet [5] dan XceptionNet [6] merupakan contoh arsitektur CNN yang banyak digunakan karena kemampuannya dalam mempelajari representasi visual yang efisien dan akurat untuk klasifikasi gambar atau deteksi manipulasi wajah.

2.6.1 Arsitektur Dasar *Deep Learning*

Deep learning tersusun atas jaringan saraf tiruan berlapis banyak (*multi-layer neural networks*) yang bekerja dengan memproses data melalui neuron-neuron buatan. Setiap neuron menerima *input*, mengalikan bobot (*weight*), menjumlahkan bias, lalu melewatkannya melalui fungsi aktivasi seperti *Rectified Linear Unit* (ReLU), *sigmoid*, atau *softmax*. Melalui pelatihan (*training*), jaringan akan melakukan pembaruan bobot (*weight update*) dengan algoritma optimasi seperti *Stochastic Gradient Descent* (SGD) atau *Adaptive Moment Estimation* (ADAM) sehingga model dapat mempelajari pola dari data [7].

Dalam konteks pengolahan citra, lapisan awal cenderung mempelajari fitur sederhana seperti tepi, garis, dan tekstur, sedangkan lapisan yang lebih dalam mempelajari pola kompleks seperti bentuk objek, ekspresi wajah, dan struktur wajah manusia. Hierarki pembelajaran fitur inilah yang membuat *deep learning* unggul pada tugas deteksi visual.

2.6.2 Peran *Deep Learning* dalam Pembuatan *Deepfake*

Deep learning merupakan fondasi utama dari teknologi *deepfake* modern. Arsitektur generatif seperti GAN dan *encoder-decoder* digunakan untuk menghasilkan manipulasi wajah, ekspresi, atau gerakan bibir secara realistis. GAN bekerja dengan dua jaringan generator dan *discriminator* yang dilatih secara *adversarial* sehingga generator terus meningkatkan kualitas citra sintetis hingga sangat mirip dengan citra asli [8, 9, 19].

Kemampuan GAN untuk menghasilkan detail wajah dan tekstur kulit yang halus menjadikan *deepfake* berkembang pesat. Namun proses generatif ini juga meninggalkan jejak matematis seperti artefak frekuensi tinggi dan distorsi spektral, yang kemudian menjadi dasar bagi berbagai teknik deteksi berbasis frekuensi [8, 9, 11].

2.6.3 Peran *Deep Learning* dalam Deteksi *Deepfake*

Selain digunakan untuk menghasilkan *deepfake*, *deep learning* juga digunakan untuk pendeteksiannya. Arsitektur CNN seperti ResNet, *MesoNet*, dan XceptionNet digunakan untuk mengekstraksi fitur spasial dan mendeteksi ketidak-wajaran visual pada citra wajah palsu [3, 4, 5, 6, 7].

CNN terbukti efektif mendeteksi anomali yang merupakan artefak umum dari proses manipulasi generatif seperti:

1. Ketidaksesuaian tekstur kulit.
2. Batas *blending* wajah yang tidak natural.

3. Distorsi wajah lokal akibat *face warping*.
4. Ketidakkonsistenan pencahayaan atau pose.

XceptionNet yang menggunakan DSC, telah diakui memiliki sensitivitas lebih baik terhadap pola tekstur dan detail halus, sehingga banyak digunakan dalam penelitian deteksi *deepfake* modern dan kompetisi *benchmark* seperti FFPP [3, 4, 6, 7].

Dalam penelitian ini, *deep learning* berperan sebagai ekstraktor fitur spasial melalui XceptionNet, dan dikombinasikan dengan analisis domain frekuensi (FFT) untuk menangkap artefak spektral yang tidak dapat dianalisis oleh metode spasial saja [13, 16, 17, 28].

2.6.4 Convolution Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan salah satu arsitektur utama dalam *deep learning* yang dirancang khusus untuk mengolah data berbentuk *grid*, seperti citra dua dimensi. CNN memanfaatkan operasi konvolusi untuk mengekstraksi fitur lokal secara hierarkis, sehingga mampu mengenali pola sederhana (tepi, tekstur) pada lapisan awal dan pola yang semakin kompleks (bagian wajah, objek) pada lapisan yang lebih dalam. Pendekatan ini berbeda dengan *multilayer perceptron* biasa karena CNN mengeksplorasi struktur spasial citra melalui *local receptive field* dan *weight sharing* sehingga jumlah parameter menjadi jauh lebih sedikit dan proses pelatihan lebih efisien [25, 36].

Secara umum, arsitektur CNN tersusun atas beberapa jenis lapisan utama, yaitu *convolutional layer*, *non-linear activation* (misalnya ReLU), *pooling layer*, dan *fully connected layer* di bagian akhir untuk melakukan klasifikasi. *Convolutional layer* melakukan operasi *filter* menggunakan *kernel* yang digeser secara spasial pada citra untuk menghasilkan *feature map*. *Weight sharing* membuat satu filter yang sama digunakan di seluruh posisi citra sehingga CNN bersifat *translationally equivariant* terhadap pola tertentu. *Pooling layer* (seperti *max-pooling*) kemudian mereduksi resolusi spasial *feature map* sambil mempertahankan informasi penting sehingga model menjadi lebih tahan terhadap variasi kecil posisi dan *noise*. Seluruh parameter jaringan dipelajari secara *end-to-end* menggunakan algoritma *backpropagation* dan optimasi gradien turun [25, 36].

Dalam bidang visi komputer, CNN telah menjadi *backbone* utama untuk berbagai tugas pengenalan citra dan wajah. Arsitektur-arsitektur seperti ResNet memanfaatkan *residual connection* untuk mengatasi degradasi performa pada jaringan sangat dalam dan berhasil mencapai performa *state-of-the-art* (SOTA) pada berbagai *benchmark* klasifikasi

citra [5]. Sementara itu, XceptionNet mengusulkan DSC untuk mengurangi kompleksitas komputasi tanpa mengorbankan akurasi, dan banyak diadopsi dalam berbagai studi deteksi *deepfake* sebagai *feature extractor* utama pada domain spasial [3, 6, 7]. Model ringan seperti *MesoNet* juga dirancang sebagai CNN berukuran kompak khusus untuk deteksi pemalsuan wajah pada video, dengan fokus pada artefak *mesoscopic* di area wajah [4].

Pada konteks deteksi *deepfake*, CNN digunakan untuk mempelajari artefak manipulasi yang sulit ditangkap oleh pengamatan manusia, seperti tekstur kulit yang tidak konsisten, *edge* yang tidak alami, atau pola *noise* yang ditinggalkan oleh proses generatif. Studi literatur menunjukkan bahwa sebagian besar detektor *deepfake* modern menggunakan CNN sebagai komponen utama, baik untuk citra statis maupun *frame* video, dengan performa yang sangat kompetitif pada berbagai *dataset* seperti FFPP [14, 15, 20, 7]. Dalam skenario ini, CNN biasanya menerima *input* berupa *patch* wajah yang telah dipotong, kemudian mengeluarkan probabilitas apakah wajah tersebut asli atau hasil manipulasi.

Perkembangan terbaru menunjukkan bahwa CNN tidak hanya efektif di domain spasial, tetapi juga dapat digabungkan dengan analisis frekuensi dengan tujuan meningkatkan kemampuan generalisasi. Pendekatan seperti *Thinking in Frequency* memanfaatkan CNN yang dioptimalkan untuk memanfaatkan petunjuk pada domain frekuensi (misalnya melalui representasi FFT), sehingga jaringan lebih peka terhadap distorsi tinggi frekuensi yang dihasilkan oleh proses generatif [11]. Model lain seperti SpecXNet, FSBI, serta serangkaian metode *frequency-aware deepfake detection* memodifikasi arsitektur CNN atau *input*-nya dengan memasukkan kanal-kanal frekuensi tambahan atau modul khusus untuk menonjolkan komponen frekuensi tinggi, sehingga artefak yang tidak terlihat di domain spasial dapat dimanfaatkan secara lebih efektif untuk deteksi [13, 16, 17, 28, 36].

Selain memodelkan informasi spasial, CNN juga sering dikombinasikan dengan pemodelan temporal untuk video. Salah satu pendekatan umum adalah menggunakan CNN untuk mengekstraksi fitur per-*frame*, kemudian meneruskannya ke model sekuensial seperti *Long Short-Term Memory* (LSTM) atau *Gated Recurrent Unit* (GRU) untuk menangkap dinamika temporal antar-*frame*, misalnya pada gerakan wajah dan sinkronisasi bibir. Pendekatan ini digunakan dalam berbagai karya deteksi *deepfake* berbasis *spatio-temporal*, termasuk yang mempelajari fitur pergerakan wajah dan artefak temporal pada video *deepfake* [34, 35]. Pendekatan *spatio-temporal* serupa juga diadopsi oleh *Haliassos et al.* yang memanfaatkan representasi CNN pada area mulut dan jaringan temporal untuk

mendeteksi ketidakwajaran pola gerak bibir pada video sintetis [37]. Di luar domain forensik media, CNN juga digunakan secara luas bersama model sekuensial untuk tugas-tugas lain seperti peramalan beban dan konsumsi energi, menunjukkan bahwa arsitektur ini bersifat serbaguna dan dapat menangani data spasial maupun *spatio-temporal* di berbagai domain aplikasi [38].

Untuk menstabilkan dan mempercepat pelatihan, lapisan *konvolusional* pada CNN modern umumnya diikuti oleh *batch normalization* (BatchNorm) [39], yang menormalisasi aktivasi setiap *mini-batch* lalu menskalakannya kembali melalui parameter yang dipelajari:

$$\hat{z}_i = \frac{z_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \cdot \gamma + \beta \quad (2.5)$$

di mana μ_B dan σ_B^2 adalah rata-rata dan varians *mini-batch*, γ dan β adalah parameter skala dan geser yang dipelajari (*learnable*), dan ϵ adalah konstanta kecil untuk mencegah pembagian dengan nol. BatchNorm mengurangi *internal covariate shift* sehingga jaringan dapat dilatih dengan *learning rate* lebih tinggi dan lebih tahan terhadap inisialisasi bobot.

Pada jaringan yang sangat dalam, koneksi *residual* (*residual connection*) yang diperkenalkan pada ResNet [5] digunakan untuk mengatasi degradasi gradien. Sebuah blok *residual* menjumlahkan keluaran transformasi *konvolusional* dengan jalur pintasan (*shortcut*) dari *input* sebelum operasi berikutnya:

$$Y = \text{MaxPool}_{2 \times 2}(\text{Conv}_{3 \times 3}(x) + \text{Shortcut}(x)) \quad (2.6)$$

di mana $\text{Shortcut}(x)$ berupa konvolusi 1×1 ketika jumlah kanal berubah, atau operasi identitas ketika dimensi sudah sesuai. Penjumlahan residual memungkinkan setiap blok mempelajari representasi tambahan di atas representasi yang sudah ada, sehingga pelatihan tetap stabil meskipun kedalaman jaringan bertambah.

Berikut komponen-komponen kunci DSC yang menjadi inti efisiensi XceptionNet, salah satu CNN paling banyak digunakan dalam deteksi *deepfake*:

1. *Depthwise Separable Convolution* (DSC)

Depthwise Separable Convolution (DSC) merupakan teknik konvolusi yang memisahkan operasi spasial dan operasi antarkanal menjadi dua tahap terpisah (*depthwise* dan *pointwise*). Pemisahan ini menekan jumlah parameter dan biaya komputasi secara signifikan tanpa mengurangi kapasitas representasi fitur, dan menjadi inti efisiensi arsitektur XceptionNet yang dibahas pada subbab 2.7. XceptionNet banyak digunakan dalam penelitian deteksi *deepfake* karena kemampuannya dalam mengekstraksi pola tekstur halus yang sering muncul sebagai artefak manipulasi wajah [4, 7].

2. Depthwise Convolution

Tahap pertama adalah *depthwise convolution*, yaitu proses konvolusi yang dilakukan secara independen pada setiap kanal masukan. Jika suatu citra masukan memiliki M kanal dan *kernel* konvolusi berukuran $K \times K$, maka operasi *depthwise* akan menerapkan satu *kernel* per kanal secara terpisah. Secara matematis, *depthwise convolution* dapat dituliskan sebagai:

$$D_m(x, y) = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} I_m(x+i, y+j) \cdot W_m(i, j) \quad (2.7)$$

dengan:

- I_m adalah kanal ke- m dari citra masukan.
- W_m adalah *kernel* konvolusi untuk kanal ke- m .
- D_m adalah hasil *depthwise convolution* pada kanal ke- m .

Pendekatan ini memungkinkan model mengekstraksi pola tekstur lokal secara lebih rinci, terutama pada bagian wajah yang sering mengandung artefak seperti *blending* dan *warping* pada *deepfake* [4].

3. Pointwise Convolution

Tahap kedua adalah *pointwise convolution*, yaitu penerapan *kernel* berukuran 1×1 untuk menggabungkan kembali seluruh kanal hasil *depthwise* menjadi satu set fitur baru. Tahap ini bertujuan menangkap korelasi antarkanal yang tidak diproses pada tahap sebelumnya. Secara matematis, *pointwise convolution* dapat dituliskan sebagai:

$$P(x, y) = \sum_{m=1}^M D_m(x, y) \cdot W_m^{pw} \quad (2.8)$$

dengan:

- D_m adalah keluaran *depthwise*.
- W_m^{pw} adalah bobot *kernel* 1×1 pada kanal ke- m .

Tahap ini membentuk representasi fitur yang lebih kaya karena menggabungkan informasi dari seluruh kanal, sekaligus tetap menjaga efisiensi komputasi.

4. Kompleksitas DSC

Perbedaan signifikan antara konvolusi biasa dan DSC terletak pada kompleksitas komputasi, sebagaimana diformulasikan oleh Chollet [6].

- Konvolusi Standar

Total operasi konvolusi standar:

$$C_{standard} = K^2 \times M \times N \quad (2.9)$$

dengan N adalah jumlah kanal keluaran.

b. *Depthwise Separable Convolution*

Operasi *depthwise* + *pointwise*:

$$C_{DSC} = K^2 \times M + (M \times N) \quad (2.10)$$

Sehingga efisiensi diperoleh karena:

$$C_{DSC} \ll C_{standard} \quad (2.11)$$

Efisiensi ini sangat relevan untuk penelitian deteksi *deepfake* yang menggunakan data dalam jumlah besar dan memerlukan proses pelatihan berulang-ulang.

5. Relevansi DSC pada Deteksi Deepfake

XceptionNet terbukti efektif untuk pendeteksian manipulasi wajah karena:

a. Sensitif terhadap pola tekstur lokal

Artefak *deepfake* umumnya muncul pada area transisi wajah seperti sekitar mata, hidung, dan garis rahang. *Depthwise convolution* memproses setiap kanal secara terpisah sehingga sesuai untuk menangkap detail lokal pada area tersebut.

b. Stabil pada variasi kualitas video

Variasi kompresi (contoh dari *dataset* FFPP) tidak mengganggu performa secara signifikan karena struktur konvolusi XceptionNet lebih tahan terhadap *compression noise* [7].

c. Kompatibel dengan analisis frekuensi

Dalam arsitektur *hybrid*, XceptionNet berperan sebagai cabang spasial yang memproses citra RGB, sedangkan representasi FFT ditangani cabang terpisah. Efisiensi XceptionNet membuatnya cocok sebagai *backbone* cabang spasial.

d. Performa tinggi pada *dataset* FFPP

XceptionNet dilaporkan sebagai detektor domain spasial terbaik pada FFPP [7] dan menjadi *backbone* populer pada berbagai metode deteksi lanjutan termasuk pendekatan *hybrid* [13]. Namun, performa tinggi ini umumnya bersifat *in-dataset*, yang artinya generalisasi ke *dataset* lain seperti CDF diketahui menurun, yang menjadi salah satu motivasi penelitian ini.

e. Kemampuan representasi spasial yang kuat

DSC menghasilkan fitur berlapis-lapis yang mampu memisahkan pola alami dan pola palsu pada wajah sintetis.

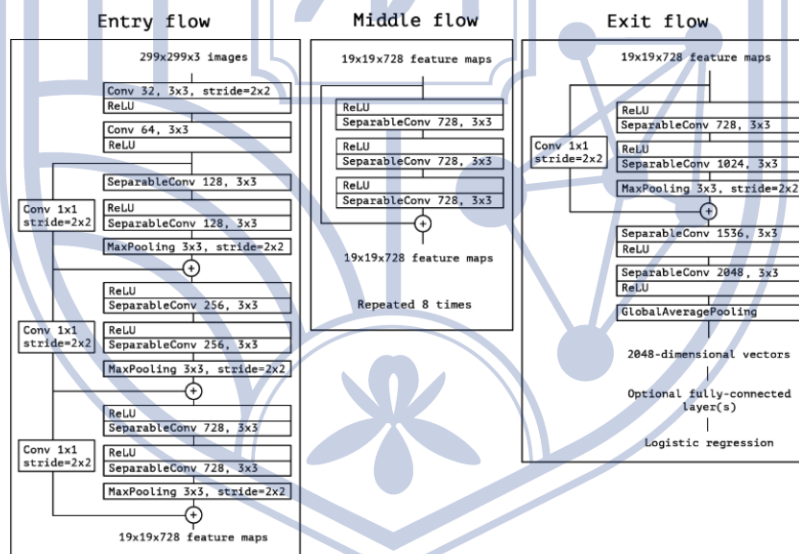
6. Peran XceptionNet sebagai Cabang Spasial dalam Arsitektur Hybrid

Dalam penelitian ini, XceptionNet memproses citra RGB sebagai cabang spasial, sedangkan FFT diproses sebagai cabang frekuensi terpisah (FreqCNN). Fitur kedua cabang digabung pada tahap *late fusion* (proyeksi *concatenation*, dan *SE gating*), bukan sebagai kanal tambahan (*early fusion*). SpecXNet [13] dan FSBI [16] juga memadukan spasial-frekuensi, melalui mekanisme yang berbeda (SpecXNet, melakukan pemisahan kanal spasial-spektral dan menambahkan atensi *Fourier* di dalam blok, dan FSBI, meningkatkan citra *self-blended* dengan frekuensi).

2.7 Extreme Inception Network (XceptionNet)

XceptionNet merupakan arsitektur CNN yang diperkenalkan oleh Chollet pada tahun 2017 sebagai pengembangan dari konsep *Inception Module*. Arsitektur ini dibangun berdasarkan premis bahwa korelasi spasial dan korelasi antar-kanal (*cross-channel*) dapat dipisahkan secara penuh melalui DSC, sehingga operasi konvolusi menjadi lebih efisien tanpa mengorbankan kapasitas representasi fitur [6].

2.7.1 Arsitektur XceptionNet



Gambar 2.3 Arsitektur XceptionNet [6]

Arsitektur XceptionNet terdiri dari 36 lapisan *convolutional* yang terorganisasi dalam tiga bagian utama:

1. Entry Flow

Bagian ini berfungsi sebagai tahap awal pemrosesan yang secara progresif mereduksi dimensi spasial citra masukan sambil meningkatkan jumlah kanal fitur. *Entryflow* terdiri dari dua lapisan konvolusi biasa di awal, diikuti oleh tiga blok *separable convolution*

dengan *residual connections*. Setiap blok menggunakan *max-pooling* untuk mengurangi resolusi spasial, sehingga citra masukan berukuran 299×299 secara bertahap dikompresi menjadi *feature map* berukuran lebih kecil dengan 728 kanal [6].

2. *Middle Flow*

Bagian ini merupakan inti dari arsitektur yang terdiri dari delapan blok *separable convolution* yang identik, masing-masing dengan tiga lapisan *separable convolution* dan *residual connection*. Seluruh blok mempertahankan dimensi spasial dan jumlah kanal (728) yang sama, sehingga fokus pemrosesan adalah pada penyempurnaan representasi fitur tanpa mengubah dimensi. Konfigurasi ini memungkinkan model mengekstraksi pola tekstur dan detail yang sangat halus, yang relevan untuk mendeteksi artefak manipulasi wajah [6].

3. *Exit Flow*

Bagian terakhir terdiri dari satu blok *separable convolution* dengan *max-pooling* untuk reduksi dimensi akhir, diikuti oleh dua lapisan *separable convolution* tambahan yang meningkatkan jumlah kanal menjadi 2.048. Setelah itu, *global average pooling* diterapkan untuk menghasilkan vektor fitur satu dimensi, yang kemudian diteruskan ke lapisan *fully connected* untuk klasifikasi [6].

Seluruh lapisan *separable convolution* dalam arsitektur ini menggunakan fungsi aktivasi ReLU dan BatchNorm, yang membantu menstabilkan proses pelatihan dan mempercepat konvergensi model.

2.7.2 *Transfer Learning* pada XceptionNet

Dalam penelitian deteksi *deepfake*, XceptionNet umumnya digunakan dengan pendekatan *transfer learning*, yaitu memanfaatkan bobot model yang telah dilatih sebelumnya pada *dataset* berskala besar seperti ImageNet. Pendekatan ini efektif karena fitur-fitur visual dasar yang dipelajari dari ImageNet, seperti deteksi tepi, tekstur, dan pola geometris, juga relevan untuk mengenali artefak manipulasi wajah pada *deepfake* [7].

Proses *transfer learning* dilakukan dengan menginisialisasi bobot XceptionNet dari model *pretrained* ImageNet, kemudian melakukan *fine-tuning* pada *dataset* deteksi *deepfake*. Strategi ini memungkinkan model untuk beradaptasi dengan pola-pola spesifik artefak manipulasi tanpa memerlukan pelatihan dari awal, sehingga lebih efisien dalam hal waktu pelatihan dan kebutuhan data [6, 7].

Normalisasi *input* pada XceptionNet menggunakan statistik ImageNet ($mean = [0.485, 0.456, 0.406]$, $std = [0.229, 0.224, 0.225]$) untuk memastikan distribusi data masukan konsisten dengan distribusi yang digunakan selama *pretraining*. Hal ini penting agar bobot *pretrained* dapat berfungsi secara optimal pada tahap awal *fine-tuning*.

2.7.3 Keunggulan XceptionNet dalam Deteksi Deepfake

XceptionNet memiliki sejumlah keunggulan yang menjadikannya arsitektur yang sangat sesuai untuk deteksi *deepfake*:

1. Sensitivitas terhadap Pola Tekstur Lokal

DSC memungkinkan XceptionNet menangkap pola tekstur halus secara independen pada setiap kanal, sehingga sangat efektif dalam mendeteksi artefak lokal seperti batas *blending*, distorsi tekstur kulit, dan ketidakwajaran pada area transisi wajah [4, 7].

2. Efisiensi Parameter

Dibandingkan dengan arsitektur konvolusi standar, XceptionNet memiliki jumlah parameter yang lebih sedikit untuk kapasitas representasi yang setara, karena pemisahan konvolusi spasial dan antar-kanal mengurangi parameter yang mubazir secara signifikan [6].

3. Performa Tertinggi pada Benchmark FFPP

Pada *benchmark* FFPP tanpa kompresi, XceptionNet mencapai akurasi deteksi 99,26% dan melampaui ResNet-50 maupun MesoNet. Namun, akurasi turun ke 81,00% pada kompresi berat (c40), menunjukkan ketergantungan model pada artefak visual berfrekuensi tinggi yang rusak oleh kompresi [3, 7].

4. Stabilitas pada Variasi Kompresi

XceptionNet terbukti mempertahankan performa yang stabil pada video dengan berbagai tingkat kompresi, termasuk kompresi tinggi yang umum ditemui pada media sosial. Hal ini penting karena sebagian besar *deepfake* di dunia nyata mengalami kompresi berulang [7].

5. Standar dalam Penelitian Deteksi *Deepfake*

XceptionNet telah diadopsi secara luas sebagai *baseline* dan *backbone* utama dalam berbagai kompetisi dan publikasi deteksi *deepfake*, termasuk SpecXNet [13], sehingga memudahkan perbandingan hasil dengan penelitian sebelumnya [7].

2.8 *Squeeze-and-Excitation Networks*

Dalam arsitektur *deep learning* modern, tidak semua fitur yang diekstraksi oleh lapisan konvolusional memiliki tingkat kepentingan yang sama untuk tugas klasifikasi tertentu. Beberapa kanal fitur mungkin mengandung informasi yang sangat diskriminatif, sementara kanal lainnya mengandung *noise* atau informasi yang tidak relevan. *Squeeze-and-Excitation Networks* (SE-Net) yang diperkenalkan oleh Hu et al. [39] mengatasi permasalahan ini melalui mekanisme *channel attention* yang memungkinkan jaringan untuk secara adaptif mempelajari bobot kepentingan setiap kanal fitur.

2.8.1 Konsep *Channel Attention*

Channel attention merupakan mekanisme yang memberikan bobot berbeda pada setiap kanal fitur berdasarkan konten *input*. Berbeda dengan konvolusi standar yang memperlakukan semua kanal secara seragam, *channel attention* memungkinkan jaringan untuk secara selektif menekan (*suppress*) kanal yang tidak informatif dan memperkuat (*enhance*) kanal yang diskriminatif. Pendekatan ini terinspirasi dari mekanisme atensi selektif pada sistem visual manusia yang secara alami memfokuskan perhatian pada informasi yang paling relevan [39].

2.8.2 Arsitektur SE Block

SE block terdiri dari tiga operasi berurutan yang diterapkan pada peta fitur keluaran lapisan *konvolusional*:

1. ***Squeeze (Kompresi Global)***, informasi spasial dari setiap kanal fitur dikompresi menjadi satu nilai skalar melalui operasi *global average pooling*. Untuk peta fitur $U \in \mathbb{R}^{C \times H \times W}$ dengan C kanal, operasi *squeeze* menghasilkan vektor deskriptor $z \in \mathbb{R}^C$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (2.12)$$

Setiap elemen z_c merepresentasikan respons global dari kanal ke- c , sehingga menangkap distribusi informasi pada kanal tersebut secara keseluruhan [39].

2. **Excitation (Pemodelan Hubungan Antar Kanal)**, vektor deskriptor z diproses melalui dua lapisan *fully connected* (FC) yang membentuk arsitektur *bottleneck* untuk mempelajari hubungan non-linear antar kanal:

$$s = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \quad (2.13)$$

di mana $W_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ adalah matriks bobot reduksi, $W_2 \in \mathbb{R}^{c \times \frac{c}{r}}$ adalah matriks bobot ekspansi, r adalah rasio reduksi (*reduction ratio*), dan σ adalah fungsi *sigmoid* yang menghasilkan bobot gerbang pada rentang $[0, 1]$. Rasio reduksi r mengendalikan kapasitas dan biaya komputasi, nilai r yang lebih besar mengurangi jumlah parameter dalam *bottleneck* [39].

3. **Scale (Pengalihan Elemen)**, bobot gerbang s yang telah dipelajari dikalikan secara elemen-per-elemen (*element-wise*) dengan peta fitur asli:

$$\tilde{u}_c = s_c \cdot u_c \quad (2.14)$$

di mana s_c adalah bobot skalar untuk kanal ke- c dan u_c adalah peta fitur kanal ke- c . Kanal dengan bobot s_c mendekati 1 dipertahankan, sedangkan kanal dengan bobot mendekati 0 ditekan [39].

2.8.3 Relevansi SE Block untuk Fusi Multi-Domain

Dalam konteks deteksi *deepfake* berbasis pendekatan *hybrid*, SE block memiliki peran penting sebagai mekanisme fusi adaptif. Ketika fitur dari domain spasial (XceptionNet) dan domain frekuensi (FreqCNN) digabungkan melalui *concatenate*, vektor fusi yang dihasilkan mengandung dimensi-dimensi fitur dari kedua domain yang mungkin memiliki tingkat relevansi berbeda tergantung pada jenis manipulasi yang dihadapi.

Mekanisme SE memungkinkan model untuk:

1. Menyesuaikan kontribusi relatif kedua domain secara adaptif berdasarkan *input* misalnya, memberikan bobot lebih tinggi pada fitur frekuensi ketika artefak spasial sulit dideteksi, atau sebaliknya.
2. Menekan fitur yang mubazir atau mengandung *noise* dari salah satu cabang, sehingga proses klasifikasi tidak terganggu oleh informasi yang tidak relevan.
3. Mengoptimalkan interaksi lintas domain, SE block dapat mempelajari korelasi antara fitur spasial dan frekuensi yang tidak dapat ditangkap oleh *concatenation* sederhana.

Dibandingkan dengan mekanisme fusi sederhana seperti *concatenation* atau penjumlahan (*addition*), penggunaan SE block sebagai gerbang adaptif telah terbukti

meningkatkan performa pada arsitektur *multi-domain* karena memungkinkan model untuk secara dinamis menyesuaikan strategi fusi berdasarkan karakteristik *input* [39].

Dalam arsitektur fusi *multi-domain*, fitur dari cabang yang berbeda kerap memiliki dimensi yang tidak seimbang sehingga salah satu cabang dapat mendominasi. Untuk menyeimbangkannya, setiap vektor fitur diproyeksikan ke dimensi yang sama melalui lapisan linear yang diikuti BatchNorm dan aktivasi ReLU:

$$h = \text{ReLU}(\text{BN}(W \cdot f + b)) \quad (2.15)$$

di mana f adalah vektor fitur masukan, W dan b adalah bobot dan bias proyeksi, dan h adalah vektor terproyeksi berdimensi seragam. Setelah proyeksi, fitur dari kedua cabang dapat digabungkan (*concatenation*) secara seimbang sebelum diproses oleh mekanisme *channel attention* seperti *SE gating*.

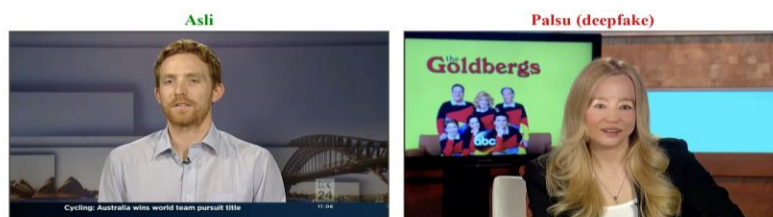
2.9 FaceForensics++ (FFPP)

Dataset FaceForensics++ (FFPP) merupakan salah satu *dataset* standar yang paling banyak digunakan dalam penelitian deteksi *deepfake* karena menyediakan kumpulan video manipulasi wajah dengan kualitas, metode pemalsuan, dan tingkat kompresi yang beragam. *Dataset* ini pertama kali diperkenalkan oleh Rössler et al. [7] sebagai pengembangan dari *dataset* sebelumnya, FaceForensics, dengan meningkatkan variasi teknik pemalsuan serta memperluas jumlah sampel agar lebih representatif terhadap karakteristik *deepfake* modern.

Dalam konteks penelitian ini, FaceForensics++ dipilih karena menyediakan kondisi pengujian yang realistis serta mencakup berbagai teknik generatif yang umum digunakan pada sistem pemalsuan wajah berbasis GAN maupun *encoder-decoder*.

2.9.1 Struktur dan Komposisi Dataset

FFPP terdiri dari dua komponen utama, yaitu *original videos* dan *manipulated videos*. Kumpulan video asli diambil dari rekaman wawancara publik yang menampilkan variasi ekspresi, sudut pandang, dan kondisi pencahayaan. Video ini kemudian dimanipulasi menggunakan empat teknik utama:



Gambar 2.4 Contoh Frame Real dan Palsu dari Dataset FaceForensics++

1. *Deepfakes*, dengan menggunakan model *encoder-decoder* untuk menukar wajah sumber dengan wajah target.
2. *Face2Face*, mengubah ekspresi wajah target berdasarkan gerakan wajah sumber menggunakan teknik *facial reenactment*.
3. *FaceSwap*, menggantikan wajah target menggunakan metode penyelarasan geometris dan *blending* manual.
4. *NeuralTextures*, menggunakan *neural rendering* untuk menghasilkan ekspresi wajah sintesis berbasis peta tekstur.

Variasi ini penting karena setiap teknik menghasilkan jenis artefak yang berbeda, mulai dari kesalahan *warping*, ketidakselarasan detail, ketidakkonsistenan tekstur, hingga pola spektral yang tidak wajar akibat proses generatif [7].

Dataset ini juga disediakan dalam tiga tingkat kompresi:

1. RAW (*lossless*), tanpa kompresi, tekstur paling utuh.
2. HQ (*high-quality*), kompresi H.264 dengan *Constant Rate Factor* (CRF) 23 yang tergolong kompresi ringan.
3. LQ (*low-quality*), kompresi H.264 dengan CRF 40 yang tergolong kompresi berat dan meniru kondisi unggahan media sosial.

Pembagian kualitas ini membuat *dataset* sangat relevan untuk menguji *robustness* model terhadap *noise* kompresi dan degradasi kualitas yang umum ditemui pada *deepfake* di dunia nyata. Penelitian ini menggunakan versi HQ (c23) sesuai standar *benchmark* FFPP.

2.9.2 Karakteristik Artefak pada FFPP

Setiap teknik manipulasi pada FFPP menghasilkan pola artefak yang berbeda, baik pada domain spasial maupun domain frekuensi. Hal ini membuat *dataset* ini sangat berguna untuk penelitian *hybrid* seperti XceptionNet-FFT.

1. Artefak Spasial

Artefak utama yang sering muncul meliputi:

- a. Ketidaksesuaian warna kulit antara wajah sintesis dan latar.
- b. Tepi wajah yang tidak rapi akibat *blending* yang buruk.
- c. Distorsi geometri pada bagian mata dan mulut.
- d. Ketidakkonsistenan pencahayaan dan bayangan akibat *face warping*.
- e. *Blurring* lokal pada area rekonstruksi.

Penelitian sebelumnya menunjukkan bahwa artefak ini paling jelas terlihat pada area transisi wajah, sehingga algoritma deteksi perlu fokus pada *patch* wajah, bukan keseluruhan citra [4, 7].

2. Artefak Frekuensi

Artefak pada domain frekuensi yang muncul karena proses generatif tidak mampu mempertahankan distribusi spektral alami (*natural image statistics*). Pengukuran energi spektral *azimuth* pada citra GAN menunjukkan kelebihan energi pada *band* frekuensi tinggi dibanding citra natural, beserta penguatan harmonik akibat *upsampling* yang tidak ideal [8].

Karena mencakup empat teknik manipulasi, FFPP menghadirkan artefak spasial maupun frekuensi sebagaimana dibahas pada subbab 2.2.3.

1. *Deepfakes* menyisakan batas *blending* dan ketidaksesuaian warna pada wajah hasil *face-swap*.
 2. *Face2Face* meninggalkan *warping* berfrekuensi rendah dari *reenactment* ekspresi.
 3. *NeuralTextures* memperlihatkan tekstur frekuensi tinggi yang berulang dari *neural rendering*.
 4. *FaceSwap* sering menunjukkan *spectral drop-off* yang tidak konsisten pada tepi wajah.
- Keberagaman artefak lintas-teknik ini menjadikan FFPP ideal untuk model *hybrid* yang memadukan domain spasial dan frekuensi, sebagaimana dilakukan pada penelitian ini.

2.9.3 Alasan Pemilihan FFPP untuk Penelitian Ini

Ada beberapa alasan utama mengapa *dataset* FFPP menjadi pilihan tepat:

1. Representasi Teknik *Deepfake* yang Beragam
FFPP mencakup metode berbasis grafika komputer (*Face2Face*, *FaceSwap*) dan berbasis pembelajaran (*Deepfakes*, *NeuralTextures*) [7], sehingga model dapat mempelajari berbagai pola pemalsuan, bukan hanya satu jenis teknik. Hal ini penting agar model memiliki kemampuan generalisasi yang baik [14, 15].
2. Mendukung Evaluasi *Cross-Compression*
Penelitian deteksi *deepfake* sering gagal pada video berkualitas rendah [7]. Karena *dataset* ini menyediakan berbagai tingkat kompresi, maka model dapat diuji apakah tetap akurat pada kondisi *real-world* seperti yang ditemukan di media sosial.
3. Kompatibel dengan Pendekatan Spasial-Frekuensi

Dataset ini memiliki kualitas detail spasial dan spektral yang cukup baik, sehingga transformasi FFT dapat menangkap perbedaan distribusi frekuensi dengan jelas. Penelitian semacam SpecXNet dan FSBI juga menggunakan *dataset* ini untuk validasi [13, 16].

4. Menjadi *Benchmark* Internasional

FFPP menjadi salah satu *benchmark* yang paling sering digunakan sejak publikasinya pada 2019 [7, 15], sehingga hasil penelitian menjadi mudah dibandingkan dengan pekerjaan sebelumnya.

5. Proporsional dan Seimbang

Jumlah sampel asli dan palsu yang seimbang membantu mengurangi bias pelatihan.

2.9.4 Peran FFPP dalam Sistem *Hybrid XceptionNet-FFT*

Dalam penelitian ini, *dataset* FFPP digunakan untuk dua fungsi:

1. Pelatihan Model Spasial (XceptionNet)

XceptionNet akan menangkap pola spasial, tekstur, dan artefak lokal pada wajah.

2. Pelatihan Model Frekuensi (FFT)

FFT diterapkan pada citra hasil preprocessing FFPP, sehingga saluran frekuensi dapat menangkap:

- Ketidakstabilan spektral.
- Pola frekuensi tinggi yang berlebihan.
- Anomali distribusi amplitudo.

Fitur spasial dan frekuensi kemudian digabungkan pada tahap *late fusion* di tingkat fitur sehingga model mampu mendeteksi manipulasi baik secara visual maupun secara statistik spektral. FFPP menjadi *dataset* utama untuk mengevaluasi performa sistem *hybrid* pada penelitian ini.

2.10 *Celebrity DeepFake* (Celeb-DF)

Celebrity DeepFake (Celeb-DF atau CDF) merupakan *dataset* deteksi *deepfake* berskala besar yang dikembangkan untuk mengatasi keterbatasan *dataset* generasi sebelumnya dalam hal kualitas visual manipulasi. *Dataset* ini terdiri dari video selebriti yang dimanipulasi menggunakan teknik *face swapping* berbasis *deep learning* yang menghasilkan kualitas manipulasi yang jauh lebih tinggi dibandingkan *dataset* FFPP [18].

2.10.1 Komposisi dan Karakteristik *Dataset*

Celeb-DF terdiri dari dua versi, yaitu Celeb-DF (v1) dan Celeb-DF (v2), dengan versi kedua yang lebih luas digunakan dalam penelitian. Celeb-DF (v2) mencakup 590 video asli yang diambil dari wawancara selebriti di YouTube dan 5.639 video *deepfake* yang dihasilkan menggunakan model *encoder-decoder* yang telah disempurnakan. Video-video dalam *dataset* ini memiliki variasi yang tinggi dalam hal kondisi pencahayaan, resolusi, latar belakang, dan ekspresi wajah, sehingga lebih merepresentasikan kondisi *deepfake* di dunia nyata [18].



Gambar 2.5 Contoh Frame Real dan Palsu dari Dataset Celeb-DF v2

Dibandingkan dengan FFPP, Celeb-DF memiliki beberapa perbedaan penting:

1. Kualitas manipulasi yang lebih tinggi, sehingga artefak visual seperti batas *blending* dan distorsi wajah lebih sulit diamati secara kasat mata.
2. Variasi subjek yang lebih beragam, mencakup berbagai etnis, usia, dan jenis kelamin.
3. Kondisi perekaman yang lebih bervariasi, termasuk pencahayaan alami dan buatan, serta resolusi yang berbeda-beda.

Perbedaan-perbedaan ini menjadikan Celeb-DF sebagai *dataset* uji yang lebih menantang dibanding FFPP, sekaligus tolak ukur penting untuk menilai generalisasi model, sebagaimana diuraikan pada bagian berikut.

2.10.2 Peran Celeb-DF dalam Evaluasi *Cross-Dataset*

Dalam penelitian ini, Celeb-DF digunakan sebagai *dataset* sekunder untuk mengevaluasi kemampuan *cross-dataset generalization* dari model *hybrid* yang diusulkan. Model yang dilatih pada FFPP kemudian diuji pada Celeb-DF untuk mengukur sejauh mana model mampu mendeteksi *deepfake* dari sumber dan teknik manipulasi yang berbeda. Pengujian ini penting karena menunjukkan apakah model telah mempelajari fitur-fitur

manipulasi yang bersifat universal, bukan hanya artefak spesifik dari satu *dataset* tertentu [14, 39].

Celeb-DF menjadi *benchmark* yang menantang karena kualitas manipulasinya yang tinggi membuat artefak spasial menjadi sangat minim. Karena itu, kemampuan model memanfaatkan informasi domain frekuensi menjadi sangat menentukan ketika diuji pada *dataset* ini, sebab artefak frekuensi yang bersifat algoritmik cenderung lebih stabil meskipun kualitas visual manipulasi meningkat [8].

2.11 Analisis Citra

Analisis citra merupakan proses untuk memperoleh informasi dari sebuah citra digital melalui serangkaian operasi matematis, statistik, maupun transformasi domain. Analisis citra bertujuan mengekstraksi karakteristik penting seperti bentuk, tekstur, *edge*, *noise*, maupun struktur pola tertentu. Citra digital sendiri direpresentasikan sebagai fungsi dua dimensi $f(x,y)$ yang memetakan koordinat piksel ke intensitas cahaya, sehingga analisis citra dapat dilakukan baik dalam domain spasial (langsung pada piksel) maupun domain frekuensi (melalui transformasi seperti FFT) [25].

Metode analisis citra berbasis domain spasial berfokus pada karakteristik visual yang tampak langsung, seperti konsistensi pencahayaan, *texture descriptor*, dan informasi warna pada citra digital [25]. Pada deteksi *deepfake*, pendekatan spasial digunakan untuk mengidentifikasi anomali seperti ketidaksesuaian *blending* wajah, *warping artifacts*, *detail* kulit yang tidak konsisten, atau pola *blur* tidak alami [28]. Namun, penelitian menunjukkan bahwa artefak *deepfake* modern semakin sulit dideteksi hanya dengan pendekatan spasial, karena model generatif seperti GAN telah menghasilkan citra dengan detail visual yang sangat realistis [40].

Untuk mengatasi keterbatasan tersebut, analisis citra diperluas ke domain frekuensi yang mampu mengungkap anomali spektral yang tidak tampak langsung di domain spasial. Model GAN meninggalkan pola *high-frequency artifacts* yang konsisten lintas arsitektur, terutama akibat operasi *upsampling* dan ketidakseimbangan *kernel* konvolusi [8]. Distribusi frekuensi yang jauh berbeda dari citra asli menjadi dasar metode deteksi domain frekuensi [9]. Oleh sebab itu, kombinasi analisis spasial dan frekuensi memberikan pendekatan yang lebih menyeluruh dalam mendeteksi manipulasi pada citra *deepfake*.

2.12 Analisis Video

Analisis video merupakan perluasan dari analisis citra yang mempertimbangkan tidak hanya informasi spasial dari setiap *frame*, tetapi juga hubungan temporal antar-*frame*. Video digital dapat direpresentasikan sebagai fungsi tiga dimensi $f(x,y,t)$, di mana t merupakan urutan waktu [25]. Karena itu, analisis video melibatkan pemrosesan *frame-level* dan dinamika pergerakan (temporal *cues*) seperti perubahan ekspresi, kontinuitas gerakan, serta konsistensi pola wajah dari satu *frame* ke *frame* berikutnya [41].

Dalam konteks deteksi *deepfake*, manipulasi umumnya dilakukan pada setiap *frame* secara terpisah, sehingga artefak manipulasi dapat dideteksi pada tingkat *frame* individual. Studi oleh Sabir *et al.* menunjukkan bahwa *deepfake* sering menampilkan pola-pola tidak stabil seperti *jittering*, *flickering*, dan perubahan wajah yang tidak mulus antar-*frame* [41]. Artefak-arterfak ini dapat diidentifikasi melalui analisis *frame-by-frame* yang mengekstraksi fitur spasial dan frekuensi dari setiap *frame* secara independen.

Dalam penelitian ini, pendekatan yang digunakan adalah analisis pada tingkat *frame* individual (*frame-level analysis*), di mana setiap *frame* video diekstraksi dan diproses secara terpisah menggunakan CNN untuk fitur spasial dan FFT untuk fitur frekuensi. Pendekatan *frame-level* dipilih karena artefak manipulasi yang muncul akibat proses generatif, baik pada domain spasial maupun domain frekuensi, dapat dideteksi pada setiap *frame* tunggal tanpa memerlukan informasi sekuensial antar-*frame* [3, 7]. Selain itu, pendekatan ini lebih efisien secara komputasi dan memungkinkan integrasi yang lebih langsung dengan analisis domain frekuensi berbasis FFT.

2.13 Preprocessing

Preprocessing merupakan tahap fundamental dalam sistem deteksi *deepfake* karena kualitas data masukan sangat memengaruhi kemampuan model dalam mengekstraksi fitur baik pada domain spasial melalui XceptionNet maupun pada domain frekuensi melalui FFT. Pada penelitian deteksi *deepfake* modern, *preprocessing* tidak hanya berfungsi sebagai proses penyiapan data, tetapi juga sebagai cara untuk menghilangkan *noise* visual, menjaga konsistensi bentuk wajah, serta menstabilkan distribusi frekuensi sebelum dilakukan transformasi [3, 4, 8, 11, 7].

Tahap *preprocessing* yang baik sangat diperlukan karena *deepfake* sering menampilkan variasi kualitas video, tingkat kompresi berbeda, posisi wajah yang tidak seragam, serta pola artefak yang tidak konsisten antar-*frame*. Penelitian terdahulu

menunjukkan bahwa ketidakseragaman ini berdampak langsung pada performa model deteksi, terutama yang memanfaatkan sinyal frekuensi tinggi yang sensitif terhadap rotasi, translasi, dan perubahan intensitas [8, 9, 16].

Oleh karena itu, *preprocessing* pada penelitian ini dirancang untuk memastikan data masukan memiliki format yang seragam, representasi wajah yang stabil, dan struktur frekuensi yang tidak terdistorsi. Alur preprocessing terdiri atas tahap bersama yang kemudian bercabang menjadi dua representasi. Tahap bersama meliputi ekstraksi *frame*, deteksi wajah, dan *cropping* area wajah. Setelah itu, setiap *frame* diolah menjadi dua representasi paralel, yaitu:

1. **Cabang Spasial**, berupa citra RGB yang di-*resize* ke 224×224 piksel lalu dinormalisasi dengan statistik ImageNet
2. **Cabang Frekuensi**, berupa citra yang dikonversi ke skala keabuan (*grayscale*), di-*resize*, ditransformasikan dengan FFT (disertai *high-pass filtering* dan *log-magnitude*), lalu dinormalisasi dengan *z-score*.

Kedua representasi diproses oleh cabang jaringan yang terpisah dan baru digabungkan di dalam model pada tahap *late fusion*, bukan sebagai kanal tambahan pada masukan XceptionNet.

2.13.1 Tahapan dan Alur Preprocessing

1. Ekstraksi Frame Video

Video *deepfake* direpresentasikan sebagai rangkaian *frame*, dan setiap *frame* berpotensi mengandung tingkat manipulasi berbeda. Pendekatan *frame-level* digunakan karena artefak manipulasi tidak selalu muncul konsisten sepanjang video [3]. Afchar menunjukkan bahwa beberapa *frame* menampilkan distorsi tekstur lebih jelas sehingga pemrosesan per-*frame* membantu model menangkap sinyal manipulasi penting [4]. Selain itu, pemrosesan per-*frame* memungkinkan analisis lebih mendetail terhadap variasi temporal, seperti *flickering* dan inkonsistensi ekspresi, yang merupakan ciri khas dari banyak teknik *deepfake* [18, 35].

2. Deteksi wajah dan cropping

Bagian wajah merupakan target utama manipulasi *deepfake*. Visualisasi aktivasi pada MesoNet menunjukkan jaringan paling mengandalkan area mata, hidung, dan mulut, yang merupakan bagian paling detail pada wajah asli, karena pada wajah *deepfake* area

tersebut justru tampak buram dan kehilangan detail. Rössler melalui FaceForensics++ menegaskan bahwa latar belakang justru menjadi noise bila tidak dipotong [7].

Dengan membatasi analisis hanya pada area wajah, jumlah fitur yang tidak relevan dapat diminimalkan, sehingga model memiliki kapasitas lebih besar untuk mempelajari pola-pola manipulasi.

3. *Resize*

Citra kemudian di-*resize* menjadi 224×224 piksel agar sesuai dengan input XceptionNet. Penyeragaman ukuran penting untuk menjaga konsistensi pola tekstur, karena XceptionNet sensitif terhadap variasi struktur local [6].

Ukuran 224×224 juga dipilih karena merupakan ukuran standar *feature extractor* pada banyak arsitektur visi komputer, sehingga mempermudah integrasi dengan model pra-latih (*pretrained model*) apabila diperlukan.

4. Normalisasi Piksel

Normalisasi piksel penting karena model CNN sangat sensitif terhadap variasi intensitas cahaya, sebagaimana ditunjukkan oleh Afchar *et al.* [4]. Pada cabang spasial, citra RGB dinormalisasi menggunakan statistik ImageNet ($mean = [0,485, 0,456, 0,406]$, $std = [0,229, 0,224, 0,225]$) agar konsisten dengan distribusi data *pretraining*. Pada cabang frekuensi, peta FFT *log-magnitude* dinormalisasi menggunakan *z-score* berdasarkan statistik per-dataset. Normalisasi meningkatkan stabilitas pelatihan dan mempercepat konvergensi model sekaligus mengurangi risiko *gradient explosion* maupun *gradient vanishing*.

5. Konversi *Grayscale* dan Transformasi FFT (cabang frekuensi)

Untuk cabang frekuensi, citra wajah lebih dahulu dikonversi ke skala keabuan (*grayscale*) karena analisis frekuensi cukup dilakukan pada informasi luminansi, lalu di-*resize* ke 224×224 pixel dan ditransformasikan ke domain frekuensi menggunakan FFT dua dimensi. Peta magnitude yang dihasilkan dilewatkan melalui *Gaussian high-pass filter* untuk menekan dominasi komponen frekuensi rendah, kemudian dikompresi rentang dinamisnya dengan *log-magnitude*. Rincian matematis transformasi ini diuraikan pada subbab 2.13.2.

Tabel 2.4 Tahapan Preprocessing

Tahap	Operasi	Cabang	Tujuan Utama	Output
1	Ekstraksi <i>frame</i>	Bersama	Mengubah video menjadi deretan <i>frame</i>	Citra per- <i>frame</i>

Tahap	Operasi	Cabang	Tujuan Utama	Output
2	Deteksi wajah (MTCNN)	Bersama	Menemukan lokasi (<i>bounding box</i>) wajah	Koordinat wajah
3	<i>Cropping</i> (margin 0,3)	Bersama	Memotong area wajah dari latar	<i>Patch</i> wajah
4	<i>Resize</i> 224×224	Spasial (RGB)	Menyeragamkan resolusi	Citra RGB 224×224
5	Normalisasi ImageNet	Spasial (RGB)	Menstabilkan intensitas dan kontras	Citra RGB ternormalisasi
6	Konversi <i>grayscale</i>	Frekuensi	Mengambil informasi luminansi	Citra 1-kanal
7	<i>Resize</i> 224×224	Frekuensi	Menyeragamkan resolusi	Citra 1-kanal 224×224
8	Transformasi FFT (<i>high-pass, log</i>)	Frekuensi	Mengubah ke domain frekuensi	Peta <i>log-magnitude</i>
9	Normalisasi <i>z-score</i>	Frekuensi	Menyeragamkan skala spektral	Peta FFT ternormalisasi

Tahap 1–3 bersifat bersama untuk kedua cabang. Tahap 4–5 menghasilkan representasi spasial (RGB) yang menjadi masukan XceptionNet, sedangkan tahap 6–9 menghasilkan representasi frekuensi (peta FFT *log-magnitude*) yang menjadi masukan FreqCNN. Kedua representasi diproses oleh cabang yang terpisah dan digabungkan di dalam model pada tahap *late fusion*, bukan pada tahap *preprocessing* dan bukan sebagai kanal tambahan.

2.13.2 Fast Fourier Transform (FFT)

Fast Fourier Transform (FFT) merupakan algoritma efisien untuk menghitung DFT, yaitu proses yang mengubah citra dari domain spasial menjadi domain frekuensi. Transformasi ini memungkinkan analisis distribusi energi frekuensi pada citra sehingga pola-pola artefak halus yang tidak terlihat pada domain spasial dapat teridentifikasi. Pada penelitian deteksi *deepfake*, FFT menjadi komponen penting karena manipulasi berbasis GAN cenderung menimbulkan ketidakwajaran sinyal frekuensi, terutama pada rentang *high-frequency* akibat proses *upsampling*, *konvolusi*, dan operasi *blending* [8, 11, 16].

Dalam konteks deteksi, informasi frekuensi menjadi sangat berharga karena generator GAN secara matematis tidak mampu mempertahankan *natural image statistics* sehingga meninggalkan *spectral fingerprint* yang khas dan konsisten. Sebelum transformasi *Fourier*

diterapkan, citra RGB terlebih dahulu dikonversi menjadi citra satu kanal (*grayscale*), karena analisis frekuensi cukup dilakukan pada informasi luminansi yang merepresentasikan distribusi intensitas spasial. Konversi menggunakan standar ITU-R BT.601:

$$Y = 0,299 \cdot R + 0,587 \cdot G + 0,114 \cdot B \quad (2.16)$$

di mana Y adalah nilai luminansi, sedangkan R, G, dan B adalah nilai kanal warna merah, hijau, dan biru. Bobot yang berbeda pada setiap kanal mencerminkan sensitivitas mata manusia terhadap masing-masing warna.

Citra *grayscale* tersebut kemudian ditransformasi ke domain frekuensi menggunakan DFT dua dimensi sebagaimana telah didefinisikan pada persamaan 2.1, dengan $f(x,y)$ sebagai intensitas piksel dan $F(u,v)$ sebagai representasi frekuensinya. Besarnya energi frekuensi (*magnitude spectrum*) dihitung menggunakan persamaan 2.3.

Pada citra wajah, energi frekuensi sangat terkonsentrasi pada komponen DC dan frekuensi rendah yang merepresentasikan struktur global wajah (bentuk, pencahayaan) dan kurang informatif untuk deteksi *deepfake*. Sebaliknya, artefak sintesis seperti *checkerboard pattern* dan *spectral rolloff* yang tidak normal terutama muncul pada frekuensi menengah dan tinggi. Oleh karena itu *magnitude spectrum* dapat dilewatkan melalui *Gaussian high-pass filter* untuk menekan dominasi komponen frekuensi rendah:

$$H(u,v) = 1 - \exp\left(-\frac{[(u - u_c)^2 + (v - v_c)^2]}{2 \cdot (\beta \cdot N)^2}\right) \quad (2.17)$$

di mana (u_c, v_c) adalah koordinat pusat peta frekuensi setelah *fftshift*, N adalah dimensi peta, dan β adalah *cutoff fraction* yang mengontrol radius penekanan ($\sigma = \beta \cdot N$). *Magnitude* terfilter diperoleh dari $F'(u,v) = F(u,v) \cdot H(u,v)$. Pada penelitian ini nilai β ditetapkan sebesar 0,15, sehingga hanya sekitar 15% frekuensi terendah yang diredam. Nilai ini cukup untuk menekan dominasi komponen frekuensi rendah, yaitu bentuk global wajah dan pencahayaan, tanpa menghilangkan informasi struktural yang masih relevan bagi deteksi.

Untuk memudahkan analisis, spektrum frekuensi dikompresi rentang dinamisnya menggunakan *log scaling* pada persamaan 2.4. Dengan ϵ sebagai konstanta kecil untuk menghindari $\log(0)$. Selain itu, digunakan pula *frequency shifting* (*fftshift*) sehingga komponen frekuensi rendah berada di pusat tampilan spektrum:

$$F_{shift(u,v)} = F\left(\left(u + \frac{M}{2}\right) \bmod M, \left(v + \frac{N}{2}\right) \bmod N\right) \quad (2.18)$$

Karena setiap *dataset* memiliki karakteristik spektral yang berbeda akibat perbedaan kamera, resolusi, dan metode kompresi, peta *magnitude* perlu dinormalisasi sebelum digunakan oleh model. Normalisasi dilakukan dengan *z-score*:

$$\hat{x} = \frac{x - \mu}{\sigma} \quad (2.19)$$

di mana μ dan σ adalah rata-rata dan simpangan baku global yang dihitung dari seluruh piksel peta *magnitude* pada satu *dataset*. Tanpa normalisasi yang tepat, fitur frekuensi antar *dataset* akan memiliki skala yang tidak sebanding. Transformasi ini menghasilkan representasi frekuensi yang lebih mudah ditafsirkan dan lebih stabil untuk diproses oleh model deteksi.

Deepfake berbasis GAN menghasilkan pola-pola artefak yang tidak alami pada frekuensi tinggi akibat:

1. Operasi *upsampling* (*nearest-neighbor*, *transposed convolution*).
2. Ketidakstabilan filter pada generator.
3. Proses *blending* antara wajah asli dan hasil sintesis.
4. Kompresi video berulang kali.

Pola-pola ini sulit terdeteksi pada domain spasial namun sangat jelas pada domain frekuensi. Penelitian *Durall et al.* [8] menunjukkan bahwa citra sintetis memiliki distribusi frekuensi global yang berbeda secara konsisten dibanding citra asli. Temuan tersebut diperkuat *Qian et al.* [11] yang menunjukkan bahwa generator *deepfake* tidak mampu mereplikasi *natural image statistics* pada frekuensi tinggi.

Hasil transformasi FFT berupa spektrum magnitudo dua dimensi yang merepresentasikan distribusi energi frekuensi pada citra. Spektrum ini kemudian dikonversi menjadi satu kanal intensitas (*single-channel representation*) agar kompatibel dengan arsitektur CNN konvensional. Kanal frekuensi yang dihasilkan terutama menonjolkan pola-pola energi pada rentang frekuensi tinggi yang berkaitan dengan artefak generatif pada citra *deepfake*.

Dalam penelitian ini, FFT diterapkan pada citra hasil *cropping* dan *preprocessing* spasial, kemudian dikonversi ke skala log dan dinormalisasi. Representasi frekuensi kemudian digabungkan sebagai *frequency channel* tambahan bersama citra RGB. Pendekatan *hybrid* ini memungkinkan model mempelajari pola manipulasi baik pada domain spasial melalui XceptionNet maupun pada domain frekuensi melalui spektrum FFT.

Sabir et al. [41] menemukan bahwa sinyal frekuensi tinggi lebih efektif dalam membedakan *deepfake* terkompresi tinggi, sedangkan *Wang et al.* [9] menunjukkan bahwa penggunaan informasi frekuensi meningkatkan *robustness* model terhadap variasi kualitas

video. Oleh karena itu, integrasi FFT ke dalam *preprocessing* dianggap penting untuk meningkatkan akurasi dan stabilitas sistem deteksi pada penelitian ini. Maka dari itu, penelitian ini memproses informasi frekuensi pada cabang terpisah (FreqCNN) dan menggabungkannya dengan fitur spasial pada tahap *late fusion*, bukan sebagai kanal tambahan pada masukan XceptionNet.

Pemilihan FFT dalam penelitian ini didasarkan pada sifatnya yang mampu menangkap distribusi spektral global secara komprehensif dengan kompleksitas komputasi yang rendah. Dibandingkan metode frekuensi lain seperti DCT atau *Discrete Wavelet Transform* (DWT), FFT memberikan representasi frekuensi penuh yang lebih sesuai untuk mendeteksi pola artefak GAN yang bersifat global *spectral shift* seperti yang diidentifikasi oleh Durall et al. [8] dan Qian et al. [11]. Atas dasar itu, FFT menjadi pilihan yang efektif dan efisien untuk dikombinasikan dengan model CNN berbasis spasial.

2.14 Augmentasi Data

Augmentasi data merupakan teknik regularisasi yang memperbanyak variasi data pelatihan tanpa menambah jumlah sampel asli, sehingga membantu model menghindari *overfitting* dan meningkatkan generalisasi [4]. Pada deteksi *deepfake* yang sering dilatih dengan data terbatas, augmentasi menjadi penting untuk mencegah model menghafal karakteristik spesifik video pelatihan.

Augmentasi pada **domain spasial** umumnya diterapkan langsung pada citra RGB melalui operasi geometris dan fotometrik, seperti pemotongan acak (*random crop*), pembalikan horizontal (*horizontal flip*), perubahan kecerahan, kontras atau saturasi (*color jitter*), serta penghapusan acak sebagian area (*random erasing*). Operasi-operasi ini memanfaatkan invariansi alami citra wajah terhadap transformasi tersebut.

Augmentasi pada **domain frekuensi** memerlukan pendekatan yang berbeda. Operasi spasial seperti *random crop* atau rotasi tidak sesuai untuk peta *magnitude* FFT, karena setiap posisi piksel pada peta frekuensi merepresentasikan komponen frekuensi spesifik yang bergantung pada posisi absolutnya, dan transformasi geometris akan merusak lokalisasi frekuensi. Oleh karena itu augmentasi frekuensi dilakukan melalui dua mekanisme.

Pertama, **injeksi noise Gaussian** pada peta *magnitude* untuk menyimulasikan variasi *noise* sensor dan artefak kompresi, sekaligus mencegah penghafalan peta FFT yang identik antar-epoch:

$$\hat{x}_{fft} = x_{fft} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \quad (2.20)$$

di mana σ adalah intensitas *noise*. *Noise* hanya diterapkan selama pelatihan, setelah normalisasi.

Kedua, **spectral band masking**, yaitu menutup sebuah pita frekuensi (horizontal atau vertikal) selebar w piksel dengan nilai nol:

$$\hat{x}_{fft}[r_1 : r_1 + w, :] = 0 \quad (\text{pita horizontal}) \quad (2.21)$$

$$\hat{x}_{fft}[:, c_1 : c_1 + w] = 0 \quad (\text{pita vertikal}) \quad (2.22)$$

di mana w adalah lebar pita acak, sedangkan r_1 atau c_1 adalah posisi awal acak. Teknik ini mencegah model bergantung pada satu pita frekuensi spesifik dan mendorong representasi frekuensi yang lebih merata dan *robust*.

Pada arsitektur *hybrid* yang menerima *input* RGB dan FFT secara bersamaan, konsistensi augmentasi antar kedua cabang perlu dijaga, pembalikan horizontal pada citra RGB harus diikuti pembalikan yang sama pada peta FFT agar korespondensi spasial–frekuensi tidak rusak, karena $\text{DFT}(\text{flip}(x)) = \text{flip}(\text{DFT}(x))$.

2.15 Optimasi Model

Pelatihan model *deep learning* memerlukan algoritma optimasi yang efisien untuk meminimalkan *loss function* dan memperbarui parameter jaringan secara iteratif. Dalam konteks deteksi *deepfake*, pemilihan metode optimasi yang tepat menjadi penting karena model harus dilatih pada *dataset* besar dengan variasi kualitas citra dan jenis manipulasi yang beragam [42].

2.15.1 Stochastic Gradient Descent (SGD) dan Perkembangannya

Stochastic Gradient Descent (SGD) merupakan metode optimasi dasar yang memperbarui parameter model menggunakan estimasi gradien dari *mini-batch* data, sehingga lebih efisien dibandingkan *full-batch gradient descent* pada *dataset* besar. Pembaruan parameter pada SGD dilakukan dengan:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} l(f(x; \theta_t), y) \quad (2.23)$$

di mana η adalah *learning rate* [43, 42]. Namun, SGD memiliki kelemahan berupa sensitivitas tinggi terhadap pemilihan *learning rate* dan kecenderungan osilasi pada permukaan *loss* yang tidak seragam. Untuk mengatasi keterbatasan tersebut, dikembangkan berbagai metode optimasi adaptif yang mengatur *learning rate* secara otomatis per parameter [44].

2.15.2 AdamW (Decoupled Weight Decay Regularization)

Adaptive Moment Estimation (Adam) merupakan metode optimasi yang menggabungkan konsep *momentum* dan *adaptive learning rate* dari *Root Mean Square Propagation* (RMSprop). Adam menghitung estimasi momen pertama (rata-rata gradien) dan momen kedua (rata-rata kuadrat gradien) untuk setiap parameter, kemudian menggunakan estimasi tersebut untuk menyesuaikan langkah pembaruan secara individual [45].

Persamaan pembaruan Adam adalah:

1. Estimasi momen pertama:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2.24)$$

2. Estimasi momen kedua:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2.25)$$

3. Koreksi bias:

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (2.26)$$

dan

$$\widehat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (2.27)$$

4. Pembaruan parameter:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v}_t} + \epsilon} \widehat{m}_t \quad (2.28)$$

di mana β_1 dan β_2 adalah koefisien peluruhan momen (umumnya $\beta_1 = 0.9$, $\beta_2 = 0.999$), dan ϵ adalah konstanta stabilitas numerik [45].

AdamW merupakan varian dari Adam yang diperkenalkan oleh *Loshchilov* dan *Hutter* [46] dengan mekanisme *decoupled weight decay*. Berbeda dengan implementasi *weight decay* pada Adam standar yang menambahkan regularisasi L2 ke gradien sebelum penskalaan adaptif, AdamW menerapkan *weight decay* secara langsung pada bobot setelah langkah pembaruan Adam. Perbedaan ini memastikan bahwa efek regularisasi konsisten untuk semua parameter, tidak bergantung pada *learning rate* adaptif masing-masing parameter. Dalam skenario *transfer learning*, di mana parameter *pretrained* dan parameter baru memiliki skala gradien yang sangat berbeda, konsistensi ini sangat penting untuk menjaga stabilitas pelatihan.

2.15.3 Keunggulan AdamW untuk Deteksi Deepfake

AdamW dipilih dalam penelitian ini sebagai metode optimasi utama karena beberapa alasan:

1. *Adaptive learning rate* per parameter memungkinkan konvergensi yang lebih stabil pada arsitektur *hybrid* yang memiliki parameter dengan skala gradien yang berbeda antara cabang spasial dan frekuensi.
2. Koreksi bias pada momen pertama dan kedua membuat AdamW efektif bahkan pada tahap awal pelatihan ketika estimasi momen masih tidak akurat.
3. AdamW mengimplementasikan *decoupled weight decay* secara native, sehingga regularisasi lebih konsisten dan efektif pada arsitektur *hybrid* dengan parameter *pretrained* dan baru.
4. Dalam praktik pelatihan model deteksi *deepfake*, Adam telah terbukti menghasilkan konvergensi yang cepat dan stabil pada berbagai arsitektur CNN, termasuk XceptionNet [7].

2.15.4 Gradient Clipping

Untuk mencegah ledakan gradien (*gradient explosion*) yang dapat membuat pelatihan tidak stabil, diterapkan *gradient clipping* berdasarkan norma L2 [42]. Apabila norma vektor gradien melebihi ambang c , gradien diskalakan kembali agar normanya tepat sama dengan:

$$\begin{aligned} g &\leftarrow g && \text{jika } \|g\|_2 \leq c \\ g &\leftarrow g \cdot \left(\frac{c}{\|g\|_2}\right) && \text{jika } \|g\|_2 > c \end{aligned} \quad (2.29)$$

di mana g adalah vektor gradien dan c adalah ambang batas maksimum (*max_norm*). Teknik ini menjaga arah gradien tetap sama namun membatasi besarnya, sehingga pembaruan bobot tidak melonjak terutama pada tahap awal pelatihan atau saat *backbone pretrained* dilepaskan (*unfreezing*).

2.15.5 Binary Cross-Entropy sebagai Fungsi Loss

Dalam deteksi *deepfake*, tugas klasifikasi bersifat biner, yaitu membedakan citra asli (*real*) dan citra hasil manipulasi (*fake*). Fungsi *loss* yang digunakan adalah *Binary Cross-Entropy with Logits* (BCEWithLogitsLoss), yang menggabungkan fungsi *sigmoid* dan *binary cross-entropy* dalam satu operasi yang stabil secara numerik:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\sigma(z_i)) + (1 - y_i) \cdot \log(1 - \sigma(z_i))] \quad (2.30)$$

di mana z_i adalah *logit* keluaran model, $\sigma(z)$ adalah fungsi *sigmoid*, dan $y_i \in \{0,1\}$ adalah label kelas (0 untuk asli, 1 untuk palsu). Fungsi *loss* ini memastikan bahwa model dioptimalkan untuk menghasilkan probabilitas yang akurat dalam membedakan konten asli dan palsu [42].

Apabila distribusi kelas tidak seimbang, BCEWithLogitsLoss dapat dilengkapi dengan bobot kelas positif (*pos_weight*) yang menambah pengaruh kelas minoritas pada pembaruan gradien:

$$w_p = \frac{n_{neg}}{n_{pos}} \quad (2.31)$$

di mana n_{neg} dan n_{pos} adalah jumlah sampel kelas negatif (*real*) dan positif (*fake*). Bobot ini dikalikan pada komponen *loss* untuk sampel positif. Apabila *dataset* seimbang ($n_{neg} = n_{pos}$) maka $w_p = 1$ dan *loss* kembali menjadi BCEWithLogitsLoss standar.

Label *smoothing* merupakan teknik regularisasi [47] untuk mencegah model menjadi terlalu percaya diri (*overconfident*) dengan melunakkan label target:

$$y'_i = y_i \cdot (1 - \alpha) + \alpha \cdot 0,5 \quad (2.32)$$

di mana α adalah faktor *smoothing*. Dengan α kecil (mis. 0,05 yang dipakai pada penelitian ini), label biner 0 dan 1 dilunakkan menjadi 0,025 dan 0,975, sehingga model tidak perlu mendorong *logit* ke nilai ekstrem ($\pm \infty$).

2.16 Metrik Evaluasi Model

Evaluasi kinerja model merupakan tahap penting dalam pengembangan sistem deteksi *deepfake*. Melalui evaluasi, peneliti dapat menilai sejauh mana model mampu membedakan citra asli dari citra hasil manipulasi, sekaligus mengidentifikasi kelemahan model ketika dihadapkan pada variasi distribusi data atau skenario *cross-dataset*. Hal ini penting karena model generatif modern seperti GAN dan variannya menghasilkan konten yang makin realistis dan makin sulit dibedakan secara visual [8, 9, 14].

Pada penelitian ini, metrik evaluasi yang digunakan meliputi *Confusion Matrix* (CM), *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *Area Under the ROC Curve* (AUC). Metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dihitung dari hasil prediksi model pada ambang tertentu ($\theta = 0,5$) terhadap kelas sebenarnya, sedangkan AUC dihitung dari skor probabilitas model sehingga bersifat independen terhadap ambang. AUC digunakan sebagai metrik utama untuk seleksi model dan *early stopping* karena tahan terhadap ketidakseimbangan kelas dan tidak bergantung pada pemilihan ambang.

2.16.1 Confusion Matrix

Confusion matrix merupakan representasi berbentuk tabel yang menggambarkan performa model klasifikasi dengan membandingkan hasil prediksi dengan kondisi aktual. Melalui matriks ini, dapat terlihat jenis kesalahan apa yang dilakukan model serta apakah model lebih cenderung melakukan *false alarm* (FP) atau justru meloloskan *deepfake* (FN), yang berpotensi menimbulkan risiko keamanan signifikan [14, 15].

Tabel 2.5 Struktur Confusion Matrix

	Prediksi: Palsu	Prediksi: Asli
Aktual: Palsu	True Positive (TP)	False Negative (FN)
Aktual: Asli	False Positive (FP)	True Negative (TN)

Keterangan:

1. TP: Model benar mendeteksi *deepfake*.
2. TN: Model benar mendeteksi konten asli.
3. FP: Konten asli salah diklasifikasikan sebagai *deepfake* (*false alarm*).
4. FN: *Deepfake* gagal terdeteksi (paling berbahaya).

2.16.2 Accuracy

Accuracy mengukur seberapa besar proporsi prediksi model yang benar terhadap seluruh data.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2.33)$$

Meskipun sederhana dan umum digunakan, akurasi tidak selalu memberikan gambaran yang representatif, terutama pada *dataset* yang tidak seimbang (misalnya jika jumlah citra asli lebih banyak daripada citra palsu). Pada penelitian *deepfake*, hal ini sering terjadi karena *dataset* asli umumnya memiliki distribusi lebih besar [14].

2.16.3 Precision

Precision mengukur ketepatan model dalam memberikan label “palsu”. Nilai ini menunjukkan seberapa banyak prediksi *deepfake* yang benar-benar merupakan *deepfake*.

$$Precision = \frac{TP}{(TP + FP)} \quad (2.34)$$

Nilai *precision* yang tinggi menunjukkan bahwa model jarang memberikan *false accusation* terhadap citra asli. Hal ini penting pada skenario forensik digital, keamanan media, maupun investigasi yang tidak mengizinkan kesalahan dalam menandai konten asli sebagai palsu [14, 15].

2.16.4 Recall

Recall menunjukkan kemampuan model mendeteksi seluruh *deepfake* yang ada pada data uji.

$$Recall = \frac{TP}{TP+FN} \quad (2.35)$$

Nilai *recall* yang tinggi berarti model jarang meloloskan *deepfake*. Mengingat *deepfake* sering menampilkan artefak halus yang sulit dikenali secara visual [8, 9], *recall* menjadi metrik penting untuk memastikan tidak ada manipulasi berbahaya yang lolos dari pendeteksian.

2.16.5 F1-Score

F1-Score adalah rata-rata harmonik dari *Precision* dan *Recall*, dan digunakan untuk memberikan penilaian yang lebih stabil ketika terjadi ketidakseimbangan antara jumlah data asli dan palsu atau ketika *Precision* dan *Recall* tidak seimbang.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.36)$$

Dalam penelitian *deepfake*, F1-Score sering dipilih sebagai metrik utama karena lebih mencerminkan performa realistis model pada kondisi data yang kompleks dan tidak seimbang [14, 16].

2.16.6 Area Under the ROC Curve

Area Under the Curve (AUC) merupakan metrik evaluasi yang mengukur kemampuan model dalam membedakan dua kelas (asli dan palsu) secara keseluruhan, tanpa bergantung pada satu nilai *threshold* tertentu. AUC dihitung berdasarkan *Receiver Operating Characteristic* (ROC) Curve, yaitu kurva yang menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai *threshold* klasifikasi. TPR yang identik dengan *Recall*, dihitung sebagai:

$$TPR = \frac{TP}{TP+FN} \quad (2.37)$$

Sedangkan FPR dihitung sebagai:

$$FPR = \frac{FP}{FP+TN} \quad (2.38)$$

ROC Curve dihasilkan dengan memplot TPR terhadap FPR pada setiap kemungkinan nilai *threshold*. Model yang sempurna menghasilkan kurva yang mendekati sudut kiri atas grafik (TPR = 1, FPR = 0), dengan nilai AUC mendekati 1.0. Sebaliknya, model acak menghasilkan garis diagonal dengan AUC = 0.5.

AUC memiliki keunggulan penting dalam evaluasi deteksi *deepfake*:

1. *Threshold-independent*: AUC mengevaluasi performa model pada seluruh rentang *threshold*, sehingga memberikan gambaran yang lebih komprehensif dibanding metrik yang bergantung pada satu *threshold* tertentu seperti *accuracy* atau *precision* [14].
2. Stabil pada *dataset* tidak seimbang: AUC tidak terpengaruh oleh proporsi kelas dalam *dataset*, sehingga lebih andal ketika jumlah citra asli dan palsu tidak seimbang.
3. Standar evaluasi *cross-dataset*: Dalam skenario *cross-dataset generalization*, AUC menjadi metrik utama karena mampu menunjukkan kemampuan diskriminasi model secara keseluruhan, terlepas dari distribusi data uji yang mungkin berbeda dari data pelatihan [16, 12].

Dalam penelitian ini, AUC digunakan sebagai metrik utama untuk pemilihan model terbaik selama proses pelatihan, karena kemampuannya dalam memberikan evaluasi yang *robust* dan tidak bergantung pada pemilihan *threshold* tertentu.

2.16.7 Perhitungan Metrik Evaluasi

Untuk memberikan gambaran yang lebih jelas mengenai proses evaluasi model deteksi *deepfake*, berikut disajikan contoh perhitungan menggunakan nilai *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Contoh ini bersifat ilustratif dan tidak berasal dari hasil eksperimen penelitian.

Untuk menghitung total sampel (N), dapat digunakan persamaan di bawah:

$$N = TP + TN + FP + FN \quad (2.39)$$

Misalkan sebuah model deteksi menghasilkan:

TP = 180, yaitu *deepfake* yang benar terdeteksi sebagai palsu

FP = 20, yaitu konten asli yang *salah* diklasifikasikan sebagai palsu

TN = 160, yaitu konten asli yang benar terdeteksi sebagai asli

FN = 40, yaitu *deepfake* yang gagal terdeteksi (*missed detection*)

Untuk menghitung total sampel (N):

$$N = TP + TN + FP + FN$$

$$= 180 + 160 + 20 + 40$$

$$= 400$$

1. Menghitung *Accuracy*

Accuracy menghitung proporsi prediksi yang benar dari seluruh data:

$$\text{Accuracy} = \frac{(TP + TN)}{TP + TN + FP + FN}$$

$$= \frac{(180 + 160)}{400}$$

$$= \frac{340}{400}$$

$$= 0.85 \rightarrow 85\%$$

Interpretasi dari contoh perhitungan *Accuracy* di atas adalah, model berhasil menghasilkan prediksi yang benar pada 85% kasus.

2. Menghitung *Precision*

Precision mengukur ketepatan prediksi "palsu" model:

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

$$= \frac{180}{180 + 20}$$

$$= \frac{180}{200}$$

$$= 0.90 \rightarrow 90\%$$

Interpretasi dari contoh perhitungan *Precision* di atas adalah, model berhasil menyatakan video *deepfake* yang merupakan video *deepfake* dengan tingkatan ketepatan 90%. Kesimpulan, semakin tinggi *precision* maka semakin kecil risiko *false accusation*.

3. Menghitung *Recall* (Sensitivitas)

Recall mengukur kemampuan model dalam menangkap seluruh *deepfake*:

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

$$= \frac{180}{180 + 40}$$

$$= \frac{180}{220}$$

$$= 0.818 \rightarrow 81.8\%$$

Interpretasi dari contoh perhitungan *Recall* di atas adalah, model berhasil mendeteksi 81.8% *deepfake* yang ada. Perhitungan *Recall* digunakan untuk menghitung tingkat kelolosan video *deepfake*.

4. Menghitung *F1-Score*

F1-Score adalah rata-rata harmonik *Precision* dan *Recall*:

$$\begin{aligned} F1 - Score &= 2 \times \frac{(Precision \times Recall)}{Precision + Recall} \\ &= 2 \times \frac{(0.90 \times 0.818)}{0.90 + 0.818} \\ &= 2 \times \frac{0.7362}{1.718} \\ &= \frac{1.4724}{1.718} \\ &= 0.857 \rightarrow 85.7\% \end{aligned}$$

Interpretasi dari contoh perhitungan *F1-Score* di atas adalah, model memiliki nilai 85.7% dalam kemampuannya untuk menghindari *false positive* dan *false negative*.

Tabel 2.6 Ringkasan Hasil Perhitungan Metrik

Metrik	Rumus	Hasil
Accuracy	$\frac{(TP + TN)}{TP + TN + FP + FN}$	85%
Precision	$\frac{TP}{TP + FP}$	90%
Recall	$\frac{TP}{TP + FN}$	81.8%
F1-Score	$2 \times \frac{(Precision \times Recall)}{Precision + Recall}$	85.7%

Berdasarkan hasil perhitungan, model menunjukkan performa yang baik dengan tingkat ketelitian tinggi dalam mendeteksi *deepfake* (*Precision* = 90%). Meskipun demikian, masih terdapat *deepfake* yang tidak terdeteksi (*Recall* = 81.8%). Nilai *F1-Score* (85.7%) mengindikasikan bahwa model mampu mempertahankan keseimbangan antara menghindari *false positive* dan *false negative*.

Dengan hasil tersebut, model berada dalam kategori performa baik, tetapi peningkatan pada aspek *recall* tetap diperlukan, terutama untuk mencegah *deepfake* lolos dari sistem deteksi, sebagaimana direkomendasikan dalam studi-studi terkini [14, 16, 12].

2.16.8 Relevansi Metrik Evaluasi terhadap Penelitian *Deepfake*

Dalam konteks penelitian deteksi *deepfake* berbasis pendekatan *hybrid FFT + XceptionNet*, penggunaan metrik evaluasi tidak hanya menjadi alat ukur performa, tetapi juga indikator penting untuk menilai kemampuan generalisasi dan ketahanan model ketika diuji pada manipulasi generatif yang semakin kompleks. Hal ini relevan karena model *deepfake* modern berbasis GAN mampu menghasilkan konten dengan kualitas sangat realistis, sehingga membuat proses pendeteksian menjadi jauh lebih menantang [8, 9, 14].

1. *Accuracy* memberikan gambaran umum terhadap tingkat keberhasilan prediksi model. Namun, pada kasus *deepfake*, akurasi perlu ditafsirkan hati-hati, terutama ketika *dataset* tidak seimbang atau memiliki variasi kualitas manipulasi yang berbeda-beda [14].
2. *Precision* menjadi metrik penting untuk meminimalkan *false positive*, yaitu kondisi ketika konten asli secara keliru diklasifikasikan sebagai palsu. Hal ini penting pada aplikasi seperti forensik digital dan keamanan media, di mana kesalahan dalam menandai konten asli sebagai *deepfake* dapat menimbulkan konsekuensi serius [14, 15].
3. *Recall* mengukur kemampuan model dalam mendeteksi seluruh *deepfake* yang ada. Nilai *recall* sangat penting karena banyak *deepfake* yang mengandung artefak halus pada domain frekuensi tinggi, yang sering kali sulit diamati langsung pada domain spasial [8, 11, 16]. Oleh karena itu, *recall* menjadi indikator penting dalam memastikan bahwa *deepfake* berkualitas tinggi tidak lolos dari pendeteksian.
4. *F1-Score*, sebagai *harmonic mean* antara *Precision* dan *Recall*, memberikan penilaian yang lebih stabil ketika kedua metrik tersebut tidak seimbang. Pada penelitian *deepfake*, kondisi ini sering terjadi karena *dataset* biasanya mengandung variasi kualitas citra dan beragam jenis teknik manipulasi [14, 16, 12].

Selain itu, *Confusion Matrix* (CM) digunakan untuk memperoleh gambaran menyeluruh mengenai pola kesalahan prediksi model, apakah model lebih cenderung menghasilkan *false alarm* (FP) atau melewatkan *deepfake* (FN). Analisis ini penting terutama pada skenario *cross-dataset*, di mana distribusi data uji berbeda dari data pelatihan [12, 20].

Kombinasi metrik evaluasi tersebut memberikan gambaran menyeluruh mengenai performa pendekatan *hybrid* FFT + XceptionNet, termasuk kemampuan model mempertahankan kinerja terhadap variasi jenis *deepfake*. Hal ini sejalan dengan penelitian yang menekankan pentingnya detektor yang mampu menangkap petunjuk spasial dan frekuensi secara bersamaan untuk meningkatkan generalisasi pada *dataset* dan model generatif yang berbeda [11, 16, 17, 28].

2.17 Cross-dataset Generalization

Cross-dataset generalization merupakan konsep penting dalam penelitian deteksi *deepfake* yang menggambarkan kemampuan sebuah model untuk mempertahankan performa ketika diuji pada *dataset* yang berbeda dari *dataset* pelatihan. Konsep ini muncul sebagai respons terhadap fakta bahwa *deepfake* di lingkungan nyata sering kali memiliki karakteristik yang sangat berbeda dengan *dataset* benchmark yang umum digunakan dalam penelitian akademik. Rössler *et al.* menunjukkan bahwa *dataset* seperti FFPP hanya mencakup jenis manipulasi wajah tertentu dengan kondisi visual yang relatif seragam [7]. Karena itu, model yang dilatih pada *dataset* tunggal cenderung mengalami degradasi performa secara signifikan ketika diuji pada *dataset* lain yang memiliki distribusi data berbeda.

Degradasi performa lintas *dataset* tersebut dapat diukur secara kuantitatif melalui metrik *generalization drop* (Δ), yaitu selisih antara F1-Score pada evaluasi *in-dataset* (dilatih dan diuji pada *dataset* yang sama) dengan F1-Score pada evaluasi *cross-dataset* (diuji pada *dataset* yang berbeda dari data pelatihan):

$$\Delta = F1_{in-dataset} - F1_{cross-dataset} \quad (2.40)$$

Nilai Δ yang mendekati nol menunjukkan generalisasi yang baik, yaitu model tetap mampu mendeteksi *deepfake* dari metode atau *dataset* yang belum pernah dilihat selama pelatihan. Sebaliknya, nilai Δ yang besar mengindikasikan bahwa model terlalu bergantung pada pola spesifik *dataset* pelatihan (*dataset-specific artifacts*) dan gagal menangkap karakteristik manipulasi yang bersifat umum. Dengan demikian, Δ menjadi indikator langsung sejauh mana sebuah arsitektur mampu menahan degradasi performa saat berpindah domain.

Dalam konteks deteksi *deepfake*, tantangan *cross-dataset* terutama disebabkan oleh fenomena *domain shift*. *Domain shift* terjadi ketika model menemukan pola visual atau statistik baru yang tidak muncul pada *dataset* pelatihan. Bentuk *domain shift* dapat berupa

variasi kualitas video, tingkat kompresi, perangkat perekaman, kondisi pencahayaan, hingga teknik manipulasi generatif yang berbeda [8]. *Zhang et al.* juga menegaskan bahwa setiap metode generatif berbasis GAN menghasilkan artefak sintesis yang memiliki pola distribusi berbeda, sehingga membuat model berbasis fitur spasial saja mudah terjebak pada *dataset-specific artifacts* [9]. Kasus khusus dari domain *shift* ini dikenal sebagai *cross-GAN* (*cross-generator*), yaitu ketika model dilatih pada *deepfake* dari satu metode atau arsitektur generatif lalu diuji pada metode lain yang belum pernah dilihat. Dikarenakan perbedaan generator umumnya beriringan dengan perbedaan *dataset*, evaluasi *cross-GAN* dalam praktiknya menyatu dengan evaluasi *cross-dataset* [14, 15].

Secara teoretis, kesulitan lintas *dataset* ini merupakan persoalan yang dipelajari dalam bidang domain *adaptation* dan domain *generalization*. *Ben-David et al.* [47] menunjukkan bahwa galat sebuah model pada domain target dibatasi oleh galatnya pada domain sumber ditambah suatu ukuran jarak (*divergensi*) antara distribusi kedua domain. Semakin jauh distribusi domain uji dari domain latih, semakin besar potensi penurunan performa. Perbedaan kedua pendekatan terletak pada ketersediaan data target:

1. Domain *Adaptation*, memanfaatkan sebagian data domain target (berlabel maupun tidak) selama pelatihan untuk menyelaraskan distribusi.
2. Domain *Generalization*, berupaya melatih model agar tahan terhadap pergeseran tanpa pernah melihat domain target.

Dalam kerangka ini, evaluasi *cross-dataset* pada penelitian ini termasuk skenario domain *generalization*, bukan domain *adaptation*, karena model dilatih pada satu *dataset* lalu diuji pada *dataset* lain tanpa menyertakan data uji selama pelatihan. Teknik adaptasi domain eksplisit berada di luar lingkup penelitian ini dan menjadi salah satu arah pengembangan.

Penelitian-penelitian modern menunjukkan bahwa pendekatan berbasis domain frekuensi cenderung memiliki kemampuan generalisasi lebih baik dibanding fitur spasial murni. *Durall et al.* menemukan bahwa sebagian besar model GAN gagal mereplikasi *natural spectral statistics*, sehingga muncul pola spektral yang konsisten pada berbagai jenis *deepfake*, terlepas dari *dataset* asal [8]. *Qian et al.* memperkuat temuan tersebut dengan menunjukkan bahwa petunjuk frekuensi tinggi biasanya bersifat lebih stabil terhadap variasi domain, sehingga model yang mengeksploitasi sinyal frekuensi mampu mendeteksi manipulasi pada *dataset unseen* lebih baik [11]. *Hasanaath et al.* juga membuktikan bahwa

penggunaan fitur frekuensi eksplisit meningkatkan *robustness* model terhadap kompresi dan *noise* yang bervariasi antar *dataset* [16].

Sebaliknya, pendekatan berbasis CNN murni pada domain spasial umumnya sangat sensitif terhadap distribusi tekstur *dataset* pelatihan. Afchar *et al.* dan Chollet menunjukkan bahwa CNN secara alami mempelajari pola spasial lokal yang bergantung pada tekstur permukaan, sehingga performanya dapat menurun pada *dataset* yang memiliki karakteristik visual berbeda [4, 6]. Hal ini menjelaskan mengapa model deteksi *deepfake* berbasis fitur spasial sering gagal melakukan generalisasi lintas *dataset*, terutama ketika menghadapi variasi kualitas video atau teknik manipulasi baru.

Oleh karena itu, dalam kajian literatur terkini, kemampuan *cross-dataset generalization* menjadi indikator utama kualitas model deteksi *deepfake* modern. Penelitian seperti SpecXNet [13] dan FSBI [16] melaporkan bahwa model *hybrid* yang menggabungkan domain spasial dan domain frekuensi dapat memiliki performa generalisasi lebih baik dibanding model berbasis domain tunggal, meskipun manfaat tersebut bergantung pada representasi frekuensi dan strategi fusi yang digunakan. Hal ini disebabkan oleh sifat pelengkap kedua domain:

1. Domain spasial menangkap anomali tekstur dan struktur wajah.
2. Domain frekuensi menangkap artefak spektral yang lebih universal antar *dataset*.

Konsep *cross-dataset generalization* menegaskan bahwa deteksi *deepfake* tidak dapat bergantung pada pola *dataset* tertentu, tetapi harus mampu mengidentifikasi karakteristik manipulasi yang bersifat umum dan stabil. Oleh karena itu, penelitian terkini mengarah pada integrasi fitur spasial dan frekuensi untuk memperoleh representasi fitur yang lebih robust terhadap perbedaan distribusi data antar-*dataset* [8, 13, 11].

2.18 Dasar Pemilihan Metode

Model generatif kini semakin mampu menghasilkan video dan citra *deepfake* yang sulit dibedakan secara visual dari data asli, sehingga pemilihan metode deteksi yang tepat menjadi penting dalam penelitian ini. Pendekatan spasial berfokus pada analisis fitur visual langsung seperti tekstur, warna, dan tepi wajah, sedangkan pendekatan frekuensi berfokus pada identifikasi artefak matematis yang muncul akibat proses sintesis citra oleh GAN.

Dalam penelitian ini, digunakan dua metode utama, yaitu FFT sebagai representasi domain frekuensi, dan XceptionNet sebagai model ekstraksi fitur spasial. XceptionNet dipilih karena efisiensi DSC [6] dan performanya sebagai detektor spasial terbaik pada FFPP

[7], meskipun keunggulan ini umumnya bersifat *in-dataset*. FFT dipilih karena analisis domain frekuensi dilaporkan mampu menangkap artefak sintesis yang bersifat algoritmik dan lebih stabil lintas *dataset* [8, 9, 12, 16]. Perlu ditekankan bahwa klaim ketahanan lintas-*dataset* domain frekuensi inilah yang justru diuji secara empiris pada penelitian ini, bukan diasumsikan sebagai fakta yang sudah pasti.

Sejumlah penelitian terkini telah menggabungkan domain spasial dan frekuensi, misalnya SpecXNet [13], FSBI [16], dan *Frequency-Domain Masking* [17]. Namun, pendekatan *hybrid* tersebut sebagian besar dioptimalkan untuk performa *in-dataset* pada FFPP dan belum mengevaluasi secara sistematis seberapa besar kontribusi domain frekuensi terhadap ketahanan lintas-dataset [14, 15]. Penelitian ini mengisi celah tersebut dengan merancang model *hybrid* FFT-XceptionNet, sekaligus mengukur kontribusi domain frekuensi melalui perbandingan terkontrol antar model.

2.18.1 Pemilihan Transformasi FFT dalam Deteksi Deepfake

Pendekatan berbasis FFT dipilih karena mampu mendeteksi artefak yang tidak terlihat secara visual di domain spasial, tetapi muncul sebagai pola khas di domain frekuensi. Durall *et al*, menjelaskan bahwa proses *upsampling* pada model GAN menghasilkan distribusi spektrum yang tidak alami akibat terjadinya replikasi spektral atau duplikasi pola frekuensi yang menciptakan *spectral distortion* pada citra sintetis [8]. Pola ketidaksempurnaan tersebut menghasilkan jejak digital yang dikenal sebagai GAN *fingerprints*, yaitu pola periodik khas dari proses sintesis yang tidak ditemukan pada citra asli [9].

Metode FFT memiliki keunggulan dibanding transformasi lain seperti *Discrete Cosine Transform* (DCT) dan *Wavelet Transform*. FFT memiliki kompleksitas komputasi rendah ($O(n \log n)$) sehingga efisien untuk citra beresolusi tinggi [8]. Berbeda dari DCT yang bekerja per blok atau *Wavelet* yang memerlukan dekomposisi multi-resolusi, FFT menganalisis spektrum frekuensi secara global tanpa mengubah dimensi data. Selain itu, FFT dapat menangkap informasi amplitudo dan fase dari seluruh rentang frekuensi, sehingga efektif mendeteksi *periodic noise* dan anomali spektral hasil *upsampling* GAN [9].

Tabel 2.7 Perbandingan tiga metode transformasi frekuensi [25]

Metode	Kelebihan	Kekurangan	Kompleksitas
FFT	Cepat, global, mampu tangkap pola spektrum.	Tidak lokal, sensitif rotasi.	$O(n \log n)$
DCT	Efisien untuk kompresi citra (JPEG).	Tidak menangkap fase.	$O(n^2)$
Wavelet	Analisis multi-resolusi, lokal.	Kompleks implementasi.	$O(n^2 \log n)$

Penelitian oleh *Zhang et al.* menunjukkan bahwa artefak frekuensi GAN bersifat algoritmik dan universal, membuat deteksi berbasis frekuensi lebih tangguh terhadap variasi *deepfake* dibanding pendekatan berbasis piksel [9]. *Hasanaath et al.* juga membuktikan bahwa metode frekuensi seperti FFT memiliki ketahanan lebih baik pada skenario lintas *dataset* dibanding detektor spasial murni [16]. Atas dasar itu, FFT dipilih karena mampu mendeteksi artefak tersembunyi hasil sintesis GAN secara efisien dan bersifat umum, serta dilaporkan memiliki generalisasi yang lebih tinggi, klaim yang justru diuji pada penelitian ini.

2.18.2 Pemilihan Arsitektur XceptionNet dalam Domain Spasial

Arsitektur XceptionNet (*Extreme Inception Network*) dipilih sebagai model untuk mengekstraksi fitur spasial tingkat tinggi secara efisien dengan menggunakan mekanisme DSC, yang memisahkan konvolusi per kanal (*depthwise*) dari penggabungan kanal (*pointwise*). Desain ini membuat XceptionNet lebih efisien secara parameter dibandingkan arsitektur konvensional seperti ResNet, sekaligus mempertahankan sensitivitas tinggi terhadap tekstur, tepi, dan pola detail pada citra wajah hasil manipulasi [3, 4].

Penelitian *Afchar et al.* [4] dengan model MesoNet menunjukkan bahwa arsitektur CNN dangkal dapat mendeteksi wajah palsu secara efektif, meskipun akurasi menurun ketika diuji pada *dataset* yang lebih kompleks. Sebaliknya, *Rössler et al.* [7] melaporkan bahwa XceptionNet mencapai akurasi deteksi hingga 99,26% pada *dataset* FFPP, melampaui performa ResNet dan MesoNet dalam mendeteksi video *deepfake*. Sementara itu, penelitian *Javid Al Haq* yang membandingkan XceptionNet dan ResNet-50 dengan pendekatan *Gaussian filter* serta LBP menemukan bahwa XceptionNet tidak hanya lebih akurat, tetapi juga lebih cepat dalam waktu klasifikasi [3].

XceptionNet telah diakui sebagai *baseline* kuat dan banyak digunakan dalam berbagai penelitian maupun kompetisi deteksi *deepfake*, seperti FFPP [7] dan SpecXNet, yang membuktikan keandalannya secara empiris [13]. Oleh karena itu, XceptionNet dipilih karena:

1. Arsitekturnya efisien namun tetap mampu mengekstraksi fitur spasial kompleks.
2. Terbukti memiliki akurasi lebih tinggi dibanding ResNet dan Mesonet pada FFPP [7].
3. Telah menjadi *baseline* standar dalam berbagai studi deteksi *deepfake*, sehingga memudahkan komparabilitas dengan penelitian sebelumnya.

Tabel 2.8 Perbandingan performa model CNN [3, 4, 13, 7]

Model	Akurasi (FaceForensics++)	Jumlah Parameter	Kelebihan
MesoNet	83,10% (HQ/c23)	~0,03 juta (~28 ribu)	Ringan, cepat
XceptionNet	96–99%	22M	Akurasi tertinggi, efisien

2.18.3 Perbandingan Akurasi dan Ketahanan Metode FFT dan XceptionNet

Berdasarkan pembahasan pada subbab 2.18.1 dan 2.18.2, FFT dan XceptionNet unggul pada dimensi yang berbeda sehingga berpotensi saling melengkapi. XceptionNet unggul pada akurasi *in-dataset* dan efisiensi parameter, tetapi cenderung bergantung pada artefak visual yang spesifik terhadap dataset pelatihan. Sebaliknya, representasi FFT menonjolkan artefak spektral yang bersifat algoritmik dan dilaporkan lebih stabil pada skenario lintas-dataset serta variasi kompresi, meskipun kehilangan informasi lokasi spasial. Ringkasan perbandingan kedua pendekatan disajikan pada Tabel 2.9.

Secara teoretis, kombinasi keduanya dapat menghasilkan representasi yang lebih lengkap, di mana FFT menangkap *spectral distortion* dan GAN *fingerprints* di domain frekuensi, sedangkan XceptionNet menangkap anomali spasial pada struktur wajah dan tekstur. Sejumlah penelitian terkini melaporkan potensi peningkatan akurasi dan ketahanan dari integrasi dua domain [13, 16, 17], meskipun besarnya manfaat bergantung pada representasi frekuensi, strategi fusi, dan karakteristik *dataset*. Penelitian ini bertujuan menguji secara empiris sejauh mana kombinasi tersebut memberikan peningkatan, khususnya pada skenario lintas-dataset.

Tabel 2.9 Ringkasan Perbandingan FFT vs XceptionNet [3, 8, 9, 13, 16, 7]

Aspek Evaluasi	FFT (Cabang Frekuensi)	XceptionNet (Cabang Spasial)	Hybrid
Artefak yang terdeteksi	Pola spektrum, noise periodik	Tekstur, tepi, distorsi wajah	Gabungan keduanya
<i>Robustness</i>	Tinggi terhadap kompresi	Sedang	Sangat tinggi
Kebutuhan komputasi	Rendah	Sedang–Tinggi	Sedang
Generalisasi (cross-dataset)	Baik	Baik	Sangat baik
Akurasi (rata-rata)	90–93%	96–99%	98–99%

