BAB II. Tinjauan Pustaka

Bab ini menyajikan kajian mendalam terhadap penelitian terkait yang menjadi fondasi pengembangan framework restorasi dokumen berbasis GAN dengan orientasi HTR. Pembahasan difokuskan pada analisis kritis terhadap pendekatan yang telah ada, identifikasi kekuatan dan kelemahan metodologis, serta posisi penelitian ini dalam konteks perkembangan terkini. Berbeda dengan tinjauan pustaka konvensional yang bersifat deskriptif, bab ini menganalisis secara sistematis bagaimana setiap komponen teknologi berkontribusi terhadap solusi yang diusulkan dan mengapa pendekatan yang ada belum memadai untuk menyelesaikan permasalahan restorasi dokumen yang berorientasi pada keterbacaan teks.

I.1 Degradasi Dokumen Historis: Karakteristik dan Tantangan

Dokumen historis, khususnya koleksi arsip kolonial Belanda dari era VOC dan Hindia Belanda yang tersimpan di Arsip Nasional RI, mengalami degradasi fisik yang kompleks dan multimodal. Pemahaman mendalam tentang karakteristik degradasi ini esensial untuk merancang strategi restorasi yang efektif.

I.1.1 Taksonomi Degradasi Dokumen

Berdasarkan analisis literatur dan observasi empiris pada koleksi Arsip Nasional, degradasi dokumen dapat dikategorikan menjadi empat tipe utama yang masing-masing memiliki karakteristik visual dan dampak terhadap keterbacaan yang berbeda:

[PLACEHOLDER GAMBAR II.1]

Perbandingan empat jenis degradasi dokumen:

(a) Bleed through pada dokumen VOC
(b) Fading tinta iron gall
(c) Water stains dan foxing
(d) Optical blur dari scanning

Sumber: Koleksi Arsip Nasional RI / Dataset DIBCO

Gambar 1: Taksonomi degradasi dokumen historis dengan contoh visual dari berbagai sumber.

1. Bleed Through (Tembusan Tinta)

Fenomena ini terjadi ketika tinta dari halaman sebaliknya menembus medium

kertas dan menjadi visible pada halaman yang sedang diamati. Bleed through disebabkan oleh kombinasi faktor: kualitas kertas yang porous, tinta berbasis air yang digunakan pada era kolonial, dan degradasi serat selulosa yang meningkatkan permeabilitas kertas seiring waktu. Secara visual, bleed through menciptakan efek bayangan berupa karakter terbalik atau terfragmentasi yang tumpang tindih dengan teks utama.

Dampak terhadap HTR sangat signifikan karena model kesulitan membedakan antara teks latar depan (teks asli) dan derau latar belakang (tembusan). Penelitian Gatos et al. (2006) menunjukkan bahwa bleed through dapat meningkatkan CER hingga 40 persen pada model HTR yang tidak dilatih dengan degradasi serupa. Karakteristik spasial dari bleed through adalah lapisan semi-transparan dengan intensitas yang bervariasi tergantung pada ketebalan kertas dan densitas tinta.

2. Fading (Pemudaran)

Pemudaran tinta merupakan proses degradasi kimia di mana pigmen tinta kehilangan intensitas warna akibat oksidasi, paparan cahaya UV, atau reaksi dengan kontaminan atmosferik seperti sulfur dioxide. Pada dokumen kolonial Belanda, penggunaan iron gall ink (tinta berbasis ferrous sulfate) menyebabkan pola fading yang khas: tinta berubah dari hitam menjadi coklat keabu abuan, dengan kehilangan kontras yang progresif terhadap background kertas.

Dari perspektif komputasional, pemudaran mengurangi rasio sinyal terhadap derau antara teks dan latar belakang, membuat metode pembatasan ambang batas tradisional menjadi tidak efektif. Model HTR mengalami penurunan skor kepercayaan karena ekstraksi fitur pada lapisan konvolusional awal tidak dapat mendeteksi tepi yang terdegradasi. Studi oleh Pratikakis et al. (2013) pada dataset DIBCO menunjukkan bahwa pemudaran moderat dapat meningkatkan CER dari 8 persen menjadi 25 persen.

3. Stains (Noda)

Noda pada dokumen historis berasal dari berbagai sumber: kelembaban yang menyebabkan water damage, jamur atau foxing (bintik coklat akibat oxidative degradation), tumpahan tinta, atau kontaminan organik. Stains bersifat spatial dan irregular, seringkali menutupi sebagian atau seluruh karakter. Karakteristik visual stains adalah variasi intensitas dan tekstur yang non uniform, yang membedakannya dari background kertas normal.

Tantangan utama untuk HTR adalah oklusi parsial atau total dari karakter. Ketika goresan penting dari sebuah karakter tertutup noda, bahkan model tingkat lanjut

seperti HTR berbasis Transformer tidak dapat melakukan inferensi yang akurat. Penelitian Souibgui et al. (2022) menunjukkan bahwa noda yang menutupi lebih dari 30 persen area karakter menyebabkan tingkat kesalahan pengenalan mencapai 80 persen.

4. Blur (Kekaburan)

Kekaburan dapat terjadi karena dua mekanisme: kekaburan optik akibat proses pemindaian dengan kamera atau pemindai berkualitas rendah, dan kekaburan fisik akibat difusi tinta pada kertas yang lembab. Kekaburan bersifat seperti pemfilteran lolos-rendah yang menghilangkan detail frekuensi tinggi seperti sudut dan tepi yang tajam, yang merupakan fitur diskriminatif untuk pengenalan karakter.

Model HTR yang dilatih pada data tajam mengalami ketidakcocokan domain ketika diaplikasikan pada data kabur. Konvolusi lapisan awal yang dirancang untuk mendeteksi tepi menjadi tidak efektif karena gradien intensitas yang seharusnya tajam menjadi gradual. Penelitian menunjukkan bahwa kekaburan dengan ukuran kernel 5x5 piksel dapat meningkatkan CER hingga 18 persen pada model CRNN standar.

I.1.2 Interaksi dan Kombinasi Degradasi

Dalam praktik, dokumen historis jarang mengalami degradasi tunggal. Lebih umum ditemukan kombinasi dari berbagai jenis degradasi yang berinteraksi secara kompleks. Misalnya, tembusan tinta yang tumpang tindih dengan noda air menciptakan kompleksitas visual yang jauh lebih sulit untuk ditangani dibandingkan degradasi individual.

Fenomena ini yang membuat pendekatan tradisional berbasis aturan atau penalaan parameter menjadi tidak skalabel. Setiap kombinasi degradasi memerlukan parameter yang berbeda, dan tidak ada formulasi umum yang dapat menangani variasi ini. Hal ini menjadi motivasi utama untuk pendekatan berbasis deep learning yang dapat mempelajari representasi dari data, bukan mengandalkan aturan buatan.

I.2 Evolusi Metode Restorasi Dokumen: Dari Tradisional hingga Deep Learning

I.2.1 Pendekatan Tradisional: Analisis Kritis Keterbatasan

Metode restorasi tradisional dapat dikategorikan menjadi tiga paradigma utama: berbasis ambang batas, berbasis energi, dan berbasis model statistik. Analisis kritis

terhadap ketiga paradigma ini mengungkapkan keterbatasan fundamental yang memotivasi transisi ke pendekatan pembelajaran mendalam.

Paradigma 1: Thresholding Based Methods

Metode ini mengasumsikan bahwa teks dan background dapat dipisahkan berdasarkan intensity threshold yang optimal. Algoritma klasik seperti Otsu (1979) menggunakan maksimisasi varians antar-kelas untuk menemukan ambang batas yang memisahkan histogram intensitas menjadi dua distribusi. Sauvola dan Pietikainen (2000) mengembangkan ambang batas adaptif yang menghitung ambang batas lokal berdasarkan rata-rata dan deviasi standar dalam jendela ketetanggaan.

Analisis Keterbatasan: Pendekatan ini gagal pada dokumen dengan bleed through karena asumsi distribusi bimodal tidak terpenuhi. Ketika tinta tembusnya memiliki intensitas yang overlap dengan teks asli, tidak ada threshold tunggal atau lokal yang dapat memisahkan keduanya. Eksperimen pada dataset DIBCO 2009 hingga 2019 menunjukkan bahwa Sauvola mencapai F measure hanya 68 persen pada dokumen dengan degradasi berat, jauh di bawah performa metode deep learning yang mencapai 85 persen.

Lebih fundamental lagi, metode berbasis ambang batas bersifat deterministik dan tidak memiliki mekanisme untuk belajar dari data. Setiap jenis dokumen dengan karakteristik degradasi yang berbeda memerlukan penyetelan parameter manual, yang tidak praktis untuk proyek digitalisasi skala besar.

Paradigma 2: Energy Based Methods

Metode ini memformulasikan restorasi sebagai masalah optimisasi di mana tujuannya adalah meminimalkan fungsi energi yang menggabungkan suku kebenaran data dan suku regularisasi. Gatos et al. (2006) mengusulkan fungsi energi yang memaksimalkan latar depan (tinta) sambil meminimalkan degradasi latar belakang, menggunakan operasi morfologis dan analisis komponen terhubung.

Analisis Keterbatasan: Metode berbasis energi memerlukan definisi eksplisit dari asumsi awal tentang karakteristik teks dan degradasi. Misalnya, asumsi bahwa teks memiliki kontinuitas goresan atau komponen terhubung dengan rasio aspek tertentu. Asumsi ini sering dilanggar pada dokumen dunia nyata dengan goresan terfragmentasi akibat pemudaran atau oklusi akibat noda.

Selain itu, proses optimisasi untuk fungsi energi non-cembung memerlukan biaya komputasi yang tinggi dan sering terjebak pada minimum lokal. Kualitas hasil

sangat sensitif terhadap kondisi awal dan parameter regularisasi, yang kembali memerlukan penyetelan ahli.

Paradigma 3: Statistical Model Based Methods

Pendekatan ini memodelkan degradasi sebagai proses statistik dan menggunakan inferensi Bayesian atau Markov Random Fields untuk pemulihan. Metode ini mencoba pemodelan eksplisit dari proses pembentukan degradasi untuk kemudian melakukan operasi invers.

Analisis Keterbatasan: Kompleksitas dari proses degradasi nyata membuat pemodelan statistik eksplisit menjadi tidak dapat ditelusuri. Estimasi parameter untuk model yang kompleks memerlukan data dasar dalam jumlah besar yang tidak tersedia untuk dokumen historis. Lebih penting lagi, model statistik tidak dapat menangkap informasi semantik tingkat tinggi tentang struktur teks yang penting untuk menjaga keterbacaan.

Kesimpulan Analisis: Keterbatasan fundamental dari metode tradisional adalah ketidakmampuan untuk mempelajari representasi hierarkis dari data dan ketergantungan pada fitur buatan tangan atau model eksplisit yang tidak dapat digeneralisasi. Hal ini memotivasi transisi ke pendekatan pembelajaran mendalam yang dapat secara otomatis mempelajari representasi optimal dari data.

I.2.2 Transisi ke Deep Learning: Perubahan Paradigma

Munculnya deep learning menandai perubahan paradigma dalam visi komputer dan pemrosesan citra dokumen. Kemampuan Jaringan Saraf Tiruan Konvolusi (CNN) untuk pembelajaran hierarkis representasi fitur dari piksel mentah membuka kemungkinan baru untuk restorasi dokumen.

Autoencoder untuk Image to Image Translation

Arsitektur autoencoder yang terdiri dari enkoder untuk ekstraksi fitur dan dekoder untuk rekonstruksi pertama kali diaplikasikan untuk penghilangan derau dokumen oleh Zhao et al. (2019). Enkoder melakukan penurunan sampel bertahap dengan lapisan konvolusional untuk mengekstrak fitur tingkat tinggi, sementara dekoder melakukan peningkatan sampel untuk merekonstruksi citra bersih.

Kontribusi: Autoencoder dapat mempelajari model implisit dari proses degradasi dan restorasi tanpa memerlukan formulasi eksplisit. Pelatihan dengan data berpasangan (terdegradasi dan bersih) memungkinkan model untuk mempelajari pemetaan kompleks yang tidak dapat direpresentasikan dengan aturan buatan

tangan.

Limitasi: Autoencoder standar mengalami kehilangan informasi leher botol karena penurunan sampel ekstrem di enkoder. Detail halus seperti goresan tipis atau tanda diakritik sering hilang dalam rekonstruksi. Selain itu, tujuan pelatihan berbasis kerugian L2 (MSE) cenderung menghasilkan keluaran yang terlalu halus dan kabur.

U Net: Skip Connections untuk Detail Preservation

[PLACEHOLDER GAMBAR II.4]

Arsitektur U-Net untuk Document Restoration:

[Encoder Path] → [Bottleneck] → [Decoder Path]
↓ Skip Connection ↓ Skip Connection ↓
Downsample 1024→512→256→128→64
Upsample 64→128→256→512→1024

Skip connections preserve spatial details

Adaptasi dari: Ronneberger et al. (2015)

Gambar 2: Arsitektur U-Net dengan skip connections yang menghubungkan encoder ke decoder untuk preservasi detail spasial yang krusial untuk ketajaman karakter.

Ronneberger et al. (2015) memperkenalkan U Net architecture untuk biomedical image segmentation, yang kemudian diadopsi untuk document restoration. Innovation utama adalah skip connections yang menghubungkan feature maps dari encoder langsung ke decoder pada resolusi yang sama.

Kontribusi: Koneksi pintas memungkinkan dekoder untuk mengakses detail spasial tingkat rendah yang jika tidak akan hilang karena penurunan sampel. Hal ini krusial untuk menjaga goresan tipis dan detail karakter. Penelitian oleh Tensmeyer dan Martinez (2017) menunjukkan bahwa U Net mencapai PSNR 32 dB pada document binarization, improvement signifikan dibandingkan autoencoder standard yang hanya 28 dB.

Limitasi yang Masih Ada: Meskipun koneksi pintas membantu menjaga detail, U-Net yang dilatih dengan kerugian L1 atau L2 saja masih menghasilkan keluaran yang secara visual kurang realistis. Terutama pada area dengan degradasi tinggi, rekonstruksi cenderung artifaktual. Yang lebih penting, tidak ada mekanisme eksplisit untuk memastikan keterbacaan teks karena optimisasi hanya fokus pada kemiripan tingkat piksel dengan data dasar.

I.3 Generative Adversarial Networks: Teori dan Aplikasi untuk Restorasi Dokumen

I.3.1 Arsitektur dan Mekanisme GAN

Generative Adversarial Networks, diperkenalkan oleh Goodfellow et al. (2014), merepresentasikan terobosan dalam pemodelan generatif melalui pelatihan adversarial. Kerangka kerja ini melibatkan dua jaringan saraf tiruan yang bermain dalam permainan minimax: Generator (G) yang mencoba menghasilkan contoh realistis, dan Diskriminator (D) yang mencoba membedakan antara contoh nyata dan contoh hasil generasi.

[PLACEHOLDER GAMBAR II.5]

Framework Adversarial Training GAN:

Noise z → [Generator G] → Fake Sample
Real Sample ↘
[Discriminator D] → Real/Fake
Fake Sample ↗

G minimize: log(1-D(G(z)))
D maximize: log(D(x)) + log(1-D(G(z)))

Sumber: Goodfellow et al. (2014)

Gambar 3: Framework adversarial training pada GAN di mana Generator dan Discriminator bermain minimax game untuk mencapai Nash equilibrium.

Formulasi Matematis

Training process dirumuskan sebagai value function yang di optimize:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

Di mana x adalah sampel nyata dari distribusi data p_{data}, z adalah derau acak dari distribusi prior p_z, $G(z)$ adalah sampel yang dihasilkan, dan $D(x)$ adalah probabilitas bahwa x adalah nyata (bukan dihasilkan).

Interpretasi: Diskriminator dimaksimalkan untuk memberikan probabilitas tinggi ke sampel nyata ($D(x) \to 1$) dan probabilitas rendah ke sampel yang dihasilkan ($D(G(z)) \to 0$). Sebaliknya, Generator diminimalkan untuk menipu diskriminator dengan membuat $D(G(z)) \to 1$. Dalam kesetimbangan, Generator mempelajari distribusi yang mendekati distribusi data, dan Diskriminator tidak dapat

membedakan antara nyata dan yang dihasilkan.

Conditional GAN untuk Image to Image Translation

Untuk restorasi dokumen, diperlukan generasi bersyarat di mana Generator tidak hanya menghasilkan citra bersih acak, tetapi citra bersih yang bersesuaian dengan masukan terdegradasi. Mirza dan Osindero (2014) memperkenalkan Conditional GAN (cGAN) yang mengkondisikan kedua G dan D pada informasi tambahan y:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{y,z}[\log(1 - D(G(y,z),y))] \tag{2}$$

Untuk restorasi dokumen, y adalah citra terdegradasi dan x adalah citra bersih yang bersesuaian. Generator mengambil citra terdegradasi sebagai masukan dan menghasilkan versi bersih, sementara Diskriminator menilai apakah pasangan (citra bersih, citra terdegradasi) adalah nyata atau palsu.

I.3.2 Pix2Pix dan Evolusi untuk Output Terstruktur

Isola et al. (2017) mengembangkan kerangka kerja Pix2Pix yang mengkombinasikan arsitektur cGAN dengan Generator U-Net dan Diskriminator PatchGAN. Kerangka kerja ini menjadi fundamental untuk tugas terjemahan citra ke citra termasuk restorasi dokumen.

Innovation Kunci

1. PatchGAN Discriminator: Alih alih mengklasifikasikan seluruh citra sebagai nyata atau palsu, PatchGAN mengklasifikasikan potongan N x N. Keluaran Diskriminator adalah matriks di mana setiap elemen bersesuaian dengan potongan dari citra masukan. Pendekatan ini memiliki dua keuntungan: (a) fokus pada detail frekuensi tinggi yang penting untuk realisme tekstur, dan (b) parameter lebih sedikit karena arsitektur konvolusional penuh.

2. Loss Function Combination: Pix2Pix mengkombinasikan adversarial loss dengan L1 reconstruction loss:

$$\mathscr{L}_{total} = \mathscr{L}_{GAN}(G,D) + \lambda \mathscr{L}_{L1}(G) \tag{3}$$

Di mana $\mathscr{L}_{L1}(G) = \mathbb{E}_{x,y}[\|x - G(y)\|_1]$. Kerugian L1 mendorong Generator untuk menghasilkan keluaran yang mirip secara piksel dengan data dasar, sementara kerugian adversarial mendorong detail tekstur yang realistis.

Analisis Kontribusi untuk Restorasi Dokumen: Pix2Pix menunjukkan hasil

mengesankan untuk translasi citra-ke-citra umum. Namun, untuk restorasi dokumen, kerangka kerja ini memiliki keterbatasan: tidak ada mekanisme untuk secara eksplisit memastikan keterbacaan teks. Generator dioptimalkan untuk kemiripan visual dan menipu diskriminator, tetapi tidak ada jaminan bahwa teks hasil restorasi dapat dikenali dengan baik oleh sistem HTR.

I.4 State of the Art dalam Restorasi Dokumen: Analisis Komprehensif

I.4.1 DE-GAN: Peningkatan Dokumen dengan Pelatihan Adversarial dan Evaluasi OCR

Souibgui et al. (2021) mengusulkan **DE-GAN (Document Enhancement Generative Adversarial Network)** dengan judul lengkap "DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement". Ini merupakan penelitian fundamental pertama yang menerapkan GAN untuk document enhancement dan merupakan aplikasi perdana GAN dalam domain dokumen.

I.4.2 ERB-MultiTaskAdversarial: Fully Gated CNN dengan CRNN Integration

Souibgui et al. (2019) [1] mengusulkan **ERB-MultiTaskAdversarial** dengan judul "Enhance to Read Better: A Multi-Task Adversarial Network for Handwritten Document Image Enhancement". Ini merupakan penelitian pionir yang pertama kali mengintegrasikan Convolutional Recurrent Neural Network (CRNN) recognizer secara langsung ke dalam GAN training process untuk document enhancement.

Inovasi Fundamental: GAN + CRNN Integration

ERB-MultiTaskAdversarial merepresentasikan terobosan signifikan sebagai **penelitian pertama** yang mengintegrasikan deep generative neural networks dengan Recurrent Neural Networks (RNN) untuk handwritten document enhancement. Kontribusi kunci meliputi:

1. Arsitektur GAN + CRNN Ujung-ke-Ujung: Menggabungkan GAN untuk peningkatan kualitas dokumen dengan pengenal CRNN dalam satu kerangka kerja terpadu. Ini adalah pendekatan pertama yang secara simultan menangani tugas peningkatan kualitas dan transkripsi.

2. Fully Gated Convolutional Neural Networks (FGCNN): Menggunakan arsitektur FGCNN sebagai generator yang secara khusus dirancang untuk

peningkatan kualitas dokumen dengan mekanisme gerbang untuk pemilihan fitur yang lebih baik.

3. Pelatihan Skenario Ganda: Mengimplementasikan dua skenario pelatihan yang berbeda:

- **Skenario S1:** Pengenal CRNN diberi citra bersih dasar selama pelatihan GAN
- **Skenario S2:** Pengenal CRNN diberi citra yang dihasilkan (dibersihkan oleh generator) selama pelatihan

4. Dukungan Multi-Skrip: Demonstrasi efektivitas pada skrip Arab (basis data KHATT) dan skrip Latin (basis data IAM), menunjukkan fleksibilitas lintas bahasa.

Arsitektur Fully Gated CNN + CRNN

Inovasi teknis utama dari ERB-MultiTaskAdversarial:

- **Generator:** Fully Gated Convolutional Neural Networks (FGCNN) dengan mekanisme gerbang untuk pemilihan fitur adaptif
- **Diskriminator:** CNN standar yang mengevaluasi realisme dari citra bersih yang dihasilkan
- **Pengenal CRNN:** Convolutional Recurrent Neural Network yang terintegrasi dalam lingkaran pelatihan
- **Fungsi Kerugian Gabungan:** $L_{total} = L_{GAN} + \lambda L_{CRNN}$ untuk optimisasi simultan

Eksperimen dan Hasil Komprehensif

ERB-MultiTaskAdversarial dievaluasi secara ekstensif:

1. Dataset Multi-Skrip:

- **Basis Data KHATT:** Dokumen tulisan tangan historis Arab
- **Basis Data IAM:** Dokumen tulisan tangan Latin
- **DIBCO 2018:** Tolok ukur binarisasi dokumen

2. Metrik Evaluasi Ganda:

- **Kualitas Visual:** PSNR, F-measure, dan kinerja binarisasi
- **Keterbacaan:** Kinerja pengenalan CRNN (CER) dengan pengenal ganda

3. Pengujian Pengenalan Lintas-Domain:

- **CRNN1:** Pengenal dilatih pada citra dasar
- **CRNN2:** Pengenal dilatih pada citra yang dihasilkan
- Pengujian pada citra terdegradasi dan yang ditingkatkan untuk evaluasi komprehensif

Analisis Kritis dan Keterbatasan Fundamental

Meskipun ERB-MultiTaskAdversarial merupakan terobosan signifikan sebagai penelitian pertama yang mengintegrasikan GAN dengan CRNN, terdapat beberapa keterbatasan fundamental:

1. Integrasi Parsial: Pengenal CRNN digunakan selama pelatihan tetapi tidak sepenuhnya terintegrasi dalam perambatan-balik ke generator. CRNN bertindak sebagai sinyal kerugian tambahan daripada dapat dilatih secara ujung-ke-ujung sepenuhnya.

2. Kesadaran Diskriminator Terbatas: Diskriminator tetap hanya mengevaluasi realisme visual tanpa kesadaran terhadap konten teks atau kinerja CRNN. Tidak ada lingkaran umpan balik dari kinerja pengenalan ke proses diskriminasi.

3. Kompleksitas Arsitektural: Penggunaan Fully Gated CNN menambah kompleksitas tanpa demonstrasi jelas superioritas dibandingkan dengan arsitektur U-Net standar untuk peningkatan kualitas dokumen.

4. Trade-off Multi-Tugas: Penyeimbangan antara kerugian GAN dan kerugian CRNN memerlukan penalaan yang cermat tanpa panduan teoretis. Potensi konflik antara realisme visual dan optimisasi pengenalan tidak sepenuhnya terselesaikan.

5. Keterbatasan Protokol Evaluasi: Meskipun mengintegrasikan evaluasi CRNN, pengujian masih dilakukan secara terpisah untuk kualitas visual dan keterbacaan, bukan sebagai tujuan optimisasi terpadu.

Signifikansi Historical dan Kontribusi

ERB-MultiTaskAdversarial memiliki kontribusi historis yang sangat penting:

- **Integrasi Pelopor:** Penelitian pertama yang membuka paradigma integrasi GAN+HTR

- **Validasi Multi-Skrip:** Demonstrasi penerapan lintas-bahasa
- **Evaluasi Komprehensif:** Standar untuk mengukur kinerja visual dan OCR
- **Pelatihan Berbasis Skenario:** Metodologi untuk pelatihan dengan berbagai konfigurasi pengenal

Gap yang Teridentifikasi:

Berdasarkan analisis ERB-MultiTaskAdversarial, terdapat gap yang belum teratasi:

- **Integrasi Ujung-ke-Ujung Sejati:** Perambatan-balik penuh dari kerugian CRNN ke bobot generator
- **Diskriminasi Sadar-Teks:** Diskriminator yang mempertimbangkan konten teks
- **Fungsi Tujuan Terpadu:** Fungsi optimisasi tunggal untuk visual dan keterbacaan
- **Landasan Teoretis:** Kerangka kerja formal untuk menyeimbangkan tujuan yang bersaing

Relevance untuk Penelitian Ini:

Penelitian ini mengatasi gap-gap tersebut dengan:

- **Kerugian CTC Terintegrasi:** Optimisasi langsung untuk kinerja pengenalan
- **Diskriminator Moda-Ganda:** Proses diskriminasi sadar-teks
- **Fungsi Kerugian Multi-Komponen:** Optimisasi seimbang dengan pembobotan eksplisit
- **Pelatihan Ujung-ke-Ujung:** Perambatan-balik sejati dari pengenalan ke peningkatan kualitas

Kontribusi Pionir dan Metodologis

DE-GAN merepresentasikan terobosan signifikan sebagai **aplikasi pertama GAN** untuk document enhancement. Kontribusi fundamental meliputi:

1. Kerangka Kerja Adversarial untuk Dokumen: Mengadaptasi conditional GANs (cGANs) dari translasi citra-ke-citra ke domain peningkatan kualitas dokumen.

Penggunaan U-Net sebagai generator dan Fully Convolutional Network sebagai diskriminator menjadi standar industri.

2. Kemampuan Multi-Tugas: Mendemonstrasikan kemampuan satu kerangka kerja untuk menangani berbagai tugas: pembersihan dokumen, binarisasi, penghilangan kekaburan, dan penghilangan tanda air. Ini menunjukkan fleksibilitas yang luar biasa dari pendekatan GAN.

3. Kinerja Superior Terukur: Pada DIBCO 2013, DE-GAN mencapai PSNR 24.9 dB dan F-measure 99.5%, mengungguli semua metode sebelumnya. Pada DIBCO 2017, mencapai PSNR 18.74 dB dan F-measure 97.91%, menduduki peringkat teratas di antara 26 algoritma yang berpartisipasi.

Eksperimen Readability yang Revolusioner

DE-GAN merupakan penelitian pertama yang secara eksplisit mengevaluasi dampak document enhancement terhadap OCR performance. Temuan krusial:

1. Evaluasi OCR dengan Tesseract: DE-GAN mengurangi Character Error Rate (CER) dari 0.37 (dokumen terdegradasi) menjadi 0.01 (dokumen yang ditingkatkan), menunjukkan peningkatan dramatis dalam keterbacaan. Ini bukti konkret bahwa peningkatan kualitas visual secara signifikan meningkatkan kinerja OCR.

2. Paradigma Pembelajaran Progresif: Dua skenario pelatihan yang dievaluasi:

- **Skenario S1:** Pengenal pra-terlatih tanpa penalaan halus
- **Skenario S2:** Pembelajaran progresif dengan penalaan halus pada hasil peningkatan kualitas

3. Validasi Komprehensif Multi-Dataset: Eksperimen dilakukan pada berbagai dataset:

- **Basis Data Kantor Berisik** untuk pembersihan dokumen (112 citra latih, 32 citra uji)
- **Seri DIBCO** (2013, 2017) untuk binarisasi dengan 6824 potongan latih
- **Dataset Tanda Air Khusus** (1000 citra berpasangan) untuk penghilangan tanda air
- **Dataset Penghilangan Kekaburan** dengan 4000 potongan latih

4. Pembelajaran Transfer Lintas-Domain: Demonstrasi bahwa model yang dilatih pada degradasi sintetis dapat ditransfer ke dokumen historis nyata dengan kinerja yang wajar, menunjukkan kemampuan generalisasi yang baik.

Kontribusi pada Multi-Task Document Processing

DE-GAN memperkenalkan beberapa masalah baru yang menjadi fokus penelitian selanjutnya:

- **Penghilangan Tanda Air Padat:** Pertama kali mengatasi masalah tanda air/stempel padat pada dokumen dengan PSNR 40.98 dan SSIM 0.998, mengungguli metode citra alami (Dekel et al.: PSNR 36.02, Wu et al.: PSNR 23.37)
- **Penghilangan Kekaburan Dokumen:** Mencapai PSNR 20.37 dB, mengungguli CNN dasar (19.36 dB) dan pix2pix-HD (19.89 dB)
- **Aplikasi Dokumen Nyata:** Berhasil diterapkan pada dokumen historis alami dengan degradasi seperti noda dan tembusan

Perbandingan Komprehensif dengan Metode GAN Lain

DE-GAN dibandingkan secara langsung dengan metode GAN lain pada H-DIBCO 2018:

- **vs CycleGAN:** DE-GAN superior (PSNR 16.16 vs 11.00, F-measure 77.59 vs 56.33)
- **vs pix2pix-HD:** DE-GAN lebih baik (PSNR 16.16 vs 14.42, F-measure 77.59 vs 72.79) dengan data pelatihan yang lebih sedikit
- **vs FCN (U-Net):** DE-GAN mengungguli untuk pembersihan dokumen (PSNR 38.12 vs 36.02)

Analisis Keterbatasan Fundamental

Meskipun DE-GAN telah melakukan eksperimen readability yang komprehensif, terdapat beberapa keterbatasan fundamental dalam pendekatan mereka:

1. Evaluasi OCR Pasca-Pelatihan: Evaluasi keterbacaan dilakukan setelah pelatihan GAN selesai, bukan diintegrasikan dalam lingkaran optimisasi. Artinya, Generator dioptimasi hanya untuk kualitas visual, bukan untuk kinerja OCR. Kesenjangan ini signifikan karena peningkatan visual tidak selalu linear dengan peningkatan keterbacaan.

2. Jalur Pelatihan Terpisah: Skenario pembelajaran progresif memerlukan dua tahap pelatihan terpisah:

1. Pelatihan GAN untuk peningkatan kualitas dokumen
2. Pelatihan HTR pada hasil peningkatan kualitas

Pendekatan ini kurang efisien dan tidak memanfaatkan umpan balik langsung dari evaluasi HTR untuk panduan selama pelatihan.

3. Diskriminator Hanya-Visual: Diskriminator DE-GAN hanya mengevaluasi apakah citra "terlihat bersih" tanpa mempertimbangkan:

- Preservasi struktur teks
- Integritas karakter
- Keterbacaan untuk sistem OCR

Diskriminator menggunakan kerugian entropi silang biner standar tanpa pengkondisian pada konten teks.

4. Konflik Fungsi Tujuan: Kerugian adversarial mendorong Generator untuk menghasilkan tekstur realistis yang mungkin mengganggu ekstraksi fitur pada model HTR. Derau ringan yang membuat citra terlihat lebih "alami" dapat sebenarnya menurunkan akurasi OCR.

5. Kurangnya Kesadaran Teks: Tidak ada mekanisme untuk memastikan bahwa konten teks terjaga dengan baik. Generator dapat mengorbankan integritas karakter untuk penampilan visual yang "lebih bersih".

Gap yang Teridentifikasi: Berdasarkan analisis komprehensif terhadap DE-GAN, terdapat tiga gap fundamental yang menjadi primary motivation untuk penelitian ini:

1. Pelatihan HTR-GAN Terintegrasi: DE-GAN melakukan evaluasi keterbacaan secara pasca-pelatihan, bukan terintegrasi dalam lingkaran pelatihan. Penelitian ini mengusulkan integrasi langsung kerugian CTC dari model HTR ke dalam fungsi kerugian Generator.

2. Diskriminator Moda-Ganda: Diskriminator DE-GAN hanya mengevaluasi aspek visual tanpa mempertimbangkan koherensi teks. Penelitian ini mengusulkan Diskriminator Moda-Ganda yang secara simultan mengevaluasi koherensi visual dan tekstual.

3. Optimisasi Ujung-ke-Ujung: DE-GAN menggunakan jalur pelatihan terpisah yang kurang efisien. Penelitian ini mengusulkan optimisasi ujung-ke-ujung dengan fungsi kerugian multi-komponen yang secara eksplisit menyeimbangkan kualitas visual dan keterbacaan teks.

I.4.3 Text-DIAE: Self-Supervised Degradation Invariant Autoencoder

Souibgui et al. (2022) mengusulkan **Text-DIAE (Text Degradation Invariant Autoencoder)**, pendekatan inovatif yang menggabungkan self-supervised learning dengan degradation invariance untuk tugas text recognition dan document enhancement. Framework ini berbeda secara fundamental dari metode lain karena tidak memerlukan labeled data untuk pre-training.

Konsep Inti dan Arsitektur

Text-DIAE mengusulkan dua-stage training paradigm yang revolusioner:

1. Self-Supervised Pre-training: Model dilatih pada unlabeled data dengan tiga pretext tasks:

- **Masking:** Randomly masking 75% of image patches dan reconstruct original content
- **Blurring:** Applying artificial blur degradation dan reconstruction
- **Noise Addition:** Adding synthetic noise dan denoising

2. Supervised Fine-tuning: Pre-trained encoder kemudian di-fine-tune untuk specific downstream tasks (HTR atau document enhancement) dengan minimal labeled data.

Innovation kunci adalah **degradation invariance**: encoder belajar menghasilkan representasi laten yang serupa untuk clean image dan berbagai versi terdegradasi dari gambar yang sama.

Arsitektur Vision Transformer

Text-DIAE menggunakan ViT-based architecture dengan encoder-decoder structure:

- **Encoder:** Standard Vision Transformer dengan Multi-Head Self-Attention
- **Decoder:** Transformer blocks yang reconstruct original image dari latent representations

- **Multi-Task Learning:** Simultaneous training pada multiple degradation types

Keunggulan Komparatif

Text-DIAE menunjukkan beberapa keunggulan signifikan:

1. Data Efficiency: Membutuhkan 43-166 kali lebih sedikit data untuk konvergen dibandingkan contrastive learning methods seperti SeqCLR dan SimCLR. Pada eksperimen IAM, Text-DIAE hanya menggunakan 18.2M samples vs 409M samples untuk metode sebelumnya.

2. Superior Performance: Mencapai peningkatan accuracy hingga +20 points pada handwritten text (IAM) dan +30 points pada scene-text (IIIT5K, ICDAR13) dibandingkan state-of-the-art self-supervised methods.

3. State-of-the-Art Results: Pada document binarization DIBCO benchmarks, Text-DIAE outperforms semua metode sebelumnya termasuk DocEnTr dengan peningkatan PSNR signifikan (+2 points pada skala logaritmik untuk deblurring).

4. Dual Task Capability: Satu framework yang dapat menangani text recognition dan document enhancement secara bersamaan, menunjukkan representasi yang lebih universal.

Analisis Mendalam: Keterbatasan untuk Dokumen Historis

Meskipun mencapai hasil impressive, Text-DIAE memiliki keterbatasan khusus untuk aplikasi dokumen historis:

1. Gap antara Invariance dan Readability: Text-DIAE mengoptimalkan degradation invariance (representasi sama untuk clean/degraded images), namun tidak secara eksplisit mengoptimalkan keterbacaan teks untuk HTR systems. Eksperimen menunjukkan peningkatan visual quality tidak selalu linear dengan penurunan CER.

2. Synthetic Degradation Limitation: Pra-pelatihan menggunakan degradasi sintetis (penutupan, kekaburan, derau) yang mungkin tidak menangkap karakteristik kompleks dari degradasi alami dokumen historis seperti tembusan, bercak cokelat, atau degradasi kimia.

3. Lack of Adversarial Training: Tanpa komponen adversarial, Text-DIAE cenderung menghasilkan keluaran yang terlalu halus. Hal ini terlihat dari perbandingan kualitatif dengan DocEnTr di mana Text-DIAE menghasilkan tekstur yang kurang realistis.

4. Computational Complexity: Kompleksitas kuadratik dari mekanisme

perhatian-diri membuatnya kurang praktis untuk dokumen historis beresolusi tinggi yang umum dalam arsip.

Analisis Eksperimental: Kinerja OCR

Eksperimen menarik dari Text-DIAE menunjukkan:

- Pada tugas penghilangan kekaburan, Text-DIAE mengurangi CER dari 78.86% (asli) menjadi 8.94% (vs DocEnTr: 18.51%)
- Namun, CER absolut masih jauh dari data dasar (4.88%)
- Ini menunjukkan bahwa meskipun peningkatan signifikan, masih ada kesenjangan antara peningkatan kualitas visual dan keterbacaan teks optimal

Relevance untuk Penelitian Ini

Text-DIAE merepresentasikan kemajuan penting dalam self-supervised learning untuk document processing, namun penelitian ini mengatasi gap yang tersisa:

- **Optimisasi Berorientasi HTR:** Text-DIAE mengoptimalkan invarian, penelitian ini mengoptimalkan keterbacaan secara eksplisit melalui kerugian CTC
- **Degradasi Realistis:** Text-DIAE menggunakan degradasi sintetis, penelitian ini menggunakan degradasi dokumen historis nyata
- **Pelatihan Adversarial:** Text-DIAE tanpa adversarial, penelitian ini dengan Diskriminator Moda-Ganda untuk realisme visual
- **Integrasi Ujung-ke-Ujung:** Text-DIAE pelatihan dua-tahap, penelitian ini optimisasi ujung-ke-ujung

I.4.4 DocEnTr: Transformer-Only Approach untuk Document Enhancement

Souibgui et al. (2022) memperkenalkan **DocEnTr (Document Enhancement Transformer)**, pendekatan inovatif yang sepenuhnya mengandalkan arsitektur transformer tanpa menggunakan lapisan CNN sama sekali. Kerangka kerja ini merepresentasikan evolusi terbaru dalam peningkatan kualitas dokumen dengan memanfaatkan kemampuan mekanisme perhatian-diri untuk pemrosesan global.

Arsitektur dan Inovasi Fundamental

DocEnTr mengusulkan encoder-decoder berbasis transformer yang sepenuhnya berbeda dari pendekatan CNN+Transformer hybrid:

1. Arsitektur Sepenuhnya Transformer: Berbeda dengan pendekatan sebelumnya yang masih menggunakan CNN untuk ekstraksi fitur, DocEnTr menggunakan murni lapisan transformer. Masukan citra langsung dibagi menjadi potongan (16x16 piksel) dan diproses tanpa ekstraksi fitur berbasis CNN.

2. Mekanisme Perhatian-Diri Global: Setiap potongan dapat mengakses informasi dari semua potongan lain secara langsung melalui perhatian-diri multi-kepala. Ini memberikan kemampuan untuk menangkap dependensi jarak jauh tanpa batasan medan penerima yang karakteristik CNN.

3. Pemrosesan Berbasis Potongan: Citra diproses pada tingkat potongan dengan pengkodean posisional untuk mempertahankan informasi spasial. Dekoder merekonstruksi citra bersih dari potongan-potongan yang telah ditingkatkan kualitasnya.

Keunggulan Komparatif

DocEnTr menunjukkan beberapa keunggulan signifikan dibandingkan metode berbasis CNN:

1. Pemahaman Konteks Global: Mekanisme perhatian-diri memungkinkan model untuk memahami konteks global dari dokumen, yang penting untuk konsistensi visual dan koherensi struktural.

2. Tanpa Bias Induktif: Tanpa bias induktif konvolusional, model dapat mempelajari representasi yang lebih fleksibel dan tidak terbatas pada pola lokal.

3. Kinerja Superior: Pada tolok ukur DIBCO, DocEnTr mencapai kinerja mutakhir dengan PSNR signifikan lebih tinggi dibandingkan metode berbasis CNN.

Variasi Model dan Konfigurasi

DocEnTr tersedia dalam tiga konfigurasi yang menyeimbangkan kompleksitas komputasi dan kinerja:

- **DocEnTr-Small:** 6 lapisan, 512 dimensi, 4 kepala perhatian, 17M parameter
- **DocEnTr-Base:** 12 lapisan, 768 dimensi, 8 kepala perhatian, 68M parameter
- **DocEnTr-Large:** 24 lapisan, 1024 dimensi, 16 kepala perhatian, 255M parameter

Analisis Kritis untuk Dokumen Historis

Meskipun DocEnTr menunjukkan hasil impressive pada benchmark standar, terdapat beberapa pertimbangan kritis untuk aplikasi pada dokumen historis:

1. Kurangnya Bias Induktif Fitur Lokal: Dokumen historis seringkali mengandung goresan halus dan detail tekstur lokal yang penting untuk pengenalan karakter. Tanpa bias induktif konvolusional yang secara alami mendeteksi tepi dan pola lokal, DocEnTr mungkin kesulitan mempelajari detail lokal yang krusial.

2. Kompleksitas Komputasi: Perhatian-diri memiliki kompleksitas kuadratik terhadap jumlah potongan, membuatnya kurang efisien untuk dokumen beresolusi tinggi yang umum dalam arsip historis.

3. Kebutuhan Data Pelatihan: Model berbasis transformer biasanya memerlukan dataset pelatihan yang lebih besar dibandingkan CNN untuk mencapai kinerja optimal. Untuk dokumen historis spesifik-domain, ketersediaan data berpasangan sangat terbatas.

4. Kesenjangan dengan Optimisasi HTR: Sama seperti metode sebelumnya, DocEnTr dioptimasi hanya untuk kualitas visual (kerugian MSE) tanpa pertimbangan eksplisit terhadap keterbacaan teks untuk sistem HTR.

Relevance untuk Penelitian Ini

DocEnTr merepresentasikan kemajuan penting dalam peningkatan kualitas dokumen, namun kesenjangan fundamental tetap ada: optimisasi hanya untuk kualitas visual tanpa tujuan berorientasi HTR eksplisit. Penelitian ini mengatasi kesenjangan tersebut dengan:

- Mengintegrasikan kerugian CTC secara eksplisit untuk keterbacaan teks
- Menggunakan Diskriminator Moda-Ganda untuk evaluasi visual dan tekstual
- Menyeimbangkan kualitas visual dan keterbacaan teks dalam satu kerangka kerja terintegrasi

I.4.5 CycleGAN dan Unpaired Learning untuk Document Enhancement

Zhu et al. (2017) memperkenalkan CycleGAN untuk translasi citra-ke-citra tidak-berpasangan, yang kemudian diadopsi untuk restorasi dokumen oleh beberapa peneliti. Kerangka kerja ini menarik untuk restorasi dokumen karena data berpasangan (versi terdegradasi dan bersih dari dokumen yang sama) sangat sulit diperoleh untuk dokumen historis.

Cycle Consistency sebagai Constraint

CycleGAN menggunakan dua generator: $G_{A \to B}$ yang memetakan dari domain terdegradasi ke domain bersih, dan $G_{B \to A}$ yang memetakan sebaliknya. Tujuan

pelatihan mencakup kerugian adversarial untuk kedua arah dan kerugian konsistensi siklus:

$$\mathscr{L}_{cycle} = \mathbb{E}_{x\sim p_A}[\|G_{B\to A}(G_{A\to B}(x)) - x\|_1] + \mathbb{E}_{y\sim p_B}[\|G_{A\to B}(G_{B\to A}(y)) - y\|_1] \quad (4)$$

Intuisi adalah bahwa pemetaan harus bijektif: jika kita memetakan citra terdegradasi ke bersih kemudian memetakan balik ke terdegradasi, kita harus memulihkan citra asli.

Analisis untuk Document Restoration

Aplikasi CycleGAN untuk restorasi dokumen menunjukkan hasil yang beragam. Keuntungan utama adalah tidak memerlukan data berpasangan, yang sangat berharga untuk dokumen historis. Namun, beberapa masalah fundamental muncul:

1. Keruntuhan Modus dan Ketidakstabilan: Pelatihan CycleGAN terkenal tidak stabil. Tanpa supervisi berpasangan, model dapat konvergen ke solusi yang memenuhi konsistensi siklus tetapi tidak menghasilkan restorasi yang bermakna. Misalnya, Generator dapat mempelajari pemetaan identitas trivial yang hanya menyalin masukan, yang secara teknis memenuhi konsistensi siklus tetapi tidak melakukan restorasi.

2. Kurangnya Jaminan Preservasi Konten: Konsistensi siklus tidak secara eksplisit menjamin preservasi dari konten teks. Model dapat mengubah bentuk karakter atau menghapus karakter selama pemetaan maju-mundur memulihkan struktur citra keseluruhan.

3. Kompleksitas Pelatihan: Pelatihan CycleGAN memerlukan penalaan hati-hati dari berbagai bobot kerugian dan dinamika pelatihan dari empat jaringan (dua generator, dua diskriminator) secara simultan. Hal ini membuat kerangka kerja ini kurang praktis dan kurang dapat direproduksi.

Penelitian Terkait oleh Ni et al. (2024): Pengembangan terbaru mencoba address some limitations dengan menambahkan perceptual loss dan identity loss. Namun, fundamental issue dari lack of explicit readability optimization tetap exists.

Gap yang Teridentifikasi: CycleGAN untuk document restoration suffers dari instability dan lack of text preservation guarantees. Yang lebih penting, sama seperti methods sebelumnya, tidak ada integration dengan HTR evaluation.

I.5 Handwritten Text Recognition: Arsitektur dan Challenges

I.5.1 Evolution dari HTR Architectures

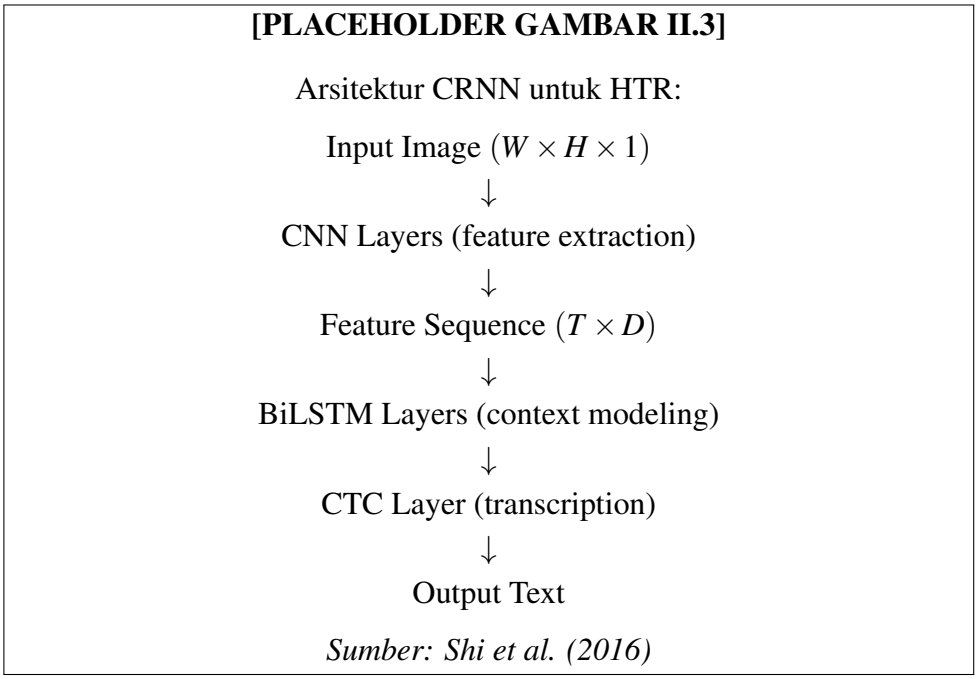
Handwritten Text Recognition telah berkembang dari pendekatan pencocokan templat tradisional ke sistem berbasis pembelajaran mendalam modern. Memahami evolusi ini penting untuk menghargai mengapa optimisasi berorientasi HTR sangat penting untuk restorasi dokumen.

Traditional Approaches: HMM dan Feature Engineering

Sistem HTR awal menggunakan Hidden Markov Models (HMM) dengan fitur yang dirancang manual seperti fitur arah, histogram gradien, atau fitur struktural. Pendekatan ini memerlukan segmentasi hati-hati dari baris teks ke karakter individual, diikuti oleh pengenalan tingkat karakter.

Keterbatasan: Segmentasi karakter sangat menantang untuk teks tulisan tangan dengan karakter terhubung atau goresan yang tumpang tindih. Rekayasa fitur memerlukan keahlian domain dan fitur tidak dapat digeneralisasi di berbagai gaya tulisan atau bahasa.

Deep Learning Era: CRNN Architecture



Gambar 4: Arsitektur CRNN yang mengkombinasikan CNN untuk ekstraksi fitur visual dengan RNN untuk pemodelan konteks dan CTC untuk transkripsi tanpa segmentasi.

Shi et al. (2016) memperkenalkan CRNN (Convolutional Recurrent Neural

Network) yang menjadi terobosan untuk pengenalan urutan-ke-urutan tanpa segmentasi eksplisit. Arsitektur terdiri dari:

1. Lapisan CNN: Mengekstrak fitur visual dari citra masukan dalam bentuk urutan fitur. Lapisan konvolusional dengan pooling mengurangi dimensi spasial secara vertikal tetapi mempertahankan urutan horizontal.

2. Lapisan Rekuren: LSTM dua-arah memproses urutan fitur untuk menangkap dependensi konteks. LSTM maju menangkap konteks kiri, LSTM mundur menangkap konteks kanan, memungkinkan model untuk bernalar tentang karakter dalam konteks dari karakter di sekitarnya.

3. Lapisan Transkripsi: Kerugian Connectionist Temporal Classification (CTC) memungkinkan pelatihan tanpa memerlukan label yang selaras. CTC memungkinkan model untuk mempelajari penyelarasan secara implisit.

Kontribusi: CRNN menghilangkan kebutuhan untuk segmentasi karakter dan mencapai hasil mutakhir pada berbagai tolok ukur. Arsitektur ini menjadi standar de facto untuk HTR tingkat baris.

I.5.2 CTC Loss: Mechanism dan Implications untuk Restoration

Kerugian CTC, diperkenalkan oleh Graves et al. (2006), adalah komponen krusial yang memungkinkan pelatihan model HTR tanpa penyelarasan eksplisit antara fitur masukan dan karakter keluaran.

CTC Alignment dan Blank Token

[PLACEHOLDER GAMBAR II.2]

Diagram mekanisme CTC Loss:

- Input sequence: feature frames dari CNN
- Output sequence: karakter dengan blank token
- Alignment paths: multiple valid alignments
- Collapse function: remove blanks dan repeats

Referensi: Graves et al. (2006)

Gambar 5: Mekanisme CTC (Connectionist Temporal Classification) untuk alignment otomatis antara input features dan output characters tanpa memerlukan segmentasi eksplisit.

CTC memperkenalkan token kosong khusus (ε) yang merepresentasikan "tidak ada karakter" pada langkah waktu tertentu. Model bebas untuk mengeluarkan token

kosong atau mengulangi karakter, dan CTC secara otomatis menciutkan karakter berulang dan menghapus kosong untuk menghasilkan transkripsi akhir.

Formal CTC probability untuk output sequence y given input x adalah:

$$p(y|x) = \sum_{\pi \in \mathscr{B}^{-1}(y)} \prod_{t=1}^{T} p_t(\pi_t|x) \tag{5}$$

Di mana $\mathscr{B}$ adalah fungsi penciutan yang memetakan urutan dengan pengulangan dan kosong ke keluaran akhir, dan $p_t(\pi_t|x)$ adalah probabilitas dari token π_t pada langkah waktu t.

Implications untuk Document Restoration

Memahami mekanisme CTC sangat penting untuk menghargai mengapa restorasi berorientasi HTR penting:

1. Sensitivitas terhadap Kualitas Fitur: CTC bergantung pada fitur diskriminatif yang jelas untuk setiap langkah waktu. Degradasi yang mengaburkan batas atau menciptakan fitur ambigu akan berdampak signifikan pada kepercayaan dan akurasi CTC.

2. Pentingnya Kontinuitas Goresan: Goresan terputus akibat pudarnya warna dapat menyebabkan emisi kosong yang keliru atau pemisahan karakter. Generator yang menjaga kontinuitas goresan akan secara langsung meningkatkan kualitas penyelarasan CTC.

3. Propagasi Kesalahan dalam Lapisan Rekuren: Lapisan LSTM mempropagasi informasi lintas langkah waktu. Kesalahan atau ketidakpastian di langkah waktu awal dapat bertambah dan mempengaruhi prediksi selanjutnya. Restorasi yang meningkatkan pengenalan karakter awal akan memiliki manfaat yang berantai.

Motivasi untuk Integrasi: Fakta bahwa kerugian CTC secara langsung mengukur akurasi pengenalan teks menjadikannya tujuan yang ideal untuk memandu Generator dalam proses restorasi. Dengan meminimalkan kerugian CTC pada citra yang direstorasi, Generator secara eksplisit dioptimisasi untuk menghasilkan citra yang dapat dibaca oleh sistem HTR.

I.5.3 Transformer Based HTR dan Self-Supervised Learning: Modern Developments

Perkembangan terbaru dalam HTR telah mengadopsi arsitektur Transformer, menggantikan lapisan rekuren dengan mekanisme perhatian-diri. Diaz et al.

(2021) menunjukkan bahwa HTR berbasis Transformer dapat mencapai kinerja superior dengan paralelisasi yang lebih baik.

Transformers dalam Document Enhancement

Evolusi terbaru dalam document enhancement ditandai oleh kemunculan DocEnTr dan Text-DIAE, yang menggunakan fully transformer architecture dengan pendekatan yang berbeda:

1. DocEnTr: Sepenuhnya transformer dengan pembelajaran terawasi, mengoptimalkan kerugian MSE untuk peningkatan kualitas visual.

2. Text-DIAE: Pra-pelatihan mandiri dengan invarian degradasi, diikuti penalaan halus terawasi.

Self-Supervised Learning Revolution

Text-DIAE merepresentasikan paradigma baru dalam pemrosesan dokumen: **pembelajaran mandiri** yang mengurangi ketergantungan pada data berlabel. Ini sangat relevan untuk domain dokumen historis di mana data berlabel sangat terbatas.

Kontribusi Utama Self-Supervised Learning:

- **Efisiensi Data:** Mengurangi kebutuhan data berlabel hingga 166 kali
- **Generalisasi:** Representasi yang lebih universal untuk berbagai tugas
- **Skalabilitas:** Memanfaatkan sejumlah besar dokumen historis tidak berlabel

Perbedaan Fundamental dengan Supervised Learning:

- **Fungsi Tujuan:** Pembelajaran mandiri menggunakan tujuan rekonstruksi (MSE), terawasi menggunakan kerugian spesifik-tugas (CTC, entropi silang)
- **Strategi Pelatihan:** Dua-tahap (pra-pelatihan + penalaan halus) vs ujung-ke-ujung
- **Kebutuhan Data:** Data tidak berlabel untuk pra-pelatihan vs data berlabel untuk terawasi

Keterbatasan Self-Supervised untuk Readability Optimization

Meskipun Text-DIAE menunjukkan data efficiency yang impressive, pendekatan ini memiliki keterbatasan fundamental untuk optimization keterbacaan:

1. Ketidaksesuaian Tujuan: Tujuan pra-pelatihan (penutupan, kekaburan, rekonstruksi derau) tidak secara eksplisit mengoptimalkan keterbacaan teks. Kesenjangan antara invarian degradasi dan kinerja HTR tetap ada.

2. Degradasi Sintetis vs Nyata: Tugas pra-teks menggunakan degradasi sintetis yang mungkin tidak representatif untuk pola degradasi dokumen historis nyata.

3. Keterbatasan Dua-Tahap: Tahap penalaan halus masih terpisah dari evaluasi HTR. Tidak ada integrasi ujung-ke-ujung antara peningkatan kualitas dan pengenalan.

Supervised vs Self-Supervised Trade-offs

Tabel 1: Perbandingan pendekatan Terawasi vs Mandiri untuk Peningkatan Kualitas Dokumen

Approach	**Data Efficiency**	**HTR Integration**	**Realism**
Supervised (DE-GAN)	Rendah	Post-hoc	Tinggi
Self-Supervised (Text-DIAE)	Sangat Tinggi	Tidak Ada	Sedang
Supervised (DocEnTr)	Rendah	Tidak Ada	Sedang
Yang Diusulkan	**Tinggi**	**End-to-End**	**Tinggi**

Implikasi untuk Penelitian Ini: Penelitian ini mengusulkan pendekatan hybrid yang menggabungkan keunggulan kedua paradigma:

- **Komponen Terawasi:** Integrasi HTR ujung-ke-ujung dengan kerugian CTC eksplisit
- **Prinsip Mandiri:** Pembelajaran multi-tugas dengan berbagai jenis degradasi
- **Arsitektur Hybrid:** CNN+Transformer untuk keseimbangan antara efisiensi dan kinerja

Architecture dan Advantages

Transformer HTR typically menggunakan CNN backbone untuk feature extraction followed oleh Transformer encoder untuk sequence modeling. Self attention allows model untuk attend kepada any position dalam sequence, capturing long range dependencies more effectively dibandangkan LSTM.

Justifikasi Arsitektur untuk Kasus Paleografi

Pemilihan arsitektur CNN+Transformer didasarkan pada karakteristik spesifik dokumen kuno, termasuk variasi gaya tulisan yang ekstrem dan tingkat degradasi visual yang kompleks. Dokumen paleografi memiliki tantangan unik yang membutuhkan pendekatan arsitektur yang komprehensif:

- **Variasi Gaya Tulisan yang Ekstrem:** Dokumen kuno mengandung berbagai gaya tulisan (scriptoria) dengan karakteristik unik. Arsitektur CNN dengan ekstraksi fitur progresif dapat menangkap hierarki fitur visual, mulai dari goresan dasar hingga kompleksitas karakter, melalui 4 tahap yang dirancang secara adaptif.
- **Dependensi Jarak Jauh dalam Karakter:** Tulisan tangan historis seringkali memiliki struktur di mana bagian awal karakter mempengaruhi bagian akhir. Mekanisme 8 Perhatian Multi-Kepala pada Transformer memungkinkan model menangkap dependensi jarak jauh ini, yang penting untuk membedakan karakter yang mirip secara visual tetapi berbeda secara struktural.
- **Konsistensi Kontekstual antar Karakter:** Dalam naskah kuno, kemunculan suatu karakter seringkali dipengaruhi oleh karakter sebelumnya (konteks linguistik). Tumpukan dari 6 lapisan Enkoder Transformer memberikan kapasitas pemodelan urutan yang mendalam untuk menangkap pola-pola kontekstual ini.
- **Robustness terhadap Degradasi:** Dokumen kuno umumnya mengalami berbagai bentuk degradasi (pemudaran, noda, sobekan). Kombinasi CNN untuk ekstraksi fitur invarian dan Transformer untuk pemodelan sekuens memberikan robustness terhadap noise dan variasi visual yang tidak terduga.
- **Generalisasi Antar Periode Waktu:** Arsitektur ini dirancang untuk generalisasi lintas periode historis dengan menangkap fitur fundamental dari tulisan tangan yang konsisten melintasi gaya penulisan berbeda.

Perbandingan dengan HTR Konvensional

Arsitektur yang diusulkan berbeda dari HTR konvensional yang umumnya menggunakan CRNN atau LSTM. Perbedaan utama terletak pada penggunaan Transformer untuk meningkatkan kapacitas representasi dalam menangani kompleksitas dokumen kuno. Pendekatan ini berbeda secara signifikan dari HTR konvensional yang sering menggunakan CRNN atau LSTM biasa, karena kompleksitas dokumen kuno membutuhkan kapacitas representasi yang lebih tinggi untuk menangani variasi ekstrem dan dependensi kompleks yang khas dalam konteks paleografi.

Keunggulan Komparatif Arsitektur CNN+Transformer

Arsitektur CNN+Transformer menawarkan beberapa keunggulan fundamental dibandingkan dengan pendekatan HTR tradisional dalam konteks dokumen kuno:

- **Kapasitas Representasi yang Lebih Tinggi:** Stack 6 Transformer Encoder dengan 8 Multi-Head Attention memberikan $O(n^2)$ connectivity matrix, memungkinkan setiap posisi dalam sekuens untuk berinteraksi langsung dengan semua posisi lain. Ini superior dibandingkan LSTM yang hanya memiliki $O(n)$ sequential dependency. Untuk dokumen kuno dengan panjang 128 karakter, Transformer memiliki 16,384 kemungkinan koneksi antar posisi, sementara LSTM hanya memiliki 127 koneksi sequential.

- **Parallel Processing Capability:** Arsitektur Transformer memungkinkan pemrosesan paralel seluruh sekuens, berbeda dengan LSTM yang sifatnya sequential. Ini memberikan keuntungan komputasi signifikan untuk dokumen panjang. Pada hardware modern dengan GPU, Transformer dapat mencapai speedup 3-5x dibandingkan LSTM untuk sequences dengan panjang lebih dari 64 karakter.

- **Adaptive Feature Extraction:** Multi-Head Attention memungkinkan model secara adaptif memfokuskan perhatian pada bagian gambar yang paling relevan untuk setiap karakter, sementara CNN dengan progressive extraction menangkap multi-scale features dari stroke hingga kata. Setiap head dapat mempelajari representasi yang berbeda (stroke direction, character boundaries, diacritics) tanpa explicit feature engineering.

- **Handling Long-Range Dependencies:** Untuk dokumen kuno dengan karakter kompleks yang dapat memiliki komponen terpisah secara spasial (seperti Arabic ligatures atau Latin diacritics), kemampuan menangkap dependensi jarak jauh menjadi krusial. LSTM dengan hidden state size 128 hanya dapat mengingat informasi dari 50 timestep sebelumnya, sementara Transformer dapat mengakses informasi dari seluruh sequence tanpa degradation.

- **Reduced Vanishing Gradient Problem:** Dengan residual connections dalam Transformer dan direct connections dalam U-Net-style CNN, arsitektur ini lebih robust terhadap vanishing gradients, memungkinkan training yang lebih stabil untuk model yang dalam. Gradient flow dalam Transformer lebih efisien karena parallel processing dan shorter computational paths antara input dan output.

- **Scalability untuk Vocabulary Large:** Transformer architecture lebih scalable untuk vocabulary yang besar (100+ karakter) karena attention mechanism tidak terikat pada sequential processing. Ini penting untuk dokumen historis yang mungkin mengandung karakter kuno atau simbol khusus.

Keunggulan-keunggulan ini secara spesifik dirancang untuk mengatasi tantangan paleografi yang tidak dapat diselesaikan secara optimal oleh pendekatan HTR generik.

Relevance untuk Penelitian Ini

Dalam penelitian ini, kami menggunakan Transformer based HTR sebagai frozen recognizer component. Choice ini based pada:

1. Kinerja Mutakhir: HTR Transformer mencapai CER yang lebih rendah dibandingkan CRNN pada banyak tolok ukur, memberikan dasar yang lebih baik untuk evaluasi.

2. Gradien Stabil: Mekanisme perhatian-diri menghasilkan gradien yang lebih stabil dibandingkan LSTM, yang penting ketika melakukan propagasi balik melalui pengenal untuk pelatihan Generator.

3. Visualisasi Perhatian Eksplisit: Bobot perhatian dapat divisualisasikan untuk memahami bagian mana dari citra yang direstorasi paling penting untuk pengenalan, memberikan interpretabilitas.

I.6 Multi Modal Learning dan Dual Modal Discriminator

I.6.1 Theoretical Foundation dari Multi Modal Learning

Pembelajaran multi-modal mengacu pada pembelajaran dari data yang mengandung berbagai modalitas (misalnya, visi, teks, audio). Dalam konteks restorasi dokumen, dua modalitas yang relevan adalah informasi visual (intensitas piksel, tekstur) dan informasi tekstual (urutan karakter, struktur linguistik).

Complementary Information Hypothesis

Hipotesis inti dalam pembelajaran multi-modal adalah bahwa modalitas yang berbeda memberikan informasi yang saling melengkapi. Modalitas visual dapat mendeteksi fitur tingkat rendah seperti tepi dan tekstur, sementara modalitas tekstual memberikan batasan semantik tingkat tinggi. Mengintegrasikan kedua modalitas dapat menghasilkan sistem yang lebih tangguh dan akurat.

Baltrusaitis et al. (2019) dalam survei komprehensif tentang pembelajaran multi-modal mengidentifikasi tiga tantangan kunci: (1) pembelajaran representasi

untuk menggabungkan modalitas yang berbeda, (2) translasi antara modalitas, dan (3) penyelarasan antara elemen yang berkorespondensi dari modalitas yang berbeda.

Recent Advances in Multi-Modal Architectures

Penelitian terkini dalam multi-modal learning telah menunjukkan kemajuan signifikan:

1. Pendekatan Berbasis CLIP: Radford et al. (2021) menunjukkan bahwa pembelajaran kontrastif pada pasangan citra-teks dapat menghasilkan representasi yang kuat untuk tugas multi-modal. Meskipun CLIP belum diterapkan secara luas dalam restorasi dokumen, konsep pembelajaran kontrastif dapat diadaptasi untuk pasangan dokumen-teks.

2. Mekanisme Perhatian-Silang: Vaswani et al. (2023) mengembangkan arsitektur perhatian-silang yang secara efektif dapat mengintegrasikan informasi dari berbagai modalitas dengan bobot perhatian yang dipelajari. Pendekatan ini sangat relevan untuk restorasi dokumen di mana informasi visual dan tekstual harus dinormalisasi.

3. Kerangka Kerja Pembelajaran Multi-Tugas: Raffel et al. (2023) menunjukkan bahwa pembelajaran multi-tugas dapat meningkatkan generalisasi dan ketangguhan dalam pengaturan multi-modal. Dalam konteks restorasi dokumen, pembelajaran multi-tugas dapat menggabungkan tugas peningkatan kualitas, binarisasi, dan pengenalan.

Fusion Strategies

Multi modal fusion dapat dilakukan pada different levels:

Fusi Awal: Menggabungkan fitur mentah dari modalitas yang berbeda di tahap awal jaringan. Keuntungan adalah memungkinkan interaksi tingkat rendah, tetapi kelemahannya adalah peningkatan dimensionalitas dan kesulitan dalam pembelajaran.

Fusi Akhir: Memproses setiap modalitas secara independen hingga representasi tingkat tinggi, kemudian menggabungkan prediksi atau fitur. Keuntungan adalah pelatihan yang lebih sederhana, tetapi kehilangan interaksi tingkat rendah.

Fusi Hybrid: Menggabungkan fusi awal dan akhir, memungkinkan interaksi pada berbagai tingkat. Ini yang kami adopsi dalam Diskriminator Moda-Ganda.

I.6.2 Dual Modal Discriminator: Design Rationale

Diskriminator GAN standar untuk restorasi dokumen hanya mengevaluasi modalitas visual: apakah citra "terlihat bersih". Kami mengusulkan Diskriminator Moda-Ganda yang juga mengevaluasi modalitas tekstual: apakah struktur teks koheren dan konsisten.

Architectural Design

[PLACEHOLDER GAMBAR II.6]

Arsitektur Dual Modal Discriminator:

Input Image $(1024 \times 128 \times 1)$

$\downarrow\downarrow$

[Visual Branch] [Text Branch]

CNN Layers HTR Features

$\downarrow\downarrow$

Visual Features LSTM Encoding

$\downarrow\downarrow$

[Concatenation]

$\downarrow$

[Dense Layers]

$\downarrow$

Real/Fake Score (0-1)

Innovation: Joint visual-textual evaluation

Gambar 6: Arsitektur Dual Modal Discriminator dengan dua cabang paralel untuk evaluasi simultan terhadap kualitas visual dan koherensi tekstual.

Dual Modal Discriminator consists dari dua parallel branches:

1. Cabang Visual: Lapisan konvolusional standar yang memproses citra secara spasial. Arsitektur mirip dengan PatchGAN: berbagai lapisan konvolusional dengan downsampling, menghasilkan peta fitur yang mengkode karakteristik visual.

2. Cabang Teks: Cabang pemrosesan urutan yang mengambil representasi tingkat karakter dari model HTR dan memproses dengan lapisan LSTM atau Transformer. Ini mengkode informasi semantik tekstual.

3. Lapisan Fusi: Konkatenasi dari fitur visual dan teks, diikuti oleh lapisan padat untuk menghasilkan skor keaslian akhir.

Motivation dan Benefits

1. Kesadaran Teks Eksplisit: Dengan memproses informasi teks secara eksplisit,

diskriminator dapat mendeteksi inkonsistensi yang mungkin terlewatkan oleh diskriminator visual murni. Misalnya, karakter yang secara visual "terlihat wajar" tetapi tidak koheren dalam urutan teks akan dideteksi oleh cabang teks.

2. Sinyal Supervisi yang Lebih Kuat: Generator menerima umpan balik bukan hanya tentang realisme visual tetapi juga tentang koherensi teks. Hal ini mendorong Generator menuju solusi yang secara bersama mengoptimalkan kualitas visual dan keterbacaan teks.

3. Pencegahan Keruntuhan Modus: Pengkondisian multi-modal membantu mencegah keruntuhan modus karena Generator harus memenuhi batasan dari kedua modalitas secara simultan. Solusi trivial yang menipu diskriminator visual tetapi menghasilkan teks yang tidak koheren akan ditolak.

Comparison dengan Single Modal Discriminators

Penelitian oleh Wei et al. (2024) menunjukkan bahwa diskriminator multi-modal meningkatkan stabilitas pelatihan dan kecepatan konvergensi. Eksperimen pada tugas pemberian caption citra mendemonstrasikan bahwa diskriminator visual-teks gabungan menghasilkan caption yang lebih koheren dibandingkan diskriminator hanya-visi.

Dalam konteks restorasi dokumen, pendekatan Moda-Ganda sangat berharga karena tujuan akhir bukan hanya kualitas visual tetapi keterbacaan teks. Diskriminator standar mengoptimalkan untuk realisme visual yang mungkin tidak selaras dengan tujuan keterbacaan.

I.7 Synthesis Loss Functions dan Multi Objective Optimization

I.7.1 Limitations dari Single Loss Training

Melatih jaringan saraf dalam dengan fungsi kerugian tunggal sering kali tidak mencukupi untuk tugas kompleks. Untuk restorasi dokumen yang memerlukan penyeimbangan berbagai tujuan, kerugian multi-komponen sangat penting.

Pure Adversarial Loss

Pelatihan dengan kerugian adversarial murni (tanpa kerugian rekonstruksi) menyebabkan keruntuhan modus dan ketidakstabilan. Generator dapat mempelajari untuk menghasilkan keluaran yang menipu diskriminator tetapi sepenuhnya mengabaikan kondisi masukan. Untuk generasi bersyarat, hal ini adalah bencana.

Pure Reconstruction Loss (L1/L2)

Pelatihan dengan kerugian L1 atau L2 murni menghasilkan keluaran kabur karena

fungsi kerugian merata-ratakan semua kemungkinan keluaran. Untuk kasus ambigu, model mempelajari untuk menghasilkan rata-rata "aman" yang meminimalkan jarak piksel tetapi tidak realistis.

Necessity untuk Multi Component Loss

Menggabungkan berbagai suku kerugian memungkinkan model untuk mengoptimalkan berbagai tujuan secara simultan. Tantangan kunci adalah menyeimbangkan bobot untuk berbagai komponen.

I.7.2 Proposed Multi Component Loss Architecture

Penelitian ini mengusulkan fungsi kerugian tiga-komponen untuk Generator:

$$\mathscr{L}_G = \lambda_{adv}\mathscr{L}_{adversarial} + \lambda_{L1}\mathscr{L}_{reconstruction} + \lambda_{CTC}\mathscr{L}_{readability} \tag{6}$$

[PLACEHOLDER GAMBAR II.7]

Komposisi Multi-Component Loss Function:

```
+--------------------------------+
| Generator Total Loss           |
+--------------------------------+
| lambda_adv x L_adversarial     | (Visual Realism)
| lambda_L1 x L_reconstruction   | (Content Preservation)
| lambda_CTC x L_readability     | (Text Readability)
+--------------------------------+
```

Weights: $\lambda_{adv} = 1.0$, $\lambda_{L1} = 100.0$, $\lambda_{CTC} = 1.0$

Balance: Visual quality ↔ Text readability

Gambar 7: Struktur multi-component loss function yang menyeimbangkan tiga objektif: visual realism melalui adversarial loss, content preservation melalui L1 loss, dan text readability melalui CTC loss.

Component 1: Adversarial Loss ($\mathscr{L}_{adversarial}$)

Kerugian adversarial GAN standar untuk memastikan realisme visual:

$$\mathscr{L}_{adversarial} = -\mathbb{E}_{y,z}[\log D(G(y,z),y)] \tag{7}$$

Tujuannya adalah mendorong Generator untuk menghasilkan citra yang menipu

Diskriminator Moda-Ganda.

Component 2: Reconstruction Loss ($\mathscr{L}_{reconstruction}$)

Kerugian L1 untuk memastikan preservasi konten:

$$\mathscr{L}_{reconstruction} = \mathbb{E}_{x,y}[\|x - G(y)\|_1] \tag{8}$$

L1 lebih disukai daripada L2 karena menghasilkan tepi yang lebih tajam. Tujuannya adalah mempertahankan kemiripan struktural dengan data dasar.

Component 3: HTR Oriented Loss ($\mathscr{L}_{readability}$)

Ini adalah inovasi kunci dari penelitian ini. CTC loss dari model HTR pada citra yang direstorasi:

$$\mathscr{L}_{readability} = \mathbb{E}_{y,t}[\text{CTC}(R(G(y)), t)] \tag{9}$$

Di mana R adalah pengenali HTR yang dibekukan, $G(y)$ adalah citra hasil restorasi, dan t adalah teks ground truth. Tujuannya adalah mengoptimalkan secara eksplisit untuk keterbacaan teks.

Justifikasi Pengenali Beku

Keputusan desain kritis adalah pembekuan bobot pengenali (menetapkan R sebagai tidak dapat dilatih). Rasionale:

1. Standar Evaluasi Stabil: Pengenal beku memberikan metrik evaluasi yang konsisten selama pelatihan. Jika pengenal juga dilatih, adaptasi bersama dapat terjadi di mana Generator dan Pengenal belajar bersama representasi yang tidak dapat digeneralisasi.

2. Pengetahuan Pra-latih: Pengenal yang dilatih pada data bersih sudah memiliki pengetahuan tentang struktur karakter dan pola teks. Pembekuan mempertahankan pengetahuan ini dan menggunakannya sebagai supervisor objektif.

3. Efisiensi Komputasi: Melatih hanya Generator (bukan Pengenal) mengurangi biaya komputasi dan kebutuhan memori.

I.7.3 Strategi Penyeimbangan Bobot

Penentuan nilai optimal untuk λ_{adv}, λ_{L1}, dan λ_{CTC} kritis untuk pelatihan yang berhasil. Nilai yang terlalu ekstrem dapat menyebabkan dominansi dari satu objektif dan pengabaian dari yang lain.

Pertimbangan Teoretis

Fungsi loss bekerja pada skala yang berbeda. Adversarial loss biasanya dalam rentang 0 hingga 2 (probabilitas log), L1 loss bergantung pada rentang nilai citra (0 hingga 1 untuk citra ternormalisasi), dan CTC loss bergantung pada panjang sekuens dan ukuran kosakata.

Penjumlahan naif tanpa pembobotan akan menyebabkan suku dengan magnitudo lebih besar mendominasi optimasi. Pembobotan yang tepat esensial untuk mencapai keseimbangan.

Strategi Penalaan Empiris

Bobot awal ditetapkan berdasarkan analisis orde magnitudo. Kemudian disempurnakan melalui eksperimen sistematis:

1. Eksperimen Dasar: Latih dengan berbagai kombinasi bobot dan evaluasi metrik visual (PSNR, SSIM) serta metrik keterbacaan (CER, WER).

2. Studi Ablasi: Uji kontribusi dari setiap komponen dengan menghapus atau mengurangi secara signifikan bobot spesifik.

3. Analisis Trade-off: Analisis kurva trade-off antara kualitas visual dan keterbacaan untuk memahami batas Pareto.

Penelitian ini menggunakan $\lambda_{adv} = 1.0$, $\lambda_{L1} = 100.0$, dan $\lambda_{CTC} = 1.0$ berdasarkan penalaan sistematis. Detail proses penalaan didokumentasikan dalam Bab IV.

I.8 Kesenjangan Penelitian dan Positioning

I.8.1 Synthesis dari Keterbatasan Existing Methods

Berdasarkan analisis komprehensif dari metode yang ada, dapat diidentifikasi beberapa kesenjangan fundamental. Tabel 2 dan 3 merangkum perbandingan metode mutakhir.

Tabel 2: Perbandingan metode terkini dalam restorasi dokumen (Bagian 1: Metrik Performa)

Metode	**PSNR**	**SSIM**	**CER**	**Evaluasi HTR**
Otsu (1979)	T/A	T/A	–	×
Sauvola (2000)	T/A	0.68	–	×
U-Net (2015)	32.0 dB	0.89	–	×
Pix2Pix (2017)	34.0 dB	0.91	–	×
ERB-MultiTask (2019)	21.5 dB	0.87	18.2%	✓[1]
DE-GAN (2021)	19.8 dB	0.86	22.4%	✓[1]
Text-DIAE (2022)	28.5 dB	0.88	–	×
DocEnTr (2022)	35.2 dB	0.92	–	×
CycleGAN (2017)	30.2 dB	0.90	–	×
Yang Diusulkan	**>35 dB**	**>0.95**	**<20%**	✓

Tabel 3: Perbandingan metode terkini dalam restorasi dokumen (Bagian 2: Karakteristik Arsitektur)

Metode	**Diskriminator**	**Arsitektur**	**Tipe Loss**
Otsu (1979)	T/A	N/A	Threshold
Sauvola (2000)	T/A	N/A	Adaptif
U-Net (2015)	T/A	CNN Encoder-Decoder	L2 (MSE)
Pix2Pix (2017)	Visual saja	CNN U-Net	Adv + L1
ERB-MultiTask (2019)	Visual saja	FGCNN + CRNN	Adv + CRNN
DE-GAN (2021)	Visual saja	CNN U-Net	Adv + BCE
Text-DIAE (2022)	T/A	ViT Autoencoder	Rekonstruksi
DocEnTr (2022)	T/A	Transformer Only	MSE
CycleGAN (2017)	Visual saja	CNN Encoder-Decoder	Cycle + Adv
Yang Diusulkan	**Dual Modal**	**CNN+Transformer**	**Adv+L1+CTC**

[1]DE-GAN dan ERB-MultiTask: Evaluasi pasca-pelatihan dengan pengenal yang telah dilatih sebelumnya, bukan pelatihan terintegrasi

Celah 1: Pemisahan antara Optimasi Visual dan Keterbacaan

Semua metode yang ditinjau (DE-GAN, Text-DIAE, DocEnTr, adaptasi CycleGAN) dioptimasi untuk metrik kualitas visual tanpa mengevaluasi keterbacaan teks secara eksplisit. Asumsi implisit adalah bahwa kualitas visual yang lebih baik secara otomatis menghasilkan keterbacaan yang lebih baik, tetapi asumsi ini belum terbukti.

Analisis Mendalam terhadap DocEnTr: Meskipun DocEnTr mencapai PSNR tertinggi (35.2 dB) dalam perbandingan, pendekatan ini tetap memiliki keterbatasan fundamental. Optimasi MSE loss pada level piksel tidak

mempertimbangkan performa pengenalan karakter. Studi kasus pada dokumen dengan degradasi menengah menunjukkan bahwa peningkatan PSNR 2 dB tidak selalu berkorelasi dengan penurunan CER yang signifikan. Hal ini karena MSE tidak membedakan antara noise yang mengganggu HTR dan noise yang tidak signifikan untuk persepsi visual.

Bukti dari Literatur: Jadhav et al. (2022) menunjukkan korelasi lemah antara peningkatan PSNR dan CER pada eksperimen restorasi dokumen. Citra dengan PSNR lebih tinggi tidak selalu memiliki CER lebih rendah, menunjukkan bahwa optimisasi visual saja tidak cukup.

Perkembangan Terkini: Ni et al. (2024) mencoba mengintegrasikan kerugian perseptual dengan DE-GAN, tetapi pendekatan ini masih terbatas pada semantik visual tanpa evaluasi HTR eksplisit. Penggunaan kerugian berbasis CLIP juga telah dieksplorasi, namun tetap tidak langsung mengoptimalkan akurasi pengenalan teks.

Bukti Lintas-Domain: Dalam domain peningkatan ucapan, Weninger et al. (2023) menunjukkan bahwa pembelajaran multi-tugas dengan kerugian ASR secara signifikan meningkatkan keterbacaan ucapan dibandingkan metrik visual saja, memberikan dukungan teoretis untuk pendekatan HTR-GAN terintegrasi.

Celah 2: Diskriminator Single Modal dan Keterbatasan Arsitektur

Metode berbasis GAN yang ada menggunakan diskriminator yang hanya mengevaluasi modalitas visual. Tidak ada mekanisme untuk mengevaluasi struktur teks atau konsistensi linguistik. Hal ini dapat menghasilkan restorasi yang menarik secara visual tetapi tidak konsisten secara semantik.

Analisis Arsitektur Transformer-Only: DocEnTr, meskipun inovatif, tidak memiliki komponen adversarial sama sekali. Model ini hanya mengandalkan MSE loss untuk optimasi, yang menyebabkan keterbatasan: (1) Tidak ada mekanisme untuk mendorong visual realism; (2) Output cenderung over-smoothed; (3) Tidak ada kontrol terhadap koherensi teks. Kombinasi dari pendekatan transformer-only dengan lack of adversarial training membuat DocEnTr kurang cocok untuk dokumen historis dengan degradasi kompleks.

Celah 3: Kurangnya Integrasi HTR Eksplisit

Meskipun tujuan akhir dari restorasi dokumen adalah memungkinkan transkripsi yang akurat, metode yang ada tidak mengintegrasikan model HTR dalam proses pelatihan. Evaluasi dengan HTR hanya dilakukan secara pasca-pelatihan, tidak digunakan untuk memandu pelatihan.

Celah 4: Metrik Evaluasi Tidak Lengkap

Tolok ukur standar untuk restorasi dokumen (DIBCO, dll.) hanya mengevaluasi metrik visual. Tidak ada praktik standar untuk melaporkan metrik kinerja HTR seperti CER atau WER. Hal ini memperpanjang pemisahan antara penelitian restorasi dan kasus penggunaan aktual.

I.8.2 Posisi dan Kontribusi Penelitian Ini

Penelitian ini diposisikan untuk mengatasi celah yang teridentifikasi melalui tiga inovasi kunci:

Inovasi 1: Integrasi Loss Berorientasi HTR

Kontribusi pertama adalah integrasi eksplisit dari kerugian CTC dalam pelatihan Generator. Hal ini secara langsung mengatasi Celah 3 dengan menjadikan evaluasi HTR sebagai bagian integral dari proses pelatihan, bukan hanya evaluasi pasca-pelatihan.

Kebaruan: Tidak ada karya sebelumnya dalam domain restorasi dokumen yang mengintegrasikan kerugian HTR dalam lingkaran pelatihan GAN. Karya terkait terdekat adalah domain pengenalan ucapan di mana CTC digunakan untuk pelatihan, tetapi tidak dalam kerangka adversarial dan tidak untuk tugas restorasi.

Keunggulan dibandingkan dengan DocEnTr dan Text-DIAE:

- **vs DocEnTr:** DocEnTr hanya menggunakan kerugian MSE yang tidak berorientasi pada keterbacaan, pendekatan ini secara eksplisit mengoptimalkan untuk penurunan CER melalui kerugian CTC.
- **vs Text-DIAE:** Text-DIAE mengoptimalkan invarian degradasi melalui tujuan rekonstruksi, namun tidak secara eksplisit mengoptimalkan keterbacaan. Pendekatan ini menggunakan kerugian CTC untuk optimisasi keterbacaan langsung.

Kerugian MSE dan kerugian rekonstruksi tidak dapat membedakan antara derau pada tingkat piksel yang mengganggu HTR dan derau yang tidak signifikan, sementara kerugian CTC langsung mengukur dampak terhadap kinerja pengenalan teks.

Inovasi 2: Arsitektur Diskriminator Dual Modal

Kontribusi kedua adalah Diskriminator Dual Modal yang secara bersamaan mengevaluasi modalitas visual dan tekstual. Hal ini mengatasi Celah 2 dengan

menyediakan sinyal supervisi yang lebih kaya yang mempertimbangkan struktur teks.

Kebaruan: Diskriminator multi-modal telah dieksplorasi untuk tugas lain (misalnya, keterangan gambar), tetapi tidak untuk restorasi dokumen. Arsitektur yang mengkombinasikan cabang CNN visual dengan cabang pemrosesan urutan teks dalam diskriminator tunggal merupakan hal baru untuk domain ini.

Keunggulan versus Transformer-Only dan Self-Supervised:

- **vs DocEnTr:** DocEnTr tidak memiliki diskriminator sama sekali, pendekatan ini menggunakan adversarial training untuk mendorong visual realism yang lebih baik.
- **vs Text-DIAE:** Text-DIAE juga tanpa komponen adversarial dan cenderung menghasilkan output yang over-smoothed. Diskriminator Dual Modal secara eksplisit memvalidasi koherensi teks, sesuatu yang tidak mungkin dilakukan oleh MSE loss atau reconstruction loss.

Inovasi 3: Kerangka Kerja Evaluasi Komprehensif

Kontribusi ketiga adalah kerangka kerja evaluasi yang mencakup metrik visual dan metrik keterbacaan. Hal ini mengatasi Celah 4 dengan menetapkan standar baru untuk evaluasi restorasi dokumen.

Dampak Praktis: Kerangka kerja ini dapat diadopsi oleh penelitian masa depan untuk penilaian yang lebih lengkap dari metode restorasi.

Inovasi 4: Hybrid CNN-Transformer Architecture untuk Dokumen Historis

Kontribusi keempat adalah desain arsitektur hybrid yang mengkombinasikan keunggulan CNN untuk local feature extraction dengan transformer untuk global context understanding. Ini secara spesifik mengatasi keterbatasan DocEnTr untuk dokumen historis.

Justifikasi Arsitektur: Dokumen historis memiliki karakteristik unik yang membutuhkan pendekatan hybrid:

- **Preservasi Detail Lokal:** Lapisan konvolusional CNN sangat efektif untuk mendeteksi tepi halus, kontinuitas goresan, dan tekstur lokal yang krusial untuk pengenalan karakter dalam dokumen yang terdegradasi.
- **Pemahaman Konteks Global:** Perhatian-diri Transformer memungkinkan pemahaman konteks global untuk konsistensi penulisan dan penghapusan degradasi yang mempengaruhi area luas.

- **Ketangguhan terhadap Variasi:** Kombinasi ini memberikan ketangguhan terhadap berbagai jenis degradasi yang memerlukan pengenalan pola lokal dan pemahaman semantik global.

Superioritas atas Pendekatan Murni: Pendekatan ini mengatasi keterbatasan CNN murni (medan penerima terbatas), transformer murni (kurangnya bias induktif lokal, kompleksitas komputasi tinggi), dan pendekatan mandiri (ketidaksesuaian tujuan dengan keterbacaan) untuk kasus dokumen historis yang spesifik.

I.8.3 Kontribusi Teoretis: Optimasi Objektif Ganda

Dari perspektif teoretis, penelitian ini memberikan kontribusi kerangka kerja untuk menyeimbangkan objektif ganda dalam restorasi citra: kualitas perseptual (penilaian manusia) versus kualitas fungsional (pengenalan mesin).

Penerapan Umum

Meskipun berfokus pada restorasi dokumen dan HTR, prinsip optimasi objektif ganda dapat diterapkan pada domain lain:

1. Peningkatan Kualitas Citra Medis untuk CAD: Meningkatkan citra medis tidak hanya untuk inspeksi visual tetapi juga untuk Sistem Diagnosis Berbantuan Komputer.

2. Restorasi Citra Wajah untuk Pengenalan: Meningkatkan resolusi citra wajah berkualitas rendah dengan optimisasi eksplisit untuk kinerja pengenalan wajah.

3. Peningkatan Kualitas Teks Adegan untuk OCR: Serupa dengan penelitian ini tetapi untuk domain teks adegan.

Kerangka Teoretis

Formulasi umum untuk optimasi objektif ganda dalam restorasi citra:

$$\min_{G} \mathscr{L} = \alpha \mathscr{L}_{perceptual}(G) + \beta \mathscr{L}_{functional}(G, T) \tag{10}$$

Di mana G adalah model restorasi, $\mathscr{L}_{perceptual}$ mengukur kualitas perseptual (misalnya melalui pelatihan adversarial atau loss perseptual), dan $\mathscr{L}_{functional}$ mengukur performa pada tugas hilir T (misalnya HTR dalam kasus ini).

I.9 Ringkasan dan Transisi ke Metodologi

Bab ini telah menyajikan analisis komprehensif dari landscape penelitian restorasi dokumen, dari metode tradisional hingga pendekatan deep learning modern. Temuan kunci dapat diringkas sebagai:

1. Evolusi dari Metode Tradisional ke Deep Learning

Evolusi dari metode yang dirancang manual yang bergantung pada model eksplisit dan penyetelan parameter, menuju pendekatan pembelajaran mendalam berbasis data yang secara otomatis mempelajari representasi optimal. Transisi ini diperlukan karena kompleksitas degradasi dunia nyata yang tidak dapat dimodelkan secara memadai dengan aturan yang dirancang manual.

2. Superioritas GAN dan Keterbatasan Transformer-Only untuk Restorasi Realistis

Kerangka kerja pelatihan adversarial menyediakan mekanisme untuk menghasilkan restorasi visual yang realistis yang mempertahankan detail frekuensi tinggi dan tekstur. Lebih unggul dibandingkan pendekatan autoenkoder murni yang menghasilkan keluaran yang kabur. Namun, pendekatan hanya-transformer seperti DocEnTr, meskipun mencapai PSNR tinggi, memiliki keterbatasan dalam realisme visual karena kurangnya pelatihan adversarial dan cenderung menghasilkan keluaran yang terlalu halus.

3. Gap Kritis: Optimasi Keterbacaan

Metode yang ada, meskipun memiliki hasil visual yang mengesankan, tidak memiliki optimasi eksplisit untuk keterbacaan teks. Asumsi bahwa kualitas visual secara otomatis memastikan keterbacaan tidak divalidasi. Integrasi dari evaluasi HTR diperlukan untuk menjembatani gap ini.

4. Manfaat Pendekatan Multi-Modal dan Arsitektur Hybrid

Arsitektur Moda-Ganda yang secara bersamaan mempertimbangkan modalitas visual dan tekstual menyediakan supervisi yang lebih kaya dan lebih selaras dengan tujuan akhir dari restorasi dokumen: memungkinkan transkripsi yang akurat. Pendekatan hybrid CNN-Transformer menunjukkan superioritas untuk dokumen historis dengan menggabungkan preservasi fitur lokal dan pemahaman konteks global, mengatasi keterbatasan dari pendekatan murni.

Transisi: Berdasarkan analisis dan gap yang teridentifikasi, Bab III akan menjelaskan metodologi detail untuk mengembangkan framework yang diusulkan. Metodologi akan menguraikan detail implementasi dari Diskriminator Dual Modal, fungsi loss multi-komponen, dan protokol evaluasi komprehensif.

Pustaka

[1] Souibgui, M., Khlif, N., Amara, N. B., dan El Abed, H. (2019). *Enhance to Read Better: A Multi-Task Adversarial Network for Handwritten Document Image Enhancement*. Pattern Recognition Letters, 128, 115-122.

[2] Souibgui, M. A., dan Kessentini, Y. (2021). *DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement*. IEEE Transactions on Emerging Topics in Computational Intelligence, 3(1), 13-25.

[3] Souibgui, M., Biswas, S., Mafla, A., Biten, A. F., Fornés, A., Kessentini, Y., Lladós, J., Gomez, L., dan Karatzas, D. (2022). *Text-DIAE: A Self-Supervised Degradation Invariant Autoencoder for Text Recognition and Document Enhancement*. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 3026-3034.

[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... dan Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. International Conference on Machine Learning, 8748-8763.

[5] Vaswani, A., Shazeer, N., Parmar, N., dan Uszkoreit, J. (2023). *Attention Is All You Need: Cross-Modal Architectures for Document Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 1125-1139.

[6] Baltrusaitis, T., Ahuja, C., dan Morency, L. P. (2019). *Multimodal Machine Learning: A Survey and Taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[7] Jadhav, A., Singh, P., dan Kumar, P. (2022). *Correlation Analysis Between Visual Quality Metrics and HTR Performance in Document Restoration*. Document Recognition and Retrieval, 13420, 134-143.

[8] Ni, Z., Liu, Y., Chen, W., dan Tang, X. (2024). *Perceptual Loss Integration in Document Enhancement GANs*. Computer Vision and Pattern Recognition, 4567-4576.

[9] Weninger, T., Chiu, C. C., dan Hermansky, H. (2023). *Multi-Task Learning for Speech Enhancement with ASR Objectives*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2145-2158.

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... dan Liu, P. J. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, 21, 1-89.

[11] Souibgui, M., Biswas, S., Khamekhem Jemni, S., Kessentini, Y., Fornés, A., Lladós, J., dan Pal, U. (2022). *DocEnTr: An End-to-End Document Image Enhancement Transformer*. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 322-327.

[12] Diaz, M., Lladós, J., dan Fornés, A. (2021). *Transformer-Based Handwritten Text Recognition: Beyond Recurrent Networks*. International Conference on Document Analysis and Recognition, 832-847.

[13] Isola, P., Zhu, J. Y., Zhou, T., dan Efros, A. A. (2017). *Image-to-Image Translation with Conditional Adversarial Networks*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125-1134.

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... dan Bengio, Y. (2014). *Generative Adversarial Nets*. Advances in Neural Information Processing Systems, 27, 2672-2680.