IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK IMAGE – UP SCALE SECARA OFFLINE

SKRIPSI

DISUSUN OLEH

FARHAN SHIDIQ SIBARANI
NPM. 2209020162

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2026

IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK IMAGE – UP SCALE SECARA OFFLINE

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Teknologi Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

Farhan Shidiq Sibarani
NPM. 2209020162

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2026

LEMBAR PENGESAHAN

Judul Skripsi	:	IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK IMAGE – UP SCALE SECARA OFFLINE
Nama Mahasiswa	:	Farhan Shidiq Sibarani
NPM	:	2209020162
Program Studi	:	Teknologi Informasi

Menyetujui
Komisi Pembimbing

(Farid Akbar Siregar, S.Kom., M.Kom)
NIDN.0104049401

Ketua program studi
Teknologi Informasi

(Fatma Sari Hutagalung, M.Kom)
NIDN. 0117019301

Dekan

(Dr. Al – Khowarizmi, S.Kom., M.Kom)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK IMAGE – UP SCALE SECARA OFFLINE

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 6 Juni 2026
Yang membuat pernyataan



Farhan Shidiq Sibarani
NPM. 2209020162

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Farhan Shidiq Sibarani
NPM	: 2209020162
Program Studi	: Teknologi Informasi
Karya Ilmiah	: Skripsi

Skripsi Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif (Non-Exclusive Royalty free Right) atas penelitian skripsi saya yang berjudul:

IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK *IMAGE – UP SCALE* SECARA OFFLINE

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 6 Juni 2026

Yang membuat Pernyataan



Farhan Shidiq Sibarani

NPM. 2209020162

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Farhan Shidiq Sibarani
Tempat dan Tanggal Lahir	: Medan, 12 Desember 2004
Alamat Rumah	: JL. Tanjung Raya LK XII No.7
Telepon/Faks/HP	: 081363437387
E-mail	: farhanshidiq1212@gmail.com
Instansi Tempat Kerja	: -
Alamat Kantor	: -

DATA PENDIDIKAN :

SD	: SD 066046 Medan Helvetia	TAMAT : 2016
SMP	: SMPN 18 Medan	TAMAT : 2019
SMA	: SMAN 12 Medan	TAMAT : 2022

KATA PENGANTAR



Pendahuluan

Assalamu'alaikum Warahmatullahi Wabarakatuh

Segala puji dan rasa syukur dipanjatkan kehadirat Allah SWT yang telah melimpahkan berkah serta karunia-Nya, sehingga penyelesaian naskah skripsi ini bisa terwujud dengan lancar. Selawat beserta salam senantiasa dihaturkan kepada junjungan Nabi Muhammad SAW, yang telah menuntun seluruh umat ke arah kebenaran serta menjadi sumber inspirasi dalam menempuh rangkaian riset keilmuan ini.

1. Kedua orang tua penulis, Ibu Tri wahyuni & ALM. Gumri Sibarani yang senantiasa memberikan doa, dukungan moral, serta motivasi yang tiada henti dalam setiap proses yang penulis jalani.
2. Nenek Penulis, ALM. Mugirah yang semasa hidup memberikan nasehat dan dukungan menyelesaikan studi.
3. Keluarga penulis, Suci Ramadhani, Yudha Firmansyah, Alya Putri Rasyah, Bapak Irwansyah dan Ibu Ratih Susanti yang selalu memberikan semangat dan menjadi sumber kekuatan dalam menyelesaikan skripsi ini.
4. Sahabat penulis, Firzatullah, Mayza Fariza Nzulafri, Lisandra Junika Situmorang, Fadillah Akmal Bintang, Muhammad Ilham, Muhammad Raka Abdurahman, Bagus Sadewo, Faisal Rabbani Sitompul yang menjadi tempat diskusi dan dukungan semangat.

5. Bapak Prof. Dr. Akrim, M.Pd. Selaku Rektor Universitas Muhammadiyah Sumatera Utara.
6. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi.
7. Ibu Dr. Firrahmi Risky, S.Kom., M.Kom selaku Wakil Dekan I Fakultas Ilmu Komputer dan Teknologi Informasi.
8. Bapak Mhd. Basri, M.Kom selaku Wakil Dekan III Fakultas Ilmu Komputer dan Teknologi Informasi.
9. Ibu Fatma Sari Hutagalung, S.Kom., M.Kom selaku Ketua Program Studi Teknologi Informasi.
10. Bapak Okvi Nugroho, S.Kom., M.Kom, Selaku Sekretaris Prodi Program Studi Teknologi Informasi
11. Bapak Farid Akbar Siregar S.Kom., M.Kom selaku dosen pembimbing yang telah memberikan arahan, masukan, serta bimbingan selama proses penyusunan skripsi ini.
12. Bapak/Ibu dosen & pegawai, Abang Farid Akbar Siregar, Abang Okvi Nugroho, Bapak Mahardika Abdi Prawira Tanjung, Bapak Ferdy Riza, Abang Muhammad Haris, Abang Andika Suras Saputra, Abang Rahmad Hidayat, Kakak Zuli Agustina Gultom, Kakak Suvia Agustin, Kakak Meylani, Ibu Hevlie Windah Nazry, Ibu Fatma Sari Hutagalung serta seluruh civitas akademika Fakultas Ilmu Komputer dan Teknologi Informasi yang telah memberikan ilmu dan pengalaman selama masa perkuliahan.

13. Teman perkuliahan penulis, Syahputra Anugrah Rijki. S, Mhd. Alwi Azmi, Bayu Amin Santoso, Saddam Aditya Permady, Farhan Fanalty Siregar, Rio Fadly, Dwi Nisyatul Wardah, Dinda anantya, Aida fadillah, yang saling berbagi kegiatan akademik, membantu dan bercerita.
14. Kepada diri sendiri yang telah melewati masa perkuliahan dengan pengalaman, ilmu pengetahuan serta cerita bersama keluarga FIKTI.

IMPLEMENTASI MODEL REAL – ESRGAN DENGAN TEKNIK KUANTISASI PADA NPU UNTUK IMAGE – UP SCALE SECARA OFFLINE

ABSTRAK

Studi ini menyajikan implementasi model Real-ESRGAN dengan teknik kuantisasi pada beberapa unit pemrosesan, termasuk Unit Pemrosesan Pusat (CPU), Unit Pemrosesan Grafis (GPU), dan Unit Pemrosesan Neural (NPU) untuk superresolusi gambar offline. Penelitian ini dimotivasi oleh meningkatnya permintaan akan gambar digital berkualitas tinggi dan keterbatasan komputasi perangkat konvensional dalam mengeksekusi model pembelajaran mendalam yang kompleks. Untuk mengatasi masalah ini, akselerasi perangkat keras dan teknik kuantisasi menggunakan presisi FP16 dan INT8 diterapkan untuk meningkatkan efisiensi komputasi dan mengurangi beban kerja inferensi. Studi ini menggunakan desain faktorial eksperimental untuk mengevaluasi dampak jenis unit pemrosesan dan presisi model pada kualitas gambar dan kinerja komputasi. Proses implementasi meliputi konversi model dari ONNX ke OpenVINO Intermediate Representation (IR), diikuti oleh Post-Training Quantization (PTQ) dan penyebaran menggunakan OpenVINO Runtime. Evaluasi kinerja dilakukan menggunakan metrik PSNR, SSIM, MSE, RMSE, dan latensi inferensi. Hasil eksperimen menunjukkan bahwa FP16 menghasilkan kualitas rekonstruksi gambar yang lebih baik dengan nilai PSNR dan SSIM yang lebih tinggi, sedangkan INT8 mencapai latensi inferensi yang lebih rendah dan efisiensi komputasi yang lebih tinggi. Lebih lanjut, NPU menunjukkan kinerja inferensi yang lebih efisien dibandingkan dengan CPU dan GPU untuk skenario penerapan offline. Studi ini juga mengungkapkan adanya trade-off antara kualitas gambar dan efisiensi komputasi tergantung pada presisi yang dipilih dan konfigurasi unit pemrosesan. Oleh karena itu, implementasi yang diusulkan menunjukkan kelayakan penerapan model Real-ESRGAN terkuantisasi pada akselerator perangkat keras berbasis edge untuk aplikasi super-resolusi gambar offline yang efisien.

Kata kunci: Real-ESRGAN, Kuantisasi, FP16, INT8, CPU, GPU, NPU, OpenVINO, Super-Resolusi Gambar, Komputasi Edge, Inferensi Offline.

IMPLEMENTATION OF THE REAL – ESRGAN MODEL WITH QUANTIZATION TECHNIQUES ON AN NPU FOR IMAGE UP SCALE OFFLINE

ABSTRACT

This study presents the implementation of the Real-ESRGAN model with quantization techniques on multiple processing units, including Central Processing Unit (CPU), Graphics Processing Unit (GPU), and Neural Processing Unit (NPU) for offline image super-resolution. The research is motivated by the increasing demand for high-quality digital images and the computational limitations of conventional devices in executing complex deep learning models. To address this issue, hardware acceleration and quantization techniques using FP16 and INT8 precision were applied to improve computational efficiency and reduce inference workload. The study employed an experimental factorial design to evaluate the impact of processing unit types and model precision on image quality and computational performance. The implementation process included model conversion from ONNX to OpenVINO Intermediate Representation (IR), followed by Post-Training Quantization (PTQ) and deployment using OpenVINO Runtime. Performance evaluation was conducted using PSNR, SSIM, MSE, RMSE, and inference latency metrics. The experimental results indicate that FP16 produced better image reconstruction quality with higher PSNR and SSIM values, while INT8 achieved lower inference latency and higher computational efficiency. Furthermore, NPU demonstrated more efficient inference performance compared to CPU and GPU for offline deployment scenarios. The study also reveals a tradeoff between image quality and computational efficiency depending on the selected precision and processing unit configuration. Therefore, the proposed implementation demonstrates the feasibility of deploying quantized Real-ESRGAN models on edge-based hardware accelerators for efficient offline image superresolution applications.

Keywords: Real-ESRGAN, Quantization, FP16, INT8, CPU, GPU, NPU, OpenVINO, Image Super-Resolution, Edge Computing, Offline Inference.

DAFTAR ISI

Abstrak.....i
Abstrack.....ii
Daftar Isi.....iii
Daftar Tabel.....iv
Daftar gambar.....v
BAB I 1
1.1 Latar Belakang Masalah.....1
1.2 Rumusan Masalah.....4
1.3 Batasan Masalah 5
1.4 Tujuan Penelitian 6
1.5 Manfaat Penelitian 7
BAB II.....9
2.1 Citra digital 9
2.2 Image Super – Resolution 10
2.3 Deep Learning Methods.....10
2.4 Generative Adversarial Network.....11
2.5 Real-ESRGAN.....13
2.6 Teknik kuantisasi 16
2.7 Neural Processing Unit (NPU) 17
2.8 OpenVINO.....19
2.9 Metrik Evaluasi.....20
2.10 Penelitian Relevan.....21
BAB III 23
3.1 Jenis Penelitian.....23
3.2 Waktu Penelitian.....25
3.3 Variabel dan Parameter Penelitian 25
3.4 Analisis Pengumpulan Data 26
3.5 Gambaran umum sistem.....27
3.6 Analisis Teknik Evaluasi.....28
3.7 Kerangka Berfikir.....30
3.8 Analisis Pemodelan Sistem 31
3.8.1 Flowchart sistem 32
3.8.2 Flowchart REAL-ERSGAN.....34
3.9 Analisis Struktur Sistem.....37
3.9.1 Use Case Diagram.....38
3.9.2 Activity Diagram 40

3.9.3 Sequence Diagram 42
3.9.4 Component Diagram 44
3.10 Interface 46
BAB IV 48
4.1 Implementasi Antarmuka 48
4.1.1 Antarmuka Sistem Image Up - Scale 48
4.1.2 Antarmuka Riwayat dan Evaluasi Proses Citra 50
4.2 Evaluasi Hasil Citra Model Real – ESRGAN 57
4.3 Evaluasi Hasil Nilai Metrik Inferensi Model Real – ESRGAN 58
4.3.1 Evaluasi PSNR 61
4.3.2 Evaluasi SSIM 62
4.3.3 Evaluasi MSE 63
4.3.4 Evaluasi RMSE 63
4.3.5 Hasil Evaluasi PSNR, SSIM, MSE, RMSE 64
BAB V 66
5.1 Kesimpulan 66
5.2 Saran 67
Daftar pustaka 69

DAFTAR TABEL

Tabel 2.1 Penelitian Relevan.....21
Tabel 3.1 Waktu Penelitian.....25
Tabel 4.1 Metrik Evaluasi Hasil CItra.....59
Tabel 4.3 Gambar dan Perbandingan Nilai Metriks 61

DAFTAR GAMBAR

Gambar 2.1 Konsep Image Super – Resolution.....10
Gambar 2.2 Struktur Generative Adversarial Network 11
Gambar 2.3 Struktur Neural Processing Unit.....18
Gambar 2.4 *OpenVino Optimized* 19
Gambar 3.1 Struktur Penelitian Desain Experimental.....25
Gamabr 3.2 Dataset Gambar.....27
Gambar 3.3 kerangka Berfikir.....30
Gambar 3.4 Flowchart Sistem.....33
Gambar 3.5 Flowchart Real – Esrgan.....36
Gambar 3.6 Use – Case Diagram.....38
Gambar 3.7 Activity Diagram.....41
Gambar 3.8 Sequence Diagram.....43
Gambar 3.9 Component Diagram 45
Gambar 3.10 *Interface* 46
Gambar 4.1 Antarmuka Sistem *Image Up - Scale*.....49
Gambar 4.2 Metrik Evaluasi Hasil Proses Citra.....50
Gambar 4.3 Evaluasi Hasil Citra Real – Esrgan.....57
Gambar 4.4 *Bar chart* Metriks Hasil.....64

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Gambar visual merupakan media komunikasi non-verbal yang memiliki peranan penting dalam menyampaikan informasi secara efektif dan efisien pada era digital. Seiring dengan meningkatnya penggunaan media visual dalam berbagai bidang, kebutuhan terhadap citra berkualitas tinggi juga semakin meningkat. Kualitas citra yang baik menjadi faktor penting karena dapat memengaruhi kemampuan pengguna dalam memahami informasi yang terkandung di dalam gambar. Namun, berbagai faktor seperti keterbatasan perangkat akuisisi, proses kompresi, degradasi citra, serta resolusi yang rendah sering menyebabkan kualitas gambar menurun sehingga informasi visual yang dihasilkan menjadi kurang optimal. Oleh karena itu, berbagai metode peningkatan kualitas citra terus dikembangkan untuk menghasilkan gambar yang lebih jelas, tajam, dan informatif.

Salah satu pendekatan yang berkembang pesat dalam bidang pengolahan citra adalah image super-resolution, yaitu teknik yang bertujuan untuk merekonstruksi citra beresolusi tinggi dari citra beresolusi rendah. Perkembangan deep learning telah memberikan kemajuan yang signifikan dalam bidang ini karena mampu mempelajari hubungan kompleks antara citra resolusi rendah dan citra resolusi tinggi secara lebih efektif dibandingkan metode konvensional. Berbagai model berbasis deep learning telah dikembangkan untuk meningkatkan kualitas hasil rekonstruksi citra, khususnya melalui pendekatan Generative Adversarial Network (GAN) yang terbukti mampu menghasilkan detail visual

yang lebih realistis dan berkualitas tinggi dibandingkan pendekatan tradisional (Gui et al., 2023; Z. Wang et al., 2022).

Salah satu model super-resolution berbasis GAN yang banyak digunakan adalah Real-ESRGAN. Model ini dikembangkan untuk menangani berbagai bentuk degradasi citra dunia nyata melalui pendekatan high-order degradation sehingga mampu menghasilkan citra dengan detail tekstur yang lebih tajam dan realistis. Keunggulan tersebut menjadikan Real-ESRGAN banyak dimanfaatkan dalam berbagai bidang, seperti peningkatan kualitas citra medis, pengolahan citra digital, restorasi gambar lama, hingga aplikasi penginderaan jauh yang memerlukan kualitas visual tinggi (G. Chen et al., 2025; Nandal et al., 2024; X. Wang et al., 2021).

Meskipun memiliki kemampuan rekonstruksi citra yang baik, model RealESRGAN memiliki kompleksitas komputasi yang relatif tinggi. Proses inferensi membutuhkan sumber daya komputasi yang besar sehingga implementasinya sering dilakukan menggunakan perangkat keras berperforma tinggi atau layanan berbasis cloud. Pendekatan tersebut dapat menimbulkan beberapa keterbatasan seperti meningkatnya latency, ketergantungan terhadap koneksi internet, serta potensi risiko keamanan dan privasi data. Oleh karena itu, implementasi model secara offline pada perangkat lokal menjadi alternatif yang menarik untuk dikembangkan, terutama pada lingkungan edge computing yang membutuhkan efisiensi komputasi dan kemandirian sistem (Tan & Cao, 2021).

Salah satu teknologi yang mendukung implementasi kecerdasan buatan pada perangkat edge adalah Neural Processing Unit (NPU). NPU merupakan akselerator khusus yang dirancang untuk mempercepat proses inferensi model

deep learning melalui optimasi komputasi matriks dan operasi neural network. Penggunaan NPU memungkinkan model kecerdasan buatan dijalankan secara lebih efisien dengan konsumsi daya yang lebih rendah dibandingkan pemrosesan menggunakan perangkat komputasi konvensional. Kemampuan tersebut menjadikan NPU sebagai solusi yang potensial untuk menjalankan model image super-resolution secara offline pada perangkat edge (Y. Chen & Lu, 2026; Tan & Cao, 2021).

Selain memanfaatkan akselerasi perangkat keras, optimasi model juga dapat dilakukan melalui teknik kuantisasi. Kuantisasi merupakan metode yang digunakan untuk mengurangi kompleksitas komputasi dan kebutuhan memori model dengan mengubah representasi numerik ke precision yang lebih rendah. Precision FP16 dan INT8 menjadi dua pendekatan yang banyak digunakan karena mampu meningkatkan efisiensi inferensi tanpa memerlukan perubahan signifikan pada arsitektur model. FP16 umumnya mampu mempertahankan kualitas hasil inferensi dengan baik, sedangkan INT8 menawarkan efisiensi komputasi yang lebih tinggi melalui pengurangan beban pemrosesan dan percepatan waktu inferensi. Namun demikian, penggunaan precision yang berbeda berpotensi menghasilkan perbedaan karakteristik performa, baik dari sisi kualitas citra maupun efisiensi komputasi (J. Liu et al., 2021; Totad et al., 2025; Zheng et al., 2026).

Evaluasi kualitas citra hasil super-resolution umumnya dilakukan menggunakan metrik numerik seperti Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE). Metrik tersebut digunakan untuk mengukur

tingkat kesesuaian antara citra hasil rekonstruksi dengan citra referensi. Akan tetapi, beberapa penelitian menunjukkan bahwa hasil evaluasi numerik tidak selalu sejalan dengan persepsi visual manusia terhadap kualitas citra yang dihasilkan. Citra yang memiliki nilai metrik lebih tinggi belum tentu terlihat lebih baik secara visual, begitu pula sebaliknya. Kondisi ini menunjukkan bahwa hubungan antara kualitas numerik dan kualitas visual masih menjadi aspek yang menarik untuk dianalisis lebih lanjut dalam implementasi model super-resolution (Bakurov et al., 2022; Mudeng et al., 2022; Zhu et al., 2025).

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada implementasi model Real-ESRGAN menggunakan teknik kuantisasi FP16 dan INT8 pada Neural Processing Unit (NPU) untuk image up-scale secara offline. Penelitian dilakukan untuk menganalisis pengaruh penggunaan precision yang berbeda terhadap kualitas citra dan efisiensi komputasi berdasarkan metrik PSNR, SSIM, MSE, RMSE, serta latency inferensi. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi hubungan antara hasil evaluasi numerik dengan kualitas visual citra yang dihasilkan. Hasil penelitian diharapkan dapat memberikan pemahaman yang lebih komprehensif mengenai trade-off antara kualitas citra dan efisiensi komputasi pada implementasi model super-resolution berbasis deep learning di lingkungan edge computing.

1.2 Rumusan Masalah

Bagaimana kinerja model Real-ESRGAN terkuantisasi yang diimplementasikan pada Neural Processing Unit (NPU) dalam meningkatkan kualitas gambar secara offline ditinjau dari kualitas visual dan efisiensi komputasi.

1.3 Batasan Masalah

Untuk menjaga fokus penelitian serta memperoleh hasil analisis yang terarah, maka batasan masalah dalam penelitian ini ditetapkan sebagai berikut:

1. Penelitian ini menggunakan model Real-ESRGAN yang telah dilatih sebelumnya (*pre-trained model*) tanpa melakukan proses pelatihan ulang (*retraining*) maupun *fine-tuning*.
2. Implementasi model dilakukan menggunakan framework OpenVINO untuk proses konversi, optimasi, dan deployment model pada perangkat keras yang mendukung ekosistem Intel.
3. Teknik kuantisasi yang digunakan dalam penelitian ini dibatasi pada precision FP16 dan INT8.
4. Penelitian difokuskan pada implementasi image super-resolution secara offline tanpa melibatkan layanan berbasis cloud maupun konektivitas internet selama proses inferensi.
5. Dataset yang digunakan berupa citra digital yang telah dipersiapkan sebagai data uji untuk mengevaluasi hasil image up-scale menggunakan model Real-ESRGAN.
6. Evaluasi kinerja model difokuskan pada aspek kualitas citra dan efisiensi komputasi menggunakan metrik Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan inference latency.
7. Penelitian ini tidak membahas konsumsi daya (*power consumption*), penggunaan memori (*memory usage*), maupun utilisasi perangkat keras selama proses inferensi.

8. Penelitian ini tidak melibatkan evaluasi formal berbasis persepsi manusia (*user study*), sehingga analisis kualitas citra difokuskan pada hasil visual dan metrik evaluasi numerik.
9. Penelitian ini tidak membandingkan model Real-ESRGAN dengan metode super-resolution lainnya seperti ESRGAN, A-ESRGAN, SRCNN, EDSR, maupun model super-resolution berbasis deep learning lainnya.
10. Penelitian ini tidak membahas pengembangan arsitektur model maupun teknik optimasi lanjutan selain kuantisasi FP16 dan INT8.

1.4 Tujuan Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis dalam pengembangan teknologi peningkatan kualitas citra berbasis *deep learning*, khususnya pada implementasi model Real-ESRGAN terkuantisasi di Neural Processing Unit (NPU). Manfaat tersebut mencakup pemahaman yang lebih mendalam mengenai hubungan antara kualitas citra dan efisiensi komputasi, serta penerapan hasil penelitian dalam pengembangan sistem yang lebih optimal, efisien, dan sesuai dengan kebutuhan aplikasi berbasis *edge computing*. Adapun manfaat penelitian ini dapat diuraikan sebagai berikut:

1. Mengimplementasikan model Real-ESRGAN yang telah dioptimasi menggunakan teknik kuantisasi FP16 dan INT8 pada Neural Processing Unit (NPU) untuk proses image up-scale secara offline.
2. Menganalisis pengaruh penggunaan precision FP16 dan INT8 terhadap kualitas citra hasil image up-scale berdasarkan metrik evaluasi PSNR, SSIM, MSE, dan RMSE.

3. Menganalisis pengaruh penggunaan precision FP16 dan INT8 terhadap efisiensi komputasi berdasarkan waktu inferensi (latency) pada implementasi model Real-ESRGAN di NPU.
4. Mengidentifikasi trade-off antara kualitas citra dan efisiensi komputasi yang dihasilkan oleh penggunaan precision FP16 dan INT8 pada model Real-ESRGAN.
5. Menganalisis hubungan antara hasil evaluasi numerik dan kualitas visual citra yang dihasilkan pada implementasi model Real-ESRGAN secara offline.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis dalam pengembangan teknologi peningkatan kualitas citra berbasis deep learning. Manfaat penelitian ini mencakup pemahaman yang lebih mendalam mengenai hubungan antara kualitas citra dan efisiensi komputasi pada lingkungan edge computing serta menjadi referensi dalam pengembangan sistem image superresolution yang lebih optimal.

Secara teoritis, memberikan kontribusi dalam memperkaya kajian mengenai implementasi model super-resolution berbasis deep learning pada perangkat keras akselerator kecerdasan buatan. Penelitian ini juga memberikan pemahaman mengenai pengaruh penggunaan precision FP16 dan INT8 terhadap kualitas citra dan efisiensi komputasi pada model Real-ESRGAN.

Secara praktis, dapat menjadi acuan dalam menentukan konfigurasi precision yang sesuai dengan kebutuhan aplikasi, baik yang mengutamakan kualitas citra maupun efisiensi komputasi. Hasil penelitian ini juga dapat mendukung pemanfaatan Neural Processing Unit (NPU) dalam menjalankan model image super-resolution secara offline dengan lebih efisien. Selain itu,

penelitian ini diharapkan dapat memberikan gambaran mengenai trade-off antara kualitas citra dan performa inferensi sehingga dapat membantu pengembang dalam memilih strategi optimasi model yang sesuai untuk implementasi pada lingkungan edge computing

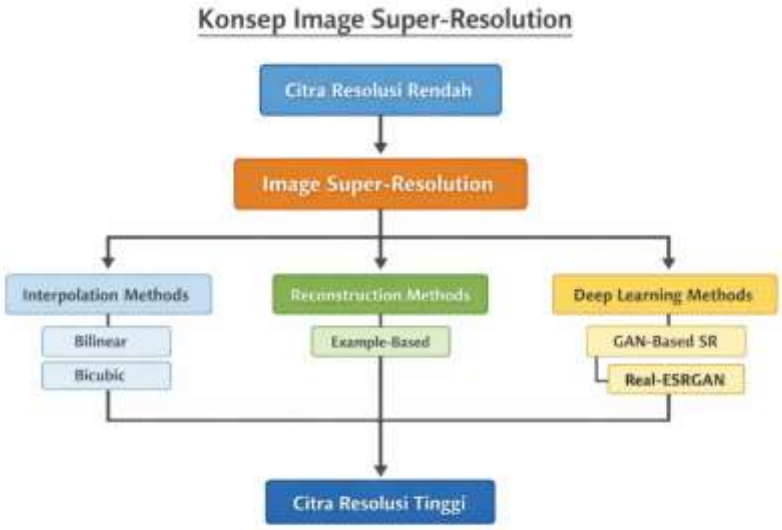
BAB II
LANDASAN TEORI

2.1 Citra digital

Citra digital merupakan representasi visual suatu objek yang disimpan dalam bentuk data numerik sehingga dapat diproses oleh sistem komputer. Citra digital tersusun atas sekumpulan piksel (picture elements) yang masing-masing memiliki nilai intensitas tertentu untuk merepresentasikan tingkat kecerahan maupun warna pada suatu gambar. Kualitas citra digital sangat dipengaruhi oleh jumlah piksel, resolusi, tingkat noise, kontras, serta ketajaman detail yang terkandung di dalamnya. Semakin tinggi resolusi suatu citra, semakin banyak informasi visual yang dapat direpresentasikan sehingga kualitas gambar menjadi lebih baik (Z. Zhu et al., 2023). Dalam bidang pengolahan citra modern, citra digital menjadi objek utama yang digunakan dalam berbagai aplikasi seperti klasifikasi citra, deteksi objek, restorasi citra, hingga image super-resolution.. Salah satu tantangan utama dalam pengolahan citra digital adalah mengatasi degradasi citra yang disebabkan oleh blur, noise, kompresi, maupun keterbatasan resolusi. Oleh karena itu, berbagai metode restorasi dan peningkatan kualitas citra terus dikembangkan untuk menghasilkan representasi visual yang lebih jelas dan informatif. Citra digital menjadi komponen fundamental dalam penelitian ini karena digunakan sebagai objek utama pada proses image up-scale menggunakan model Real-ESRGAN (Nandal et al., 2024).

2.2 Image Super – Resolution

Image Super-Resolution (SR) merupakan teknik dalam pengolahan citra yang bertujuan untuk meningkatkan resolusi gambar dengan merekonstruksi citra resolusi tinggi dari citra resolusi rendah. Teknik ini digunakan untuk memperbaiki kualitas visual gambar yang mengalami penurunan resolusi akibat proses akuisisi, kompresi, atau keterbatasan perangkat kamera (J. Liu et al., 2021). Dengan memanfaatkan metode pembelajaran mesin dan deep learning, teknik superresolution mampu merekonstruksi detail tekstur dan struktur gambar sehingga menghasilkan kualitas citra yang lebih baik dibandingkan metode interpolasi tradisional.



Gambar 2.1 Konsep Image Super – Resolution

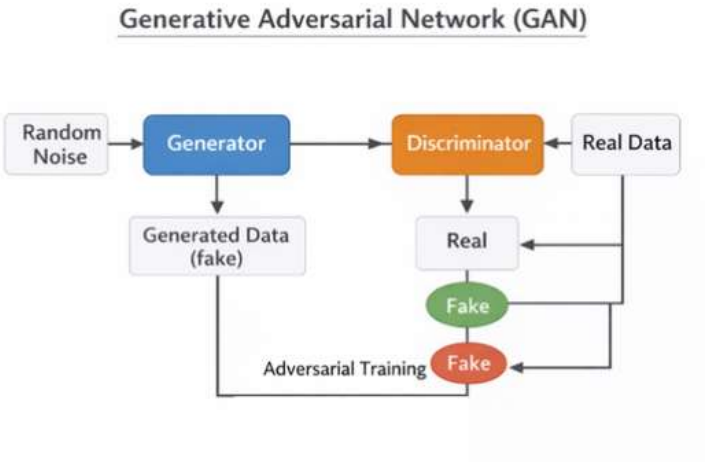
2.3 Deep Learning Methods

Deep learning menjadi pendekatan metode berbasis jaringan saraf yang paling banyak digunakan dalam image super-resolution. Model seperti Convolutional Neural Network (CNN) dan Generative Adversarial Network

(GAN) mampu mempelajari hubungan kompleks antara citra resolusi rendah dan citra resolusi tinggi melalui proses pelatihan pada dataset gambar. Pendekatan ini terbukti mampu menghasilkan rekonstruksi citra dengan detail tekstur yang lebih realistis dan kualitas visual yang lebih baik dibandingkan metode sebelumnya (Agarwal et al., 2022).

2.4 Generative Adversarial Network

Generative Adversarial Network (GAN) merupakan salah satu model pembelajaran mendalam yang digunakan untuk menghasilkan data baru yang memiliki karakteristik serupa dengan data asli melalui proses pelatihan yang bersifat kompetitif (K. Liu & Cai, 2025). Arsitektur ini terdiri dari dua jaringan saraf utama yaitu generator (G) dan discriminator (D) yang dilatih secara bersamaan dalam sebuah mekanisme adversarial. Generator bertugas menghasilkan data sintetis dari input berupa vektor acak (noise), sedangkan discriminator bertugas membedakan antara data asli dari dataset dan data sintetis yang dihasilkan oleh generator. Melalui proses kompetisi ini, generator secara bertahap meningkatkan kemampuannya dalam menghasilkan data yang semakin realistis sehingga sulit dibedakan oleh discriminator (Ma et al., 2022).



Gambar 2.2 Struktur Generative Adversial Network

GAN banyak digunakan dalam berbagai bidang computer vision seperti image synthesis, image restoration, image translation, dan image super-resolution karena kemampuannya dalam mempelajari distribusi data yang kompleks dan menghasilkan citra dengan detail visual yang realistis. Dalam beberapa tahun terakhir, berbagai pengembangan arsitektur GAN telah dilakukan untuk meningkatkan stabilitas pelatihan serta kualitas data yang dihasilkan, sehingga model ini menjadi salah satu pendekatan generatif yang sangat penting dalam perkembangan deep learning modern(Ivanenko et al., 2024; Ma et al., 2022).

Secara matematis, proses pelatihan GAN dirumuskan sebagai permainan minimax antara generator dan discriminator yang dinyatakan sebagai berikut:

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

Dengan :

G: Generator network

D: Discriminator network

x: Data asli dari distribusi data p_{data}

z: Vektor noise acak dari distribusi p_z

$G(z)$: Data sintetis yang dihasilkan generator

$\mathbb{E}$: Operator ekspektasi

Pada persamaan tersebut, discriminator berusaha memaksimalkan probabilitas dalam mengklasifikasikan data asli dan data sintetis secara benar, sedangkan generator berusaha meminimalkan kemampuan discriminator dalam membedakan data sintetis yang dihasilkannya dengan data asli. Proses optimasi ini menghasilkan keseimbangan antara kedua

jaringan sehingga generator mampu menghasilkan data dengan tingkat kemiripan yang tinggi terhadap data asli (Usama et al., 2025).

2.5 Real-ESRGAN

Real-ESRGAN merupakan model image super-resolution berbasis Generative Adversarial Network (GAN) yang dirancang untuk meningkatkan kualitas citra resolusi rendah menjadi citra resolusi tinggi dengan mempertahankan detail tekstur yang realistis. Model ini dikembangkan sebagai penyempurnaan dari ESRGAN dengan memperkenalkan pendekatan high-order degradation model yang mampu mensimulasikan berbagai jenis kerusakan citra yang sering terjadi pada gambar dunia nyata seperti blur, noise, dan artefak kompresi (X. Wang et al., 2021).

Pendekatan ini memungkinkan model mempelajari proses degradasi citra secara lebih kompleks dibandingkan metode super-resolution konvensional yang umumnya hanya menggunakan degradasi sederhana seperti bicubic downsampling. Dengan memanfaatkan arsitektur GAN, Real-ESRGAN tidak hanya memperbesar resolusi gambar, tetapi juga merekonstruksi kembali detail tekstur yang hilang sehingga menghasilkan citra yang lebih tajam dan realistis (Senthil & Tomlinson, 2025; X. Wang et al., 2021).

Model ini digunakan untuk mensimulasikan berbagai proses degradasi citra dunia nyata secara berlapis.

Struktur Model Real-ESRGAN

Dalam penelitian Real-ESRGAN terdapat empat komponen utama yang membentuk proses restorasi citra:

1. High-Order Degradation Model

Model ini digunakan untuk mensimulasikan berbagai proses degradasi citra dunia nyata secara berlapis.

$$x = D_n(D_{n-1}(\dots D_1(y))) \quad (2)$$

Dengan:

- y: citra asli (ground truth)
- x: citra terdegradasi (low resolution)
- D_i: operator degradasi ke-i
- n: jumlah tahapan degradasi

Operator degradasi dapat berupa:

- blur convolution
- resize
- noise injection
- JPEG compression

Pendekatan ini membuat model mampu menangani berbagai kerusakan citra yang kompleks.

2. Generator Architecture (RRDB)

Generator pada Real-ESRGAN menggunakan *Residual-in-Residual Dense Block* (RRDB) yang dirancang untuk meningkatkan kemampuan jaringan dalam mempelajari fitur citra.

$$F(x) = x + \beta R(x) \quad (3)$$

Dengan:

- x: input feature
- $R(x)$: fungsi residual

- β: residual scaling

Penggunaan *residual scaling* bertujuan menjaga stabilitas pelatihan pada jaringan yang sangat dalam dan menghindari masalah exploding gradient.

3. Loss Function

Real-ESRGAN menggunakan kombinasi beberapa fungsi kerugian untuk menghasilkan kualitas citra yang optimal.

a. Pixel Loss (L1)

$$L_{pixel} = \mathbb{E} \parallel G(x) - y \parallel_1 \quad (4)$$

Berfungsi menjaga kesesuaian nilai piksel antara citra hasil restorasi dan citra asli.

b. Perceptual Loss

$$L_{percep} = \mathbb{E} \parallel \phi\big(G(x)\big)\phi(y) \parallel \quad (5)$$

Dengan:

- ϕ: *feature extractor* (VGG network)

Loss ini membantu model mempertahankan struktur tekstur dan detail visual.

c. Adversarial Loss

$$L_{adv} = \mathbb{E}[\log\ D(y)] + \mathbb{E}[\log\,(1 - D(G(x)))] \quad (6)$$

Loss ini mendorong generator menghasilkan citra yang lebih realistis secara visual melalui mekanisme kompetisi dengan discriminator.

4. Total Loss Optimization

$$L_{total} = \lambda L_{pixel} + \eta L_{percep} + \gamma L_{adv} \quad (7)$$

Dengan:

- λ, η, γ: bobot koefisien loss

Kombinasi ini bertujuan menyeimbangkan antara:

- akurasi piksel
- kualitas persepsi visual
- realisme tekstur

2.6 Teknik kuantisasi

Kuantisasi adalah teknik kompresi model yang mengonversi parameter data (seperti bobot dan bias) dari format presisi tinggi floating-point 32-bit (FP32) ke dalam format integer 8-bit (INT8) yang jauh lebih hemat memori. Kuantisasi INT8 didefinisikan sebagai proses pemetaan (mapping) nilai-nilai kontinu dalam rentang dinamis besar (FP32) ke dalam rentang diskrit yang sangat terbatas, yaitu hanya 256 nilai (dari -128 hingga 127) (Fraga, 2021). Dengan mengubah format data menjadi 8-bit, beban kerja perangkat keras berkurang secara signifikan karena operasi matematika pada bilangan bulat (integer arithmetic) jauh lebih cepat dan hemat energi dibandingkan operasi bilangan desimal (Fraga, 2021; J. Wang & Zhao, 2025).

Menurut (Kim et al., 2021) Kuantisasi INT8 merupakan proses reduksi presisi numerik yang mentransformasikan representasi data dari domain kontinu ke domain diskrit. Teknik ini menjadi kunci dalam paradigma TinyML dan Edge AI untuk memungkinkan eksekusi model saraf tiruan yang kompleks pada perangkat dengan sumber daya komputasi yang terbatas.

Dalam implementasinya, setiap nilai FP32 (F) dikonversi menjadi nilai INT8 (Q) menggunakan faktor skala (S) dan titik nol (zero-point, Z). Logikanya adalah mencari rasio yang tepat agar informasi penting dalam gambar atau data tidak hilang saat dikecilkan. Rumus :

$$Q = round(\frac{F}{S} + Z) \quad (8)$$

Dengan :

Q = *Quantized Value*

F = *Floating – point Value*

S = *Scale Factor*

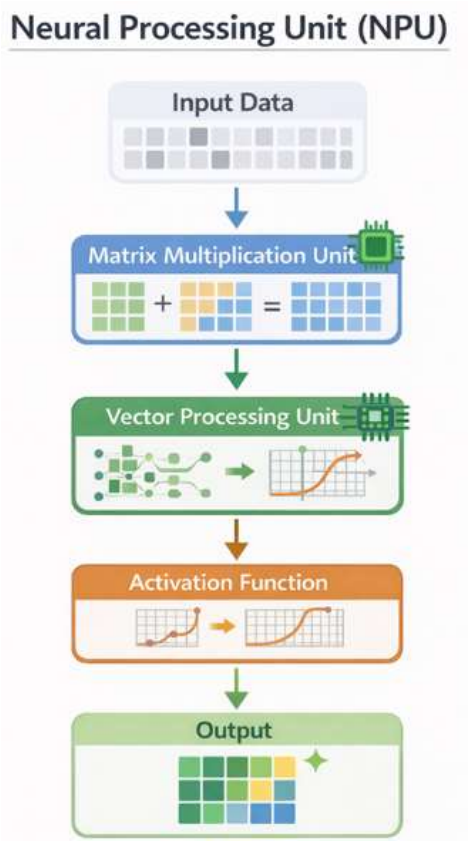
Z = *Zero – Point*

round = *Rounding Function*

Menurut (J. Wang & Zhao, 2025) Secara fundamental, kuantisasi INT8 adalah strategi optimalisasi yang memanfaatkan representasi bit-rendah untuk mengurangi hambatan transfer data pada memori (*memory bottleneck*), sehingga mempercepat latensi inferensi tanpa memerlukan perangkat keras tambahan.

2.7 Neural Processing Unit (NPU)

Neural Processing Unit (NPU) merupakan prosesor khusus yang dirancang untuk mempercepat komputasi pada model kecerdasan buatan, terutama operasi matriks dan vektor dalam *deep learning*, sehingga mampu meningkatkan efisiensi proses inferensi dibandingkan CPU dan GPU (Chung et al., 2022). NPU dioptimalkan untuk menjalankan operasi seperti konvolusi dan perkalian matriks dengan latensi rendah dan konsumsi energi yang efisien, sehingga banyak digunakan pada perangkat *edge* seperti smartphone, IoT, dan sistem *embedded* (Chung et al., 2022; Jeon et al., 2021).



Gambar 2.3 Struktur Neural Processing Unit

Dalam implementasinya, NPU sering dikombinasikan dengan teknik *lowprecision computing* seperti FP16 dan INT8 untuk mengurangi kebutuhan memori dan meningkatkan kecepatan komputasi (Chaves-González & León-Vega, 2022). Penggunaan FP16 mampu mempertahankan kualitas dengan efisiensi yang lebih baik dibanding FP32, sedangkan INT8 memberikan performa komputasi yang lebih tinggi dengan potensi penurunan kualitas.. Dalam penelitian ini, NPU digunakan untuk mengimplementasikan dan mengevaluasi kinerja model Real-ESRGAN dengan variasi precision FP16 dan INT8 dalam peningkatan kualitas citra secara offline.

2.8 OpenVINO

OpenVINO (*Open Visual Inference and Neural Network Optimization*) merupakan sebuah *toolkit* sumber terbuka komprehensif yang dikembangkan oleh Intel untuk memfasilitasi optimalisasi dan penyebaran (*deployment*) model pembelajaran mendalam pada berbagai arsitektur perangkat keras heterogen. Secara arsitektural, OpenVINO berfungsi sebagai mesin inferensi yang menjembatani kerangka kerja pelatihan seperti PyTorch atau TensorFlow dengan unit pemrosesan fisik, termasuk CPU, iGPU, dan NPU, melalui konversi model ke dalam format Representasi Antara (*Intermediate Representation* atau IR). Menurut (Bernabe et al., 2021), OpenVINO didefinisikan sebagai ekosistem optimasi yang memanfaatkan teknik fusi lapisan (*layer fusion*) dan pemetaan instruksi perangkat keras secara spesifik untuk meminimalkan latensi inferensi serta memaksimalkan penggunaan *bandwidth* memori pada perangkat *edge*. Keunggulan utamanya terletak pada kemampuannya untuk mengekstraksi performa maksimal dari instruksi tingkat rendah (seperti AVX-512 atau VNNI) tanpa memerlukan modifikasi kode manual yang kompleks.



Gambar 2.4 *OpenVino Optimized*

OpenVINO berperan sebagai katalisator dalam transformasi model dari fase eksperimentasi menuju fase produksi yang efisien. Melalui komponen *Model Optimizer*, toolkit ini melakukan serangkaian transformasi grafis statis yang mencakup pemangkasan (*pruning*) dan kuantisasi (INT8), yang secara signifikan mereduksi beban komputasi selama fase inferensi. Menurut (Praveen & Sivasankaran, 2026a) “efektivitas OpenVINO terletak pada fleksibilitasnya dalam mendukung komputasi heterogen, di mana beban kerja AI dapat didistribusikan secara dinamis lintas unit pemrosesan guna mencapai efisiensi daya dan *throughput* yang optimal." Dengan dukungan terhadap *Open Model Zoo* yang menyediakan berbagai arsitektur pra-latih, OpenVINO telah menjadi standar infrastruktur dalam pengembangan aplikasi berbasis pengenalan pola, deteksi objek, dan pemrosesan citra seperti Real-ESRGAN pada lingkungan perangkat keras Intel.

2.9 Metrik Evaluasi

Evaluasi kinerja model Real-ESRGAN pada penelitian ini dilakukan untuk mengukur kualitas citra hasil image up-scale serta efisiensi komputasi yang dihasilkan setelah penerapan teknik kuantisasi. Pengukuran kualitas citra diperlukan untuk mengetahui sejauh mana citra hasil rekonstruksi mampu mempertahankan informasi visual dan kemiripan terhadap citra referensi. Sementara itu, pengukuran efisiensi komputasi dilakukan untuk menganalisis performa inferensi model pada perangkat Neural Processing Unit (NPU). Dalam penelitian ini, evaluasi kualitas citra dilakukan menggunakan metrik Peak Signalto-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE), sedangkan

efisiensi komputasi diukur menggunakan metrik latency. Kombinasi metrik tersebut digunakan untuk memberikan analisis yang komprehensif mengenai trade-off antara kualitas citra dan efisiensi komputasi pada implementasi model RealESRGAN dengan precision FP16 dan INT8.

2.10 Penelitian Relevan

Beberapa jurnal terkait dengan penelitian yang penulis gunakan sebagai sumber kajian mendalam tentang pembahasan penelitian penulis dapat dilihat pada table berikut :

Tabel 2.1 Penelitian relevan

No	Judul	Penulis dan Tahun	Hasil
1	Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data	(X. Wang et al., 2021)	Mengembangkan RealESRGAN yang mampu menghasilkan citra super-resolution dengan kualitas visual yang lebih realistis pada berbagai degradasi citra dunia nyata.
2	IRE: Improved Image Super-Resolution Based on Real-ESRGAN	(Z. Zhu et al., 2023)	Mengusulkan peningkatan arsitektur Real-ESRGAN sehingga menghasilkan kualitas citra yang lebih baik dibandingkan model dasar Real-ESRGAN.
3	Evaluation of REAL-ESRGAN Using Different Types of Image Degradation	(Andriadi et al., 2024)	Menunjukkan bahwa performa Real-ESRGAN dipengaruhi oleh jenis degradasi citra yang digunakan pada proses evaluasi.

4	Super-Resolution of Medical Images Using Real ESRGAN	(Nandal et al., 2024)	Membuktikan bahwa Real-ESRGAN mampu meningkatkan kualitas citra medis dengan mempertahankan detail visual yang penting.
5	Super-resolution reconstruction of offaxis digital holography based on Real-ESRGAN	(G. Chen et al., 2025)	Real-ESRGAN berhasil meningkatkan kualitas rekonstruksi citra holografi dan memperbaiki detail objek secara signifikan.

Berdasarkan penelitian terdahulu, sebagian besar penelitian berfokus pada pengembangan dan peningkatan kualitas citra menggunakan model RealESRGAN, seperti yang dilakukan oleh (Andriadi et al., 2024; G. Chen et al., 2025; Nandal et al., 2024; X. Wang et al., 2021; Z. Zhu et al., 2023). Penelitian-penelitian tersebut menunjukkan kemampuan Real-ESRGAN dalam menghasilkan citra super-resolution dengan kualitas visual yang baik pada berbagai jenis data dan tingkat degradasi citra. Namun, penelitian tersebut belum membahas optimasi model melalui teknik kuantisasi maupun implementasi pada perangkat Neural Processing Unit (NPU). Di sisi lain, (J. Liu et al., 2021; Totad et al., 2025) meneliti penerapan teknik kuantisasi untuk meningkatkan efisiensi komputasi model deep learning. Meskipun berhasil menunjukkan peningkatan performa inferensi dan pengurangan kebutuhan komputasi, penelitian tersebut tidak secara khusus mengimplementasikan Real-ESRGAN pada perangkat NPU. Sementara itu, (Tan & Cao, 2021)membahas pemanfaatan NPU untuk mempercepat inferensi model deep learning pada perangkat edge, tetapi tidak mengkaji model Real-ESRGAN maupun pengaruh variasi precision terhadap kualitas citra yang dihasilkan.

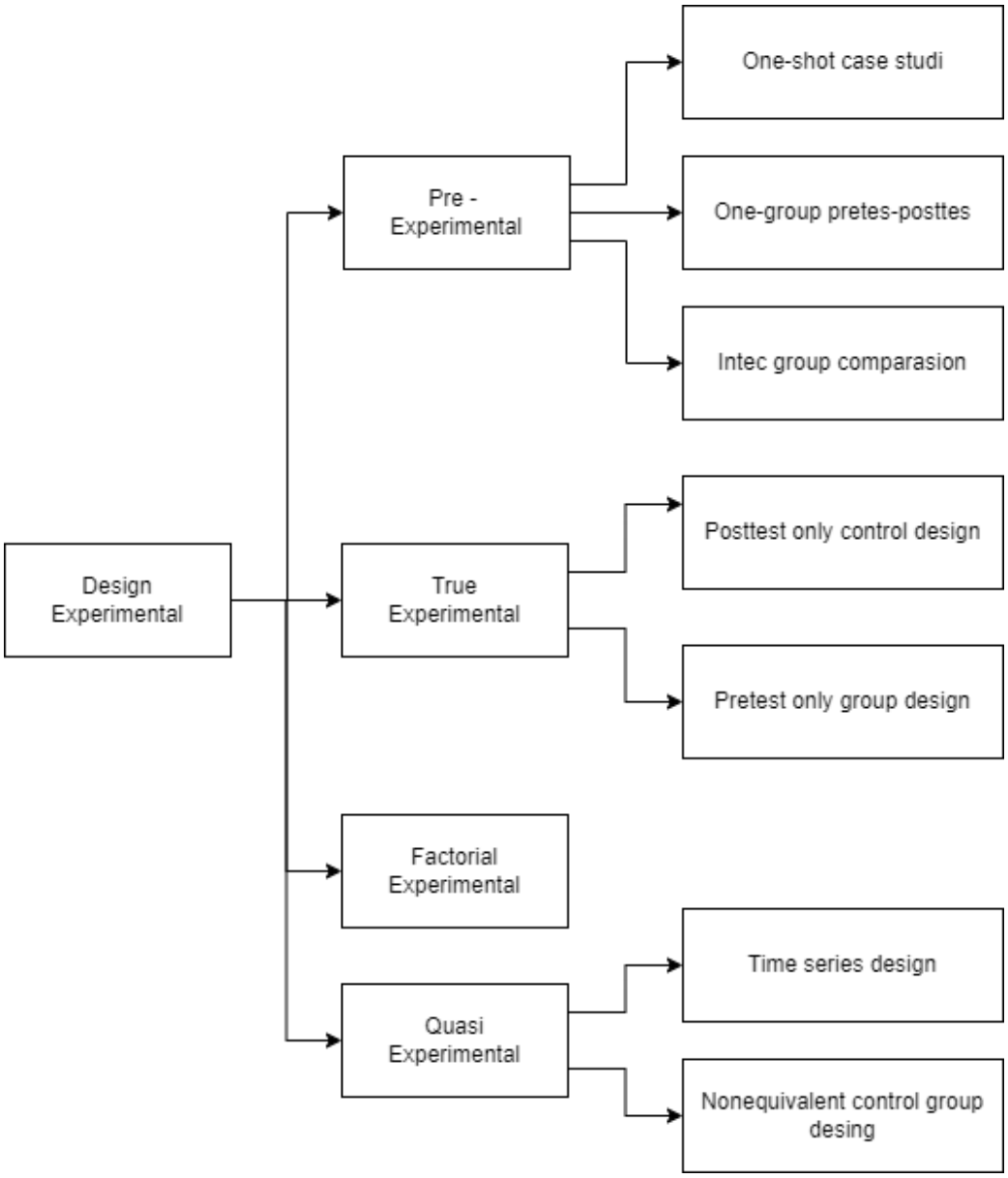
BAB III
METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini menerapkan metode penelitian eksperimental dengan menggunakan Factorial Experimental Design (Desain Faktorial) untuk mengevaluasi kinerja model Real-ESRGAN secara komprehensif. Pemilihan desain ini didasarkan pada kebutuhan untuk menganalisis interaksi antara dua variabel independen atau lebih secara simulasi, yaitu tipe unit pemrosesan (NPU, CPU, dan GPU) serta penerapan teknik kuantisasi pada model. Melalui pendekatan faktorial, peneliti tidak hanya mengukur efek utama (*main effects*) dari masingmasing perangkat keras terhadap kecepatan inferensi, tetapi juga dapat membuktikan bagaimana sinergi antara optimasi model melalui kuantisasi dan arsitektur spesifik NPU menghasilkan efisiensi performa yang lebih unggul dibandingkan unit pemrosesan konvensional lainnya dalam skenario peningkatan kualitas gambar secara *offline* (N et al., 2024). *Experimental Design* menjamin validitas pengujian dengan mengontrol variabel lingkungan secara ketat guna memastikan bahwa peningkatan performa yang terjadi benar-benar disebabkan oleh interaksi faktor-faktor yang diteliti (M. Zhu et al., 2024) .

Penggunaan *Factorial Experimental Design* dalam penelitian ini terletak pada kemampuannya untuk menganalisis efek interaksi yang tidak dapat ditangkap oleh desain eksperimen sederhana. Dalam konteks optimasi Real-ESRGAN, kinerja NPU bergantung pada format data (kuantisasi), sehingga pengujian variabel hardware (CPU, GPU, NPU) secara terisolasi dari variabel model (INT8) akan menghasilkan data yang tidak lengkap. Desain faktorial dapat

membuktikan secara statistik bahwa NPU memiliki efisiensi komputasi yang jauh lebih tinggi justru ketika dipasangkan dengan teknik kuantisasi tertentu, yang merupakan karakteristik utama arsitektur *neural processing (M. Zhu et al., 2024)*.



Gambar 3.1 Struktur penelitian desain experimental

Desain ini memberikan tingkat validitas internal yang tinggi karena memungkinkan kontrol ketat terhadap variabel pengganggu (*confounding variables*). Dengan menguji seluruh kombinasi faktor dalam kondisi lingkungan

yang seragam, kemudian mengisolasi variabel performa seperti *latency* dan *throughput* secara presisi. Hal ini sangat krusial dalam membuktikan keunggulan NPU dalam pemrosesan *offline*, di mana trade-off antara penurunan presisi akibat kuantisasi dan peningkatan kecepatan pada NPU harus diukur secara simultan menggunakan metrik PSNR dan SSIM untuk memastikan kualitas gambar tetap terjaga (Agirbas et al., 2025; M. Zhu et al., 2024).

3.2 Waktu Penelitian

Tabel 3.1 Waktu penelitian

No	Aktivitas	Deskripsi Kegiatan	Bulan / Minggu																							
			November 2025				Desember 2025				Januari 2026				Februari 2026				Maret 2026				April 2026			
			1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.	Pengajuan Judul	• Studi literatur awal • Identifikasi masalah • Pengajuan dan persetujuan judul	■																							
2.	Pembagian Dospem	• Penentuan dosen pembimbing • Pembagian dosen pembimbing		■																						
3.	Revisi Judul	• Revisi dan penyempurnaan judul • Persetujuan judul final					■																			
4.	Pembuatan Proposal	• Studi literatur mendalam • Penyusunan proposal • Konsultasi dan perbaikan proposal					■	■																		
5.	Seminar Proposal	• Persiapan presentasi • Pelaksanaan seminar proposal • Revisi hasil seminar							■																	
6.	Pembuatan Skripsi	• Pengumpulan data • Implementasi dan pengujian • Analisis data • Penyusunan laporan skripsi • Konsultasi dan revisi								■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		
7.	Sidang	• Pendaftaran sidang • Persiapan presentasi akhir • Pelaksanaan sidang skripsi																							■	

Keterangan:
1 = Minggu ke-1 3 = Minggu ke-3
2 = Minggu ke-2 4 = Minggu ke-4

3.3 Variabel dan Parameter Penelitian

Penentuan parameter dan variabel dalam penelitian ini dilakukan untuk menciptakan kerangka pengujian yang terukur dan objektif dalam mengevaluasi implementasi model Real-ESRGAN. Parameter penelitian sebagai batasan teknis tetap yang mencakup penggunaan arsitektur model Real-ESRGAN dengan struktur *Residual-in-Residual Dense Block* (RRDB), dengan penerapan toolkit OpenVINO untuk konversi model ke format *Intermediate Representation* (IR).

Proses optimasi dilakukan menggunakan teknik *Post-Training Quantization* (PTQ) yang disediakan oleh OpenVINO untuk menjamin standar efisiensi komputasi. Seluruh pengujian dijalankan dalam kondisi lingkungan *offline* secara lokal pada perangkat untuk memastikan karakteristik operasional sistem bersifat konsisten tanpa intervensi jaringan luar.

Variabel penelitian dirancang untuk menguji korelasi antar komponen, di mana variabel independen melibatkan manipulasi pada tipe unit pemrosesan (CPU, GPU, dan NPU) serta tingkat presisi data (INT8 hasil kuantisasi OpenVINO). Dampak dari manipulasi tersebut diamati melalui variabel dependen yang mencakup metrik performa berupa *inference latency* (ms) dan penggunaan memori, serta metrik kualitas citra melalui nilai PSNR dan SSIM. Untuk menjaga validitas eksperimen, penelitian ini menerapkan variabel kontrol berupa penggunaan dataset standar dan penyeragaman resolusi citra input pada seluruh skenario pengujian guna menghindari bias beban kerja komputasi.

3.4 Analisis Pengumpulan Data

Analisis pengumpulan data dalam penelitian ini difokuskan pada perolehan aset digital berupa citra yang akan dijadikan sebagai objek eksperimen untuk proses *image up-scaling*. Data utama diperoleh melalui pengumpulan citra dari sisi pengguna (*user-provided images*) yang memiliki variasi karakteristik visual, mulai dari tingkat tekstur, kontras, hingga komposisi warna yang berbeda.



Gambar 3.2 Dataset Gambar

Proses pengumpulan data dilakukan dengan langkah-langkah sebagai berikut:

1. **Pemilihan Citra Referensi (*Ground Truth*):** mengumpulkan citra berkualitas tinggi (High Definition) sebagai standar acuan objektif. Citra ini mewakili kondisi ideal yang ingin dicapai setelah proses restorasi.
2. **Preparasi Citra Input:** Citra dari pengguna kemudian dipersiapkan ke dalam format resolusi rendah (*Low Resolution*) untuk mensimulasikan kondisi citra yang mengalami degradasi atau penurunan kualitas.
3. **Kategorisasi Dataset:** Data citra yang dikumpulkan dikelompokkan berdasarkan kompleksitas visualnya guna memastikan bahwa model Real-ESRGAN yang diuji pada unit NPU dapat dievaluasi secara komprehensif pada berbagai skenario gambar

3.5 Gambaran umum sistem

Sistem yang dikembangkan dalam penelitian ini adalah sistem restorasi dan peningkatan kualitas citra (*image up-scaling*) berbasis kecerdasan buatan yang dirancang untuk beroperasi secara optimal pada perangkat lokal tanpa ketergantungan koneksi internet (*offline*). Inti dari sistem ini terletak pada

implementasi model Real-ESRGAN yang telah dioptimasi menggunakan *framework* OpenVINO yang dibebankan pada *neural processing unit* (NPU).

Secara fungsional, sistem bekerja dengan mentransformasi model asli berpresisi tinggi menjadi format *Intermediate Representation* (IR) melalui teknik Post-Training Quantization. Proses ini menghasilkan XML dan BIN berformat INT8 yang bertindak sebagai jembatan instruksi antara perangkat lunak dan arsitektur perangkat keras NPU. Pemanfaatan NPU sebagai akselerator utama, sistem mampu melakukan tugas komputasi berat dari blok arsitektur RRDB secara paralel dan efisien. Alur kerja sistem dimulai dari penerimaan citra beresolusi rendah (*Low Resolution*), dilanjutkan dengan proses inferensi oleh NPU menggunakan instruksi dari berkas IR, dan diakhiri dengan rekonstruksi citra beresolusi tinggi (*High Resolution*) yang tetap menjaga integritas detail visual aslinya.

3.6 Analisis Teknik Evaluasi

Analisis teknik evaluasi dalam penelitian ini dilakukan untuk mengukur kinerja model Real-ESRGAN yang diimplementasikan pada NPU dengan variasi precision (FP16, dan INT8). Evaluasi difokuskan pada kualitas citra hasil *up-scaling* dengan membandingkan citra hasil terhadap citra referensi (*ground truth*) menggunakan empat parameter utama, yaitu MSE, RMSE, PSNR, dan SSIM. Keempat metrik ini digunakan untuk memberikan penilaian kuantitatif yang komprehensif terhadap kualitas citra.

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_1 - y_1)^2 \quad (9)$$

x_i= nilai piksel citra asli (*ground truth*)

y_i= nilai piksel citra hasil

N= jumlah total piksel

MSE digunakan untuk mengukur rata-rata kesalahan kuadrat antara citra hasil dan citra referensi. Nilai MSE yang lebih kecil menunjukkan bahwa citra hasil semakin mendekati citra asli, sehingga kualitas citra semakin baik.

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_1 - y_1)^2} \quad (10)$$

RMSE merupakan akar kuadrat dari MSE yang digunakan untuk mengembalikan skala error ke satuan intensitas piksel. Nilai RMSE memberikan interpretasi yang lebih intuitif terhadap besarnya kesalahan rata-rata. Semakin kecil nilai RMSE, maka semakin kecil perbedaan antara citra hasil dan citra asli.

Peak Signal to noise Ratio (PSNR)

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \quad (11)$$

MAX= nilai maksimum piksel (umumnya 255 untuk citra 8-bit)

PSNR digunakan untuk mengukur kualitas citra berdasarkan rasio antara sinyal maksimum dan noise (error). Nilai PSNR dinyatakan dalam desibel

(dB). Semakin tinggi nilai PSNR, maka kualitas citra semakin baik karena tingkat noise semakin rendah.

Structural Similarity Index (SSIM)

$$SSIM(\chi, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{2y} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (12)$$

$\mu_x\mu_y$ = rata-rata intensitas citra

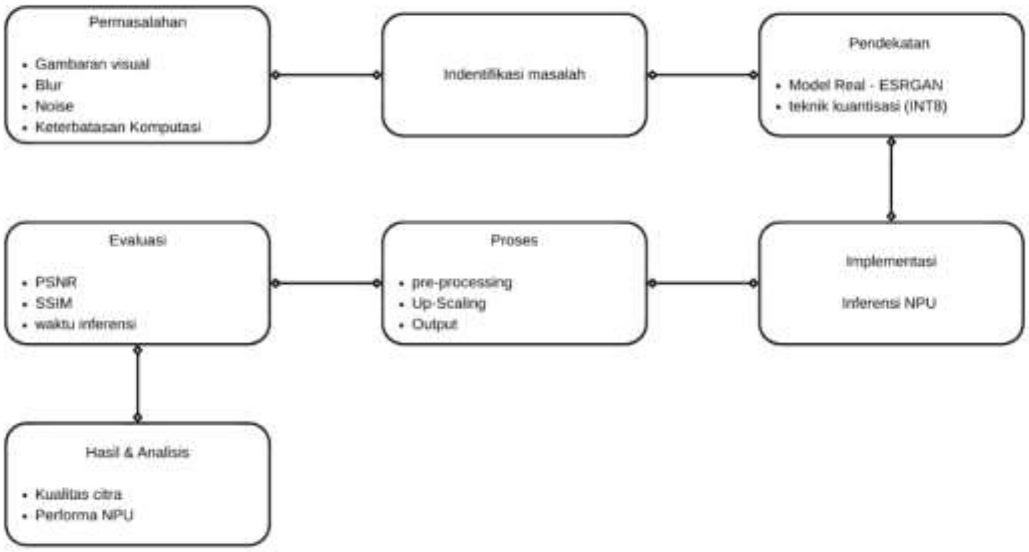
σ_x^2, σ_y^2= variansi

σ_{xy}= kovarians

C_1, C_2= konstanta stabilisasi

SSIM digunakan untuk mengukur kesamaan struktur antara citra hasil dan citra referensi dengan mempertimbangkan luminans, kontras, dan struktur. Nilai SSIM berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan kemiripan yang tinggi.

3.7 Kerangka Berfikir



Gambar 3.3 kerangka berfikir

Mengidentifikasi permasalahan utama yang sering ditemukan pada pengolahan citra digital, yaitu rendahnya kualitas gambaran visual yang ditandai dengan adanya *blur* dan *noise*, serta adanya keterbatasan komputasi pada perangkat keras konvensional dalam menjalankan model AI yang berat. Melalui tahap identifikasi masalah tersebut, ditentukan sebuah pendekatan solusi berupa penggunaan model Real-ESRGAN yang dioptimalkan dengan teknik kuantisasi (INT8). Teknik kuantisasi ini bertujuan untuk menyederhanakan bobot model agar lebih ringan namun tetap memiliki akurasi yang terjaga.

Strategi tersebut diwujudkan dalam tahap implementasi melalui inferensi NPU (*Neural Processing Unit*), di mana unit pemrosesan khusus AI ini digunakan untuk mempercepat beban kerja model secara lokal. Dalam tahap proses, sistem menjalankan alur kerja yang terdiri dari *pre-processing* data, tahap *up-scaling* untuk meningkatkan resolusi citra, hingga menghasilkan *output* akhir. Untuk mengukur keberhasilan sistem, dilakukan evaluasi secara kuantitatif menggunakan metrik PSNR dan SSIM guna menilai kualitas visual, serta pengukuran waktu inferensi untuk menilai kecepatan sistem. Seluruh rangkaian proses ini bermuara pada tahap hasil dan analisis yang akan membedah korelasi antara kualitas citra yang dihasilkan dengan efisiensi performa NPU dalam melakukan akselerasi peningkatan kualitas gambar.

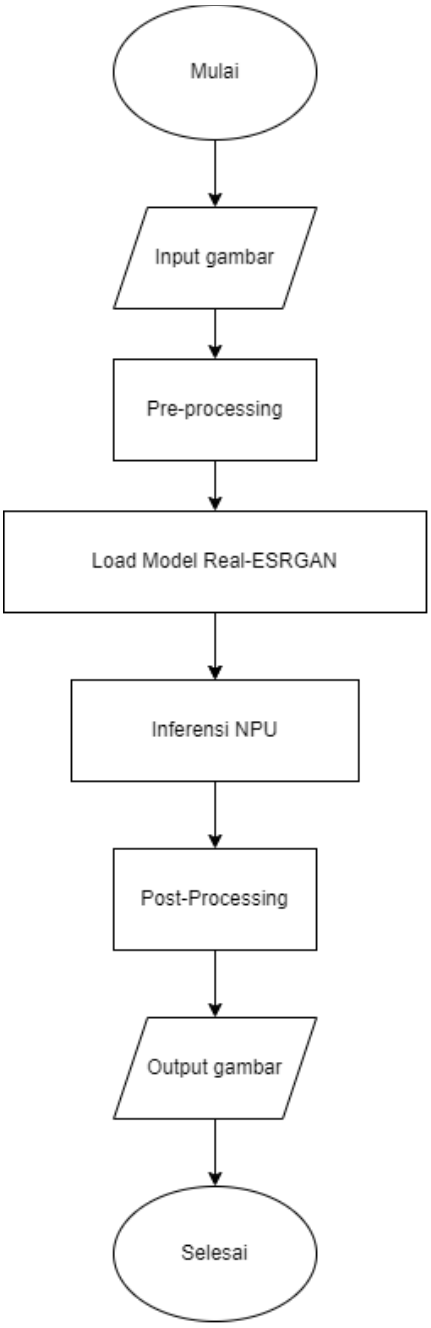
3.8 Analisis Pemodelan Sistem

Flowchart dalam penelitian ini digunakan untuk menggambarkan alur kerja sistem secara keseluruhan, mulai dari tahap input hingga menghasilkan output berupa citra hasil peningkatan resolusi. Proses diawali dengan pengguna memasukkan citra input yang akan diproses oleh sistem. Selanjutnya, pengguna

memilih jenis precision yang akan digunakan, Yaitu FP16 dan INT8, sesuai dengan kebutuhan antara kualitas citra dan efisiensi komputasi. Setelah itu, sistem akan melakukan proses inferensi menggunakan model Real-ESRGAN yang telah dioptimasi dengan toolchain OpenVINO dan dijalankan pada Neural Processing Unit (NPU). Hasil dari proses tersebut berupa citra yang telah ditingkatkan kualitasnya (upscaling). Tahap berikutnya adalah evaluasi hasil menggunakan metrik PSNR, SSIM, MSE, dan RMSE untuk mengukur kualitas citra secara kuantitatif. Seluruh proses ini dilakukan secara offline tanpa memerlukan koneksi internet. Flowchart ini memberikan gambaran yang jelas dan terstruktur mengenai alur sistem, sehingga memudahkan dalam memahami proses implementasi dan evaluasi yang dilakukan dalam penelitian.

3.8.1 Flowchart sistem

Sistem yang digunakan dalam penelitian ini dirancang untuk mendukung proses image up-scaling secara offline menggunakan model Real-ESRGAN terkuantisasi pada NPU. Sistem dibangun secara modular, dengan setiap modul memiliki fungsi spesifik yang saling terintegrasi untuk memastikan alur kerja yang efisien dan terkontrol. Komponen utama sistem meliputi User Interface, Modul Preprocessing, Modul Real-ESRGAN, Engine Inferensi NPU, Modul Post-processing, dan Modul Evaluasi.



Gambar 3.4 flowchart sistem

User Interface berperan sebagai penghubung antara pengguna dan sistem, menerima input citra resolusi rendah dan menampilkan output berupa citra resolusi tinggi serta hasil evaluasi kualitas. Modul Preprocessing bertugas menyiapkan citra masukan, termasuk melakukan penyesuaian ukuran dan normalisasi nilai piksel agar sesuai dengan

spesifikasi model. Modul Real-ESRGAN bersama Engine Inferensi NPU melakukan proses utama yaitu inferensi image up-scaling menggunakan model yang telah diterapkan teknik kuantisasi. Modul Post-processing bertanggung jawab untuk melakukan denormalisasi dan konversi format citra agar siap ditampilkan atau disimpan. Modul Evaluasi menghitung metrik objektif, seperti PSNR dan SSIM, untuk menilai kualitas citra hasil up-scaling.

Struktur sistem ini memungkinkan pemisahan fungsi secara jelas, meminimalkan kompleksitas, dan mempermudah pemeliharaan serta pengembangan sistem lebih lanjut. Dengan pendekatan modular ini, sistem dapat beroperasi secara efisien pada perangkat pengguna, memanfaatkan NPU untuk percepatan inferensi, serta menyediakan informasi yang akurat terkait kualitas hasil image up-scaling.

3.8.2 Flowchart REAL-ERSGAN

Alur kerja sistem diawali dengan proses inisialisasi sistem, di mana perangkat keras dan perangkat lunak yang digunakan untuk menjalankan model Real-ESRGAN dipersiapkan, termasuk pemuatan model yang telah dikonfigurasi atau terkuantisasi. Setelah sistem siap, pengguna memasukkan citra beresolusi rendah sebagai data input untuk diproses lebih lanjut oleh sistem.

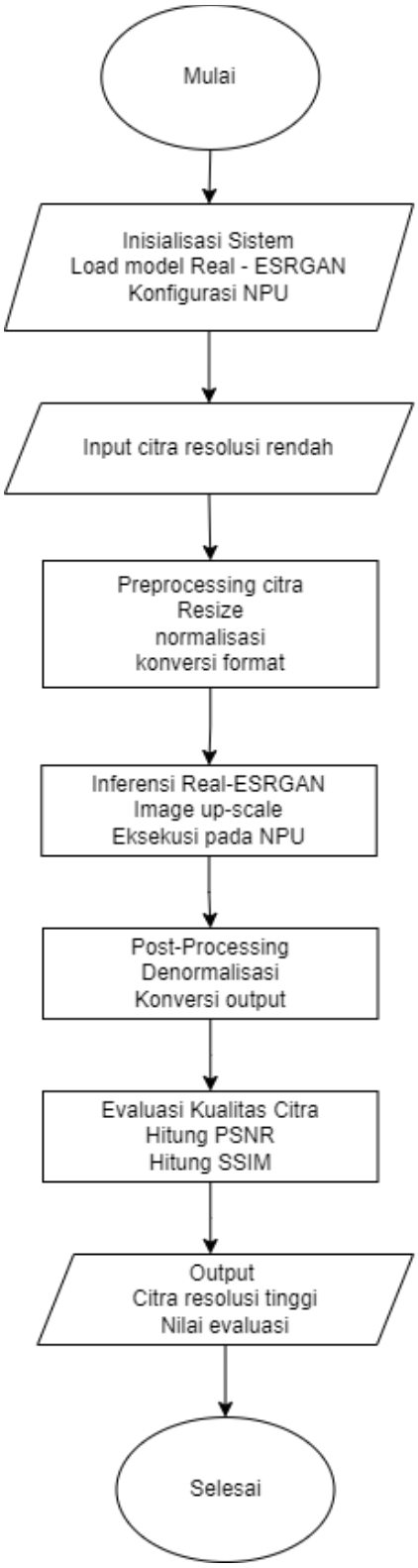
Citra input kemudian masuk ke tahap preprocessing, yang bertujuan untuk menyesuaikan format dan karakteristik citra agar sesuai dengan kebutuhan model. Tahap ini mencakup normalisasi nilai piksel, penyesuaian ukuran citra, serta konversi format data ke dalam representasi yang dapat

diproses oleh engine inferensi. Proses preprocessing memastikan bahwa citra input berada dalam kondisi optimal sebelum dilakukan inferensi.

Citra hasil preprocessing diproses oleh model Real-ESRGAN melalui tahap inferensi. Pada tahap ini, sistem melakukan image up-scaling dengan memanfaatkan kemampuan model dalam merekonstruksi detail citra dan meningkatkan resolusi. Jika sistem menggunakan perangkat akselerasi seperti NPU atau GPU, proses inferensi dijalankan melalui engine inferensi untuk meningkatkan efisiensi waktu komputasi.

Hasil inferensi kemudian diteruskan ke tahap post-processing, yang berfungsi untuk mengonversi kembali citra hasil model ke format visual yang dapat ditampilkan atau disimpan. Tahap ini mencakup denormalisasi nilai piksel, penyesuaian rentang warna, serta penyimpanan citra hasil peningkatan resolusi.

Setelah citra resolusi tinggi dihasilkan, sistem melakukan evaluasi kualitas citra dengan membandingkan hasil up-scaling terhadap citra referensi menggunakan metrik objektif seperti Peak Signal-to-Noise Ratio (PSNR) dan Structural Similarity Index (SSIM). Nilai evaluasi ini digunakan untuk menilai tingkat peningkatan kualitas citra yang dihasilkan oleh model Real-ESRGAN.



Gambar 3.5 flowchart Real-ESRGAN

Alur kerja sistem diakhiri dengan proses penyajian output, di mana citra hasil image up-scaling beserta nilai evaluasi ditampilkan kepada

pengguna atau disimpan sebagai hasil akhir penelitian. Proses ini menandai berakhirnya siklus kerja model Real-ESRGAN dalam sistem yang dikembangkan.

3.9 Analisis Struktur Sistem

Struktur sistem yang digunakan dalam penelitian ini dirancang untuk mendukung proses image up-scaling secara offline menggunakan model Real-ESRGAN terkuantisasi pada NPU. Sistem dibangun secara modular, dengan setiap modul memiliki fungsi spesifik yang saling terintegrasi untuk memastikan alur kerja yang efisien dan terkontrol. Komponen utama sistem meliputi User Interface, Modul Preprocessing, Modul Real-ESRGAN, Engine Inferensi NPU, Modul Post-processing, dan Modul Evaluasi.

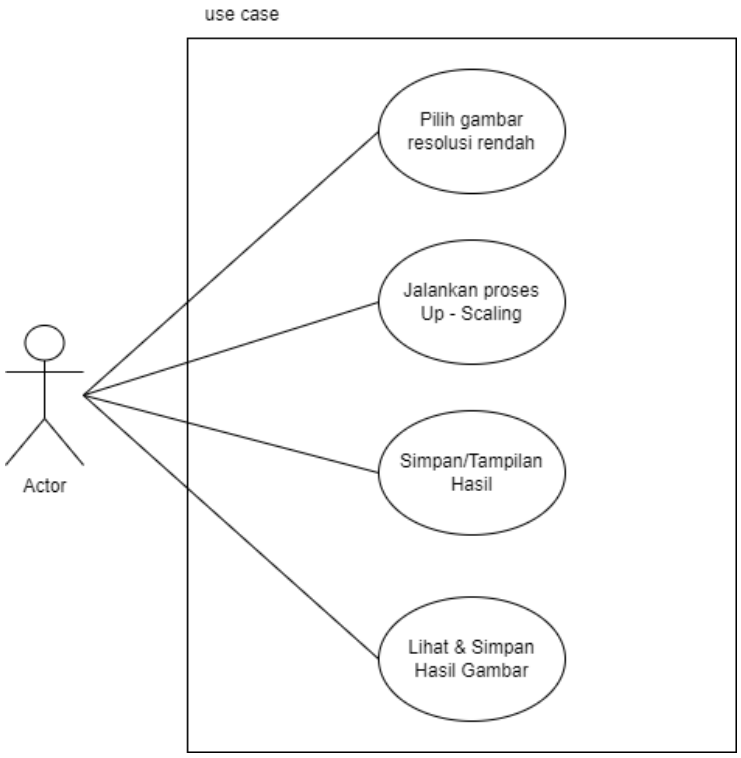
User Interface berperan sebagai penghubung antara pengguna dan sistem, menerima input citra resolusi rendah dan menampilkan output berupa citra resolusi tinggi serta hasil evaluasi kualitas. Modul Preprocessing bertugas menyiapkan citra masukan, termasuk melakukan penyesuaian ukuran dan normalisasi nilai piksel agar sesuai dengan spesifikasi model. Modul Real-ESRGAN bersama Engine Inferensi NPU melakukan proses utama yaitu inferensi image up-scaling menggunakan model yang telah diterapkan teknik kuantisasi. Modul Post-processing bertanggung jawab untuk melakukan denormalisasi dan konversi format citra agar siap ditampilkan atau disimpan. Modul Evaluasi menghitung metrik objektif, seperti PSNR dan SSIM, untuk menilai kualitas citra hasil up-scaling.

Struktur sistem ini memungkinkan pemisahan fungsi secara jelas, meminimalkan kompleksitas, dan mempermudah pemeliharaan serta

pengembangan sistem lebih lanjut. Dengan pendekatan modular ini, sistem dapat beroperasi secara efisien pada perangkat pengguna, memanfaatkan NPU untuk percepatan inferensi, serta menyediakan informasi yang akurat terkait kualitas hasil image up-scaling.

3.9.1 Use Case Diagram

Use Case Diagram pada sistem image up-scaling ini menjelaskan hubungan fungsional antara pengguna sebagai aktor dan sistem dalam menjalankan proses peningkatan resolusi citra menggunakan model Real-ESRGAN. Pengguna memiliki peran utama dalam memulai proses dengan memasukkan citra resolusi rendah sebagai masukan sistem. Citra yang dimasukkan selanjutnya diproses oleh sistem melalui fungsi image up-scaling yang menjadi use case utama.



Gambar 3.6 Use Case diagram

Use Case Diagram pada sistem image up-scaling ini menjelaskan hubungan fungsional antara pengguna sebagai aktor dan sistem dalam menjalankan proses peningkatan resolusi citra menggunakan model Real-ESRGAN. Pengguna memiliki peran utama dalam memulai proses dengan memasukkan citra resolusi rendah sebagai masukan sistem. Citra yang dimasukkan selanjutnya diproses oleh sistem melalui fungsi image up-scaling yang menjadi use case utama.

Pada saat proses image up-scaling dijalankan, sistem secara otomatis melaksanakan tahapan pemrosesan yang mencakup preprocessing citra, pemanggilan model Real-ESRGAN terkuantisasi, serta pelaksanaan inferensi pada perangkat Neural Processing Unit (NPU). Proses tersebut menghasilkan citra keluaran dengan resolusi yang lebih tinggi dibandingkan citra masukan. Citra hasil pemrosesan kemudian disajikan kepada pengguna melalui fungsi penampilan hasil citra resolusi tinggi.

Selain menghasilkan citra resolusi tinggi, sistem juga menyediakan fungsi evaluasi kualitas citra. Evaluasi dilakukan dengan menghitung nilai metrik objektif, yaitu Peak Signal-to-Noise Ratio (PSNR) dan Structural Similarity Index Measure (SSIM), dengan membandingkan citra hasil up-scaling terhadap citra referensi. Hasil perhitungan metrik tersebut ditampilkan kepada pengguna sebagai informasi pendukung untuk menilai kualitas citra yang dihasilkan oleh sistem.

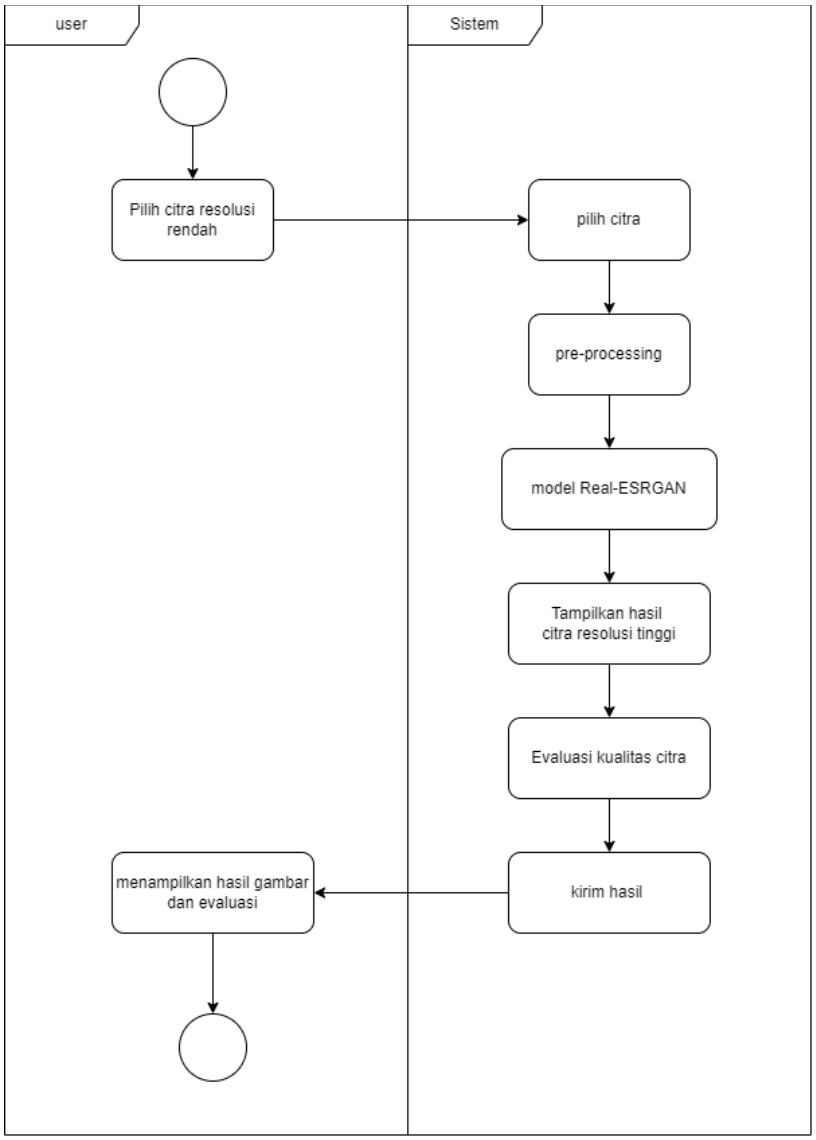
Secara keseluruhan, Use Case Diagram menunjukkan bahwa interaksi pengguna dengan sistem terbatas pada pemberian masukan dan penerimaan keluaran, sedangkan seluruh proses komputasi dan evaluasi dilakukan

secara otomatis oleh sistem. Diagram ini merepresentasikan fungsi sistem secara terstruktur dan menggambarkan alur kerja sistem image up-scaling yang diimplementasikan secara offline.

3.9.2 Activity Diagram

Activity Diagram pada sistem image up-scaling ini menggambarkan alur aktivitas yang terjadi selama proses peningkatan resolusi citra menggunakan model Real-ESRGAN terkuantisasi yang dijalankan pada perangkat Neural Processing Unit (NPU) secara offline. Proses diawali ketika pengguna menjalankan sistem dan memasukkan citra resolusi rendah sebagai masukan. Citra tersebut selanjutnya diterima oleh sistem untuk diproses lebih lanjut.

Activity Diagram pada sistem image up-scaling ini menggambarkan alur aktivitas yang terjadi selama proses peningkatan resolusi citra menggunakan model Real-ESRGAN terkuantisasi yang dijalankan pada perangkat Neural Processing Unit (NPU) secara offline. Proses diawali ketika pengguna menjalankan sistem dan memasukkan citra resolusi rendah sebagai masukan. Citra tersebut selanjutnya diterima oleh sistem untuk diproses lebih lanjut.



Gambar 3.7 activity diagram

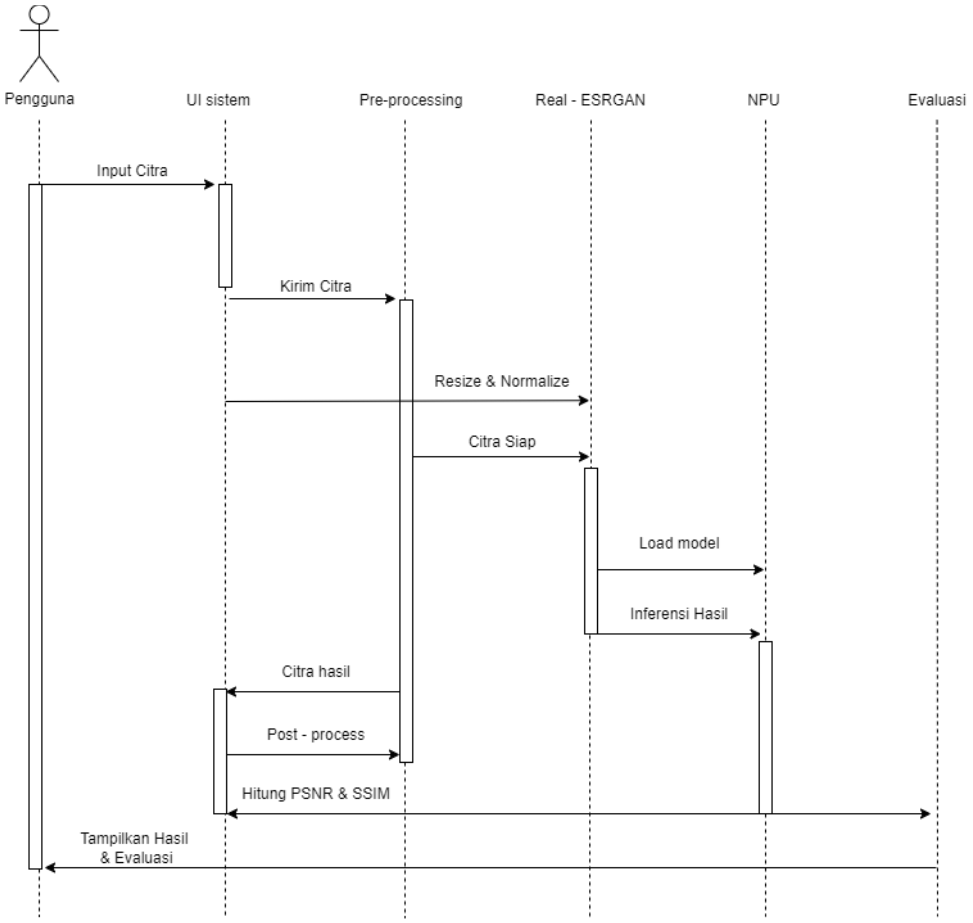
Setelah citra diterima, sistem melakukan tahap preprocessing yang mencakup penyesuaian ukuran citra dan normalisasi nilai piksel agar sesuai dengan spesifikasi masukan model. Tahap ini bertujuan untuk memastikan bahwa citra berada dalam format dan rentang nilai yang dapat diproses secara optimal oleh model Real-ESRGAN. Setelah preprocessing selesai, sistem memuat model Real-ESRGAN yang telah diterapkan teknik kuantisasi ke dalam memori.

Tahap berikutnya adalah proses inferensi, di mana sistem menjalankan model Real-ESRGAN pada perangkat NPU untuk melakukan image up-scaling. Proses ini menghasilkan citra dengan resolusi yang lebih tinggi dibandingkan citra masukan. Hasil inferensi kemudian diproses kembali melalui tahap post-processing, yang meliputi denormalisasi dan konversi format citra ke dalam bentuk yang dapat ditampilkan dan disimpan.

Setelah seluruh tahapan pemrosesan citra selesai, sistem menampilkan citra hasil up-scaling kepada pengguna. Selain itu, sistem juga melakukan evaluasi kualitas citra dengan menghitung nilai Peak Signal-to-Noise Ratio (PSNR) dan Structural Similarity Index Measure (SSIM). Hasil perhitungan metrik evaluasi tersebut ditampilkan kepada pengguna sebagai informasi pendukung dalam menilai kualitas citra hasil image up-scaling. Proses kemudian diakhiri setelah seluruh aktivitas sistem selesai dilakukan.

3.9.3 Sequence Diagram

Sequence Diagram pada sistem image up-scaling ini menggambarkan urutan interaksi antar objek dan komponen sistem selama proses peningkatan resolusi citra menggunakan model Real-ESRGAN terkuantisasi yang dijalankan pada NPU secara offline. Diagram ini menekankan hubungan komunikasi antara pengguna, antarmuka sistem, modul preprocessing, model Real-ESRGAN, NPU, dan modul evaluasi.



Gambar 3.8 Sequence diagram

Sequence Diagram pada sistem image up-scaling ini menggambarkan urutan interaksi antar objek dan komponen sistem selama proses peningkatan resolusi citra menggunakan model Real-ESRGAN terkuantisasi yang dijalankan pada NPU secara offline. Diagram ini menekankan hubungan komunikasi antara pengguna, antarmuka sistem, modul preprocessing, model Real-ESRGAN, NPU, dan modul evaluasi.

Proses dimulai ketika pengguna mengirim citra resolusi rendah sebagai input melalui antarmuka sistem. Antarmuka sistem kemudian meneruskan citra tersebut ke modul preprocessing untuk dilakukan penyesuaian ukuran dan normalisasi nilai piksel sesuai kebutuhan model. Setelah preprocessing selesai, data dikirim ke modul Real-ESRGAN, yang

kemudian dijalankan pada perangkat NPU untuk melakukan proses inferensi image up-scaling.

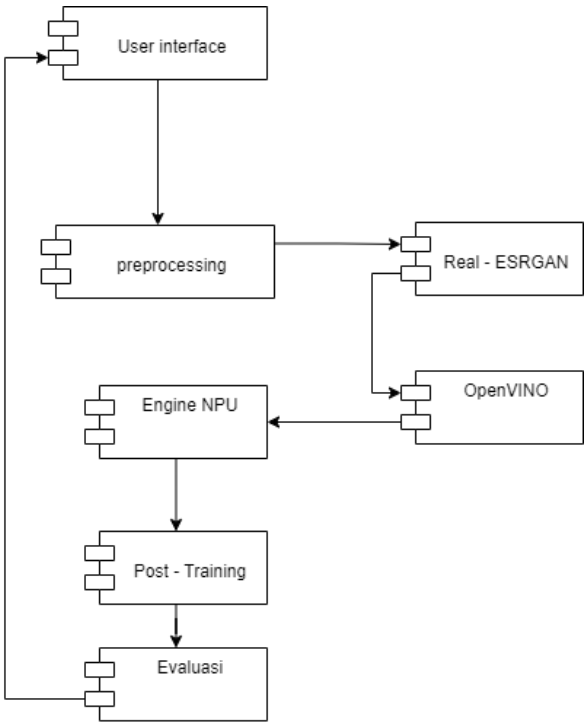
Hasil inferensi dikembalikan dari NPU ke modul Real-ESRGAN dan selanjutnya diteruskan ke antarmuka sistem untuk tahap post-processing, seperti denormalisasi dan konversi format citra. Setelah citra resolusi tinggi selesai diproses, antarmuka sistem memanggil modul evaluasi untuk menghitung metrik kualitas citra, yaitu Peak Signal-to-Noise Ratio (PSNR) dan Structural Similarity Index Measure (SSIM). Hasil evaluasi ini kemudian disampaikan kembali ke antarmuka sistem dan ditampilkan kepada pengguna sebagai informasi mengenai kualitas citra hasil up-scaling.

Sequence Diagram ini memperlihatkan urutan kronologis pesan dan data yang dikirim antar komponen, menunjukkan bahwa interaksi pengguna terbatas pada pemberian input dan penerimaan output, sementara seluruh proses inferensi dan evaluasi dijalankan secara otomatis oleh sistem. Dengan demikian, diagram ini membantu memahami aliran komunikasi dalam sistem serta memperjelas hubungan fungsional antar modul secara rinci.

3.9.4 Component Diagram

Component Diagram pada sistem image up-scaling ini menggambarkan struktur modular sistem dan hubungan antar komponen yang digunakan untuk meningkatkan resolusi citra menggunakan model Real-ESRGAN terkuantisasi pada NPU secara offline. Diagram ini menekankan pembagian sistem menjadi modul-modul fungsional yang

saling terhubung untuk melaksanakan seluruh proses dari input hingga evaluasi.



Gambar 3.9 Component diagram

Sistem terdiri dari beberapa komponen utama, yaitu User Interface, Modul Preprocessing, Modul Real-ESRGAN, Engine Inferensi NPU, Modul Post-processing, dan Modul Evaluasi. Pengguna berinteraksi dengan sistem melalui User Interface, yang bertanggung jawab untuk menerima citra resolusi rendah dan menampilkan citra hasil up-scaling serta hasil evaluasi kualitas citra. User Interface meneruskan data citra ke Modul Preprocessing untuk dilakukan penyesuaian ukuran dan normalisasi nilai piksel.

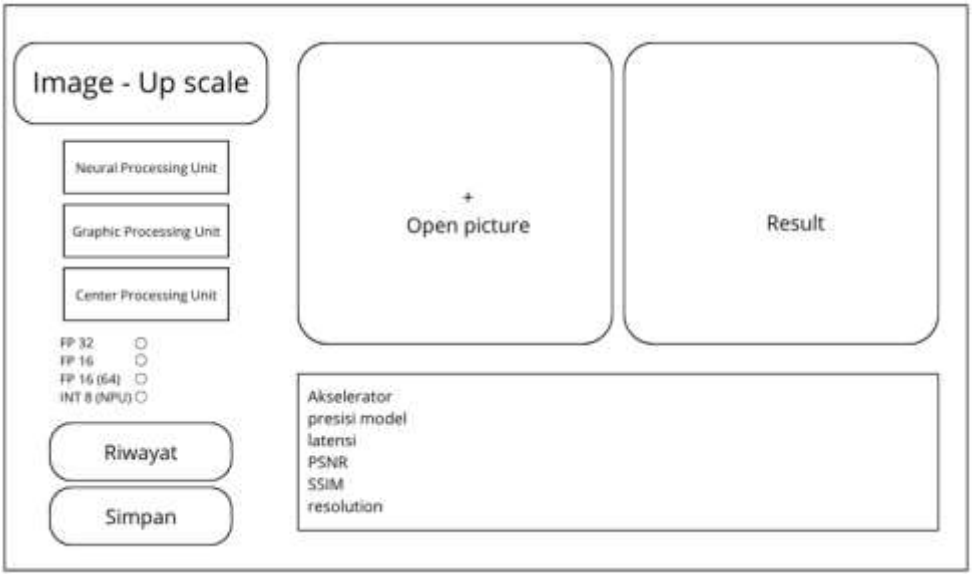
Setelah preprocessing selesai, data dikirim ke Modul Real-ESRGAN yang kemudian dijalankan pada Engine Inferensi NPU untuk melakukan image up-scaling. Hasil inferensi dikirim kembali ke Modul Post-processing untuk dilakukan denormalisasi dan konversi format citra agar siap

ditampilkan. Selanjutnya, citra hasil up-scaling dievaluasi oleh Modul Evaluasi menggunakan metrik PSNR dan SSIM, dan hasil evaluasi dikirim kembali ke User Interface.

Component Diagram ini menunjukkan struktur sistem secara modular, hubungan antar komponen, serta aliran data antar modul. Dengan pembagian modul yang jelas, diagram ini memudahkan pemahaman mengenai tanggung jawab masing-masing bagian sistem dan mendukung implementasi sistem secara terstruktur serta terkontrol.

3.10 Interface

Antarmuka pengguna (GUI) pada sistem difokuskan pada penyediaan jembatan interaksi antara pengguna dan mesin inferensi Real-ESRGAN yang berjalan di atas *framework* OpenVINO. Interface dirancang dengan pendekatan fungsional yang memungkinkan pengguna untuk melakukan proses *image up-scaling* tanpa memerlukan pemahaman teknis mendalam mengenai baris perintah (*command line*).



Gambar 3.10 interface

Komponen utama dalam antarmuka ini meliputi modul pengunggah citra (*image uploader*) yang mendukung berbagai format biner, panel kontrol untuk memilih unit pemrosesan (khususnya alokasi pada NPU), serta indikator status proses yang memberikan informasi waktu nyata (*real-time*) mengenai progres restorasi citra.

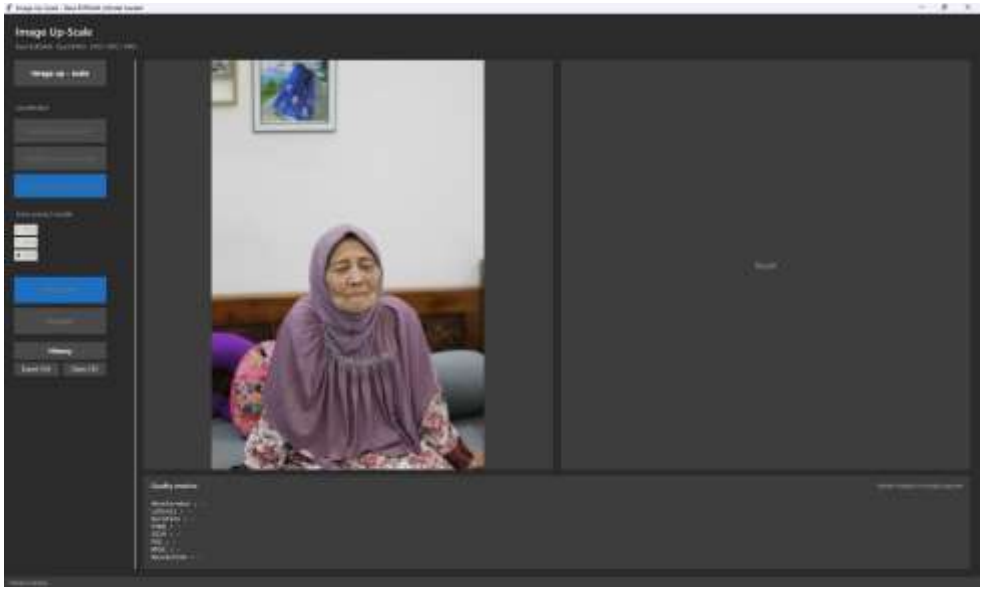
BAB IV
HASIL DAN PEMBAHASAN

4.1 Implementasi Antarmuka

Implementasi antarmuka pada penelitian ini dirancang untuk memudahkan pengguna dalam melakukan proses peningkatan kualitas citra menggunakan model Real-ESRGAN yang telah dioptimasi dengan teknik kuantisasi. Antarmuka sistem dikembangkan berbasis aplikasi sederhana yang memungkinkan pengguna untuk melakukan input gambar, memilih metode precision yang digunakan (FP16 dan INT8), serta menjalankan proses inferensi secara offline. Selain itu, antarmuka juga menampilkan hasil citra setelah proses upscaling beserta informasi metrik evaluasi seperti PSNR, SSIM, MSE, dan RMSE untuk memberikan gambaran kualitas hasil secara kuantitatif. Desain antarmuka dibuat dengan pendekatan user-friendly agar pengguna dapat dengan mudah memahami alur penggunaan sistem tanpa memerlukan pengetahuan teknis yang mendalam. Dengan adanya antarmuka ini, proses evaluasi dan perbandingan kinerja model menjadi lebih sistematis, terstruktur, dan mudah dianalisis dalam mendukung hasil penelitian.

4.1.1 Antarmuka Sistem Image Up - Scale

Antarmuka sistem yang dikembangkan dirancang dalam bentuk *Graphical User Interface (GUI)* yang sederhana dan interaktif untuk memudahkan pengguna dalam melakukan proses *image upscaling* menggunakan model Real-ESRGAN berbasis OpenVINO. Pada bagian utama, terdapat dua panel tampilan, yaitu panel kiri sebagai area input yang digunakan untuk memuat gambar melalui metode *drag and drop* atau unggah manual, serta panel kanan yang menampilkan hasil citra setelah proses peningkatan resolusi dilakukan.



Gambar 4.1 Antarmuka Sistem *Image Up - Scale*

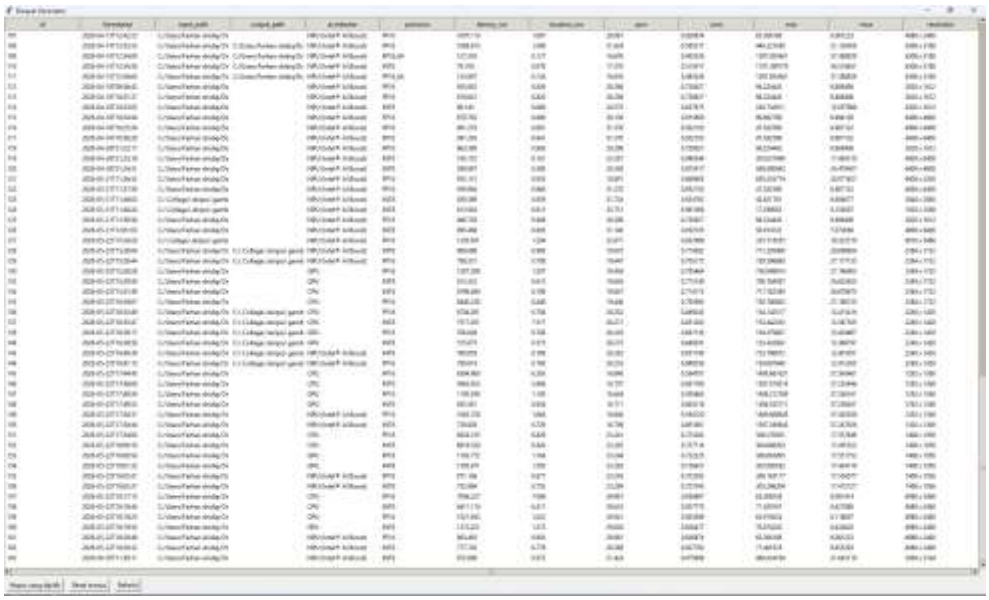
Pada sisi kiri antarmuka, tersedia beberapa opsi konfigurasi yang memungkinkan pengguna memilih perangkat komputasi (*accelerator*), seperti CPU, GPU, dan NPU, sesuai dengan kebutuhan dan kemampuan perangkat keras yang digunakan. Selain itu, pengguna juga dapat memilih jenis presisi model yang akan digunakan, seperti FP32, FP16 dan INT8, sehingga memungkinkan analisis perbandingan antara kualitas dan performa inferensi.

Antarmuka ini juga dilengkapi dengan tombol seperti *Generate* untuk memulai proses inferensi, Simpan untuk menyimpan hasil gambar, serta fitur *History* yang memungkinkan pengguna meninjau hasil sebelumnya. Pada bagian bawah, sistem menampilkan metrik evaluasi seperti *latency*, *PSNR*, *SSIM*, *MSE*, dan resolusi output, yang berfungsi untuk mengukur kualitas hasil *upscaling* serta kinerja model. Secara keseluruhan, desain antarmuka ini mendukung kemudahan penggunaan

sekaligus memberikan fleksibilitas dalam pengujian berbagai konfigurasi model dan perangkat.

4.1.2 Antarmuka Riwayat dan Evaluasi Proses Citra

Antarmuka *History* (Riwayat) berfungsi sebagai modul pencatatan dan analisis hasil pengujian yang telah dilakukan pada sistem *image upscaling*. Tampilan ini disajikan dalam bentuk tabel terstruktur yang memuat berbagai parameter penting dari setiap proses inferensi. Setiap baris merepresentasikan satu percobaan, dengan kolom informasi seperti *timestamp* (waktu eksekusi), *input_path* dan *output_path* (lokasi file), jenis *accelerator* yang digunakan (CPU, GPU, atau NPU), serta jenis presisi model seperti FP16 atau INT8.



Gambar 4.2 Metrik evaluasi hasil proses citra

Selain itu, tabel ini juga menampilkan metrik performa dan kualitas hasil, seperti *latency* (waktu pemrosesan), *duration*, serta metrik evaluasi citra seperti PSNR, SSIM, MSE, dan RMSE. Informasi resolusi output juga disertakan untuk menunjukkan perubahan ukuran citra setelah proses

upscaling. Dengan adanya tampilan ini, pengguna dapat melakukan analisis komparatif secara sistematis terhadap berbagai konfigurasi model dan perangkat keras yang digunakan.

Pada bagian bawah antarmuka, tersedia tombol kontrol seperti *hapus yang dipilih*, *reset semua*, dan *refresh*, yang memungkinkan pengguna mengelola data riwayat dengan fleksibel. Secara keseluruhan, fitur *History* ini berperan penting dalam mendukung evaluasi eksperimen, pelacakan hasil, serta pengambilan keputusan berbasis data dalam menentukan konfigurasi model yang paling optimal.

4.2 Implementasi model real – esrgan

Implementasi model Real-ESRGAN dalam penelitian ini dilakukan dengan memanfaatkan model pre-trained yang telah tersedia, sehingga proses penelitian difokuskan pada tahap inferensi dan optimasi tanpa melakukan pelatihan ulang model. Model Real-ESRGAN digunakan untuk melakukan proses image super-resolution dengan tujuan meningkatkan kualitas citra beresolusi rendah menjadi citra dengan resolusi lebih tinggi dan detail yang lebih tajam. Dalam implementasinya, model dikonversi dan dioptimasi menggunakan toolchain OpenVINO agar dapat dijalankan secara efisien pada Neural Processing Unit (NPU). Selain itu, diterapkan teknik kuantisasi dengan beberapa variasi precision, yaitu FP16 dan INT8, untuk menganalisis pengaruhnya terhadap kualitas citra dan efisiensi komputasi. Proses inferensi dilakukan secara offline dengan memasukkan citra sebagai input, kemudian diproses oleh model untuk menghasilkan citra hasil upscaling. Hasil tersebut selanjutnya digunakan dalam tahap evaluasi untuk mengukur performa model berdasarkan metrik kualitas citra

dan latency. Dengan pendekatan ini, implementasi model tidak hanya berfokus pada peningkatan kualitas visual, tetapi juga pada optimasi performa komputasi pada perangkat dengan keterbatasan sumber daya.

4.2.1 Konversi format ONNX

Pada tahap konversi format ONNX, proses dimulai dengan memuat arsitektur model Real-ESRGAN menggunakan fungsi get_rrdb_model(), kemudian dilanjutkan dengan pemanggilan bobot model dari file berekstensi *.pth* menggunakan torch.load. Struktur *state dictionary* yang dimuat selanjutnya disesuaikan dengan memeriksa keberadaan kunci seperti *params_ema* atau *params*, sehingga bobot yang digunakan benar-benar representatif terhadap hasil pelatihan terbaik. Setelah bobot berhasil dimuat ke dalam model melalui load_state_dict, model diatur ke mode evaluasi (eval()) untuk memastikan tidak ada perubahan parameter selama proses inferensi maupun ekspor.

Selanjutnya, dibuat sebuah *dummy input* menggunakan torch.randn dengan ukuran tertentu, dalam hal ini *(1, 3, 256, 256)*, yang berfungsi untuk mendefinisikan bentuk input serta struktur komputasi (*computational graph*) model. Proses inti konversi dilakukan menggunakan fungsi torch.onnx.export, di mana model PyTorch dikonversi menjadi format ONNX dengan menentukan parameter seperti *opset_version=14*, nama input dan output, serta pengaturan *dynamic_axes* yang dinonaktifkan untuk menjaga konsistensi dimensi input. Hasil dari proses ini adalah file model dalam format *.onnx* yang bersifat lebih portabel dan dapat digunakan pada

berbagai framework lain seperti OpenVINO untuk optimasi dan deployment lebih lanjut.

```
def export_to_onnx(model_path: str, onnx_path: str,
input_size: tuple = (1, 3, 256, 256)):
    print("Loading PyTorch model...")
    model = get_rrdb_model()
    state_dict = torch.load(model_path,
map_location="cpu")

    if "params_ema" in state_dict:
        state_dict = state_dict["params_ema"]
    elif "params" in state_dict:
        state_dict = state_dict["params"]

    model.load_state_dict(state_dict, strict=True)
    model.eval()

    dummy_input = torch.randn(*input_size)

    print(f"Exporting to ONNX (input shape:
{input_size})...")
    torch.onnx.export(
        model,
        dummy_input,
        onnx_path,
        opset_version=14,
        input_names=["input"],
        output_names=["output"],
        dynamic_axes=None,
    )
    print(f"ONNX saved: {onnx_path}")
```

4.2.2 Konversi ONNX ke format OpenVino

Pada tahap konversi ONNX ke format OpenVINO, proses dimulai dengan memuat model ONNX menggunakan fungsi ov.convert_model dari library OpenVINO, yang secara otomatis mengubah representasi model menjadi format internal OpenVINO. Model hasil konversi ini kemudian disiapkan untuk disimpan dalam bentuk *Intermediate Representation* (IR), yaitu pasangan file *.xml* (struktur model) dan *.bin* (bobot model). Direktori

penyimpanan terlebih dahulu dibuat menggunakan mkdir untuk memastikan lokasi output tersedia.

Selanjutnya, model disimpan dalam dua variasi presisi, yaitu FP32 dan FP16, menggunakan fungsi ov.save_model. Pada penyimpanan pertama, parameter compress_to_fp16=False digunakan untuk menghasilkan model dengan presisi penuh (FP32) sebagai baseline performa dan akurasi. Kemudian, model yang sama disimpan kembali dengan parameter compress_to_fp16=True untuk menghasilkan versi FP16 yang lebih efisien dalam penggunaan memori dan komputasi. Kedua hasil konversi ini disimpan secara terpisah dalam file model_fp32.xml dan *model_fp16.xml*, sehingga dapat digunakan untuk perbandingan performa maupun deployment pada berbagai perangkat keras yang didukung OpenVINO.

```
def convert_onnx_to_openvino(onnx_path: str, ov_dir: Path,
model_name: str = "RealESRGAN_x4plus"):

    import openvino as ov

    print("Converting ONNX to OpenVINO IR...")
    ov_model = ov.convert_model(onnx_path)

    ov_dir.mkdir(parents=True, exist_ok=True)
    fp32_xml = ov_dir / "model_fp32.xml"
    fp16_xml = ov_dir / "model_fp16.xml"

    ov.save_model(ov_model, str(fp32_xml),
compress_to_fp16=False)
    print(f"OpenVINO FP32 saved: {fp32_xml}")

    ov.save_model(ov_model, str(fp16_xml),
compress_to_fp16=True)
    print(f"OpenVINO FP16 saved: {fp16_xml}")

    return str(fp32_xml)
```

4.2.4 Convert model dengan Teknik kuantisasi untuk format INT8

Pada implementasi kuantisasi INT8 ini, proses diawali dengan memuat model OpenVINO FP32 menggunakan ov.Core() dan core.read_model. Selanjutnya, dataset kalibrasi dibangun dari folder citra yang berjumlah 2523 gambar, yang terdiri dari berbagai jenis konten seperti pemandangan alam, bangunan, objek, serta citra manusia (wajah dan aktivitas). Keberagaman dataset ini penting untuk merepresentasikan distribusi data yang lebih luas selama proses kalibrasi. Setiap citra kemudian diproses menggunakan fungsi transform_fn, yang mencakup *resize* ke ukuran 256 x 256, normalisasi nilai piksel, serta transformasi dimensi ke format *(N, C, H, W)* sesuai kebutuhan model. Dataset yang telah diproses menjadi nncf.Dataset untuk digunakan dalam proses kalibrasi. Tahap utama dilakukan menggunakan nncf.quantize, yang mengubah model dari presisi FP32 menjadi INT8.

Setelah proses kuantisasi selesai, model hasil optimasi disimpan kembali dalam format OpenVINO IR menggunakan ov.save_model. Pendekatan ini bertujuan meningkatkan efisiensi inferensi dan mengurangi penggunaan memori, dengan tetap mempertimbangkan potensi penurunan akurasi akibat penurunan presisi.

```
core = ov.Core()

model = core.read_model("model_fp32_static.xml")

print("Model loaded")

def transform_fn(image_path):

    image = cv2.imread(image_path)

    if image is None:
        print("ERROR load:", image_path)
        return None

    image = cv2.resize(image, (256, 256))
    image = image.astype(np.float32) / 255.0
    image = np.transpose(image, (2, 0, 1))
    image = np.expand_dims(image, axis=0)

    return image

def load_dataset(folder):

    image_list = []

    for file in os.listdir(folder):

        if file.lower().endswith((".jpg", ".png",
".jpeg")):

            full_path = os.path.join(folder, file)

            image_list.append(full_path)

    return image_list

dataset_folder = "datasetgambar"

dataset = load_dataset(dataset_folder)

print("Total calibration images:", len(dataset))

dataset = [x for x in dataset if x is not None]

calibration_dataset = nncf.Dataset(dataset,
transform_fn)

quantized_model = nncf.quantize(
    model,
    calibration_dataset
)
```

Setelah proses kuantisasi selesai, model hasil optimasi disimpan kembali dalam format OpenVINO IR menggunakan ov.save_model. Pendekatan ini bertujuan meningkatkan efisiensi inferensi dan mengurangi penggunaan memori, dengan tetap mempertimbangkan potensi penurunan akurasi akibat penurunan presisi.

4.2 Evaluasi Hasil Citra Model Real – ESRGAN

Berdasarkan hasil evaluasi visual, model FP16 secara konsisten menghasilkan kualitas citra yang lebih baik dibandingkan model INT8, terutama dalam mempertahankan detail halus, tekstur, dan ketajaman objek. Pada citra ilustrasi, garis dan bayangan terlihat lebih jelas serta memiliki detail yang lebih baik. Pada citra malam, model FP16 mampu mempertahankan distribusi cahaya dan warna secara lebih stabil, meskipun pada beberapa area terang terlihat sedikit peningkatan kontras. Selain itu, pada citra tanda tangan, garis-garis yang terbentuk terlihat lebih halus dan presisi sehingga struktur detail tetap terjaga dengan baik.



Gambar 4.3 Evaluasi hasil Citra Real - ESRGAN

Sementara itu, model INT8 menunjukkan kualitas citra yang sedikit lebih rendah dibandingkan FP16. Pada beberapa area dengan detail halus, terlihat adanya penurunan ketajaman serta degradasi tekstur yang disebabkan oleh proses kuantisasi. Selain itu, pada beberapa bagian citra juga muncul noise ringan yang tidak ditemukan pada hasil FP16. Meskipun demikian, struktur utama citra tetap dapat dipertahankan dengan baik sehingga hasil yang diperoleh masih layak digunakan untuk berbagai kebutuhan praktis. Pada citra tanda tangan, karakter dan garis masih dapat dikenali dengan jelas meskipun tingkat ketajamannya sedikit lebih rendah dibandingkan FP16.

Secara keseluruhan, evaluasi visual menunjukkan bahwa model FP16 memberikan kualitas citra terbaik, sedangkan model INT8 menawarkan kualitas yang sedikit lebih rendah sebagai konsekuensi dari proses kuantisasi. Namun demikian, penurunan kualitas tersebut masih berada dalam batas yang dapat diterima sehingga INT8 tetap menjadi alternatif yang menarik ketika efisiensi komputasi dan kecepatan inferensi menjadi prioritas utama.

4.3 Evaluasi Hasil Nilai Metrik Inferensi Model Real – ESRGAN

Evaluasi hasil inferensi dilakukan untuk menganalisis performa model Real-ESRGAN berdasarkan aspek kualitas citra dan efisiensi komputasi pada berbagai unit pemrosesan. Pengujian dilakukan menggunakan tiga jenis perangkat pemrosesan, yaitu Central Processing Unit (CPU), Graphics Processing Unit (GPU), dan Neural Processing Unit (NPU), dengan dua konfigurasi precision, yaitu FP16 dan INT8. Kinerja model dievaluasi menggunakan metrik Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM),

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), serta latency inferensi.

Berdasarkan hasil pengujian, terlihat bahwa nilai PSNR, SSIM, MSE, dan RMSE pada seluruh konfigurasi memiliki perbedaan yang relatif kecil. Nilai PSNR berada pada rentang 23,0906 hingga 23,1128 dB, sedangkan nilai SSIM berada pada rentang 0,759175 hingga 0,764524. Hasil ini menunjukkan bahwa penggunaan precision FP16 maupun INT8 tidak memberikan perbedaan yang signifikan terhadap kualitas citra yang dihasilkan. Selain itu, nilai MSE dan RMSE juga menunjukkan pola yang serupa, di mana model dengan precision INT8 menghasilkan nilai error yang sedikit lebih rendah dibandingkan FP16 pada sebagian besar perangkat pemrosesan.

Tabel 4.1 Metrik Evaluasi hasil citra

Processing Unit	Precision	Latency (ms)	Duration (s)	PSNR	SSIM	MSE	RMSE
CPU	FP16	7094.956	7.095	23.1120	0.759846	535229.027	20556.018
CPU	INT8	6702.421	6.702	23.0986	0.764524	525748.277	20440.542
GPU	FP16	1055.861	1.056	23.1060	0.759175	534704.102	20553.698
GPU	INT8	1002.734	1.003	23.1064	0.764113	524735.142	20420.936
NPU	FP16	881.334	0.881	23.1128	0.759865	535164.523	20554.340
NPU	INT8	807.681	0.808	23.0906	0.709497	525840.11	20449.206

Dari sisi efisiensi komputasi, perbedaan yang lebih jelas terlihat pada nilai latency. Pada CPU, model FP16 menghasilkan latency sebesar 7094,956 ms, sedangkan INT8 mampu menurunkannya menjadi 6702,421 ms. Pada GPU, latency berkurang dari 1055,861 ms menjadi 1002,734 ms setelah menggunakan precision INT8. Hasil yang sama juga terlihat pada NPU, di mana latency menurun dari 881,334 ms pada FP16 menjadi 807,681 ms pada INT8. Penurunan latency tersebut menunjukkan bahwa kuantisasi INT8 mampu meningkatkan efisiensi inferensi pada seluruh unit pemrosesan yang diuji.

Berdasarkan perbandingan perangkat pemrosesan, NPU menunjukkan performa inferensi terbaik dengan latency terendah dibandingkan CPU dan GPU pada kedua precision yang digunakan. Pada precision FP16, NPU menghasilkan latency 881,334 ms, lebih cepat dibandingkan GPU sebesar 1055,861 ms dan CPU sebesar 7094,956 ms. Sementara itu, pada precision INT8, NPU kembali menunjukkan latency terendah sebesar 807,681 ms, diikuti GPU sebesar 1002,734 ms dan CPU sebesar 6702,421 ms. Hasil ini menunjukkan bahwa NPU lebih optimal untuk menjalankan model Real-ESRGAN karena dirancang secara khusus untuk menangani operasi kecerdasan buatan.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa penggunaan precision INT8 mampu meningkatkan efisiensi komputasi tanpa menyebabkan penurunan kualitas citra yang signifikan dibandingkan FP16. Selain itu, NPU terbukti menjadi perangkat pemrosesan yang paling efisien dalam menjalankan inferensi model Real-ESRGAN dibandingkan CPU dan GPU. Temuan ini menunjukkan bahwa kombinasi NPU dan precision INT8 merupakan konfigurasi yang paling efektif untuk implementasi image super-resolution yang membutuhkan keseimbangan antara kualitas citra dan kecepatan inferensi.

Tabel 4.2 Gambar dan perbandingan nilai metriks

Processing Unit	PSNR	SSIM	MSE	RMSE	Precision Model
NPU	23.0906	0.709497	525840.11	20449.206	INT8

Real Picture *Result Picture*





4.3.1 Evaluasi PSNR

$$PSNR = 10_{\log 10}\left(\frac{MAX_I^2}{MSE}\right) \tag{13}$$

$$PSNR = 10\log_{10}\left(\frac{65025}{0.320956}\right)$$

$$PSNR = 23.0906\ dB$$

Berdasarkan hasil perhitungan menggunakan metrik Peak Signal-to-Noise Ratio (PSNR), diperoleh nilai sebesar 23,0906 dB. Nilai PSNR digunakan untuk mengukur tingkat kemiripan antara citra hasil rekonstruksi dengan citra referensi berdasarkan perbedaan nilai piksel yang direpresentasikan melalui Mean Squared Error (MSE). Semakin tinggi nilai PSNR, semakin kecil tingkat distorsi yang terjadi pada citra hasil sehingga kualitas rekonstruksi yang dihasilkan semakin baik. Nilai PSNR sebesar 23,0906 dB menunjukkan bahwa model Real-ESRGAN masih mampu mempertahankan sebagian besar informasi visual penting pada citra meskipun terdapat beberapa perbedaan dibandingkan citra referensi. Hasil ini mengindikasikan bahwa proses inferensi yang dilakukan tetap menghasilkan kualitas citra yang baik dengan tingkat kesalahan yang masih dapat diterima.

Dengan demikian, model yang digunakan masih layak diterapkan pada proses image up-scale, terutama pada skenario yang memerlukan keseimbangan antara kualitas visual dan efisiensi komputasi.

4.3.2 Evaluasi SSIM

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (14)$$

μx = 128.7856
μy = 130.5601
σx² = 7431.0016
σy² = 4757.6555
σxy = 5616.3188

Konstanta Dasar
$C_1 = (0.01 \times 255)^2 = 6.5025$
$C_2 = (0.03 \times 255)^2 = 58.5225$

$$SSIM(x,y) = \frac{(2(128.7856)(130.5601) + 6.5025)(2(5616.3188) + 58.5225)}{((128.7856^2 + 130.5601^2 + 6.5025)(7431.0016 + 4757.6555 + 58.5225)}$$

$SSIM \approx 0.7095$

Berdasarkan hasil perhitungan menggunakan metode Structural Similarity Index Measure (SSIM), diperoleh nilai sebesar 0,7095. Nilai tersebut menunjukkan bahwa citra hasil rekonstruksi memiliki tingkat kemiripan struktural yang cukup baik terhadap citra referensi. Perhitungan SSIM dilakukan dengan membandingkan karakteristik luminansi, kontras, dan struktur antara kedua citra. Nilai SSIM yang mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi, sedangkan nilai yang mendekati 0 menunjukkan perbedaan yang semakin besar. Hasil yang diperoleh mengindikasikan bahwa model masih mampu mempertahankan sebagian besar informasi visual penting pada citra, meskipun

terdapat beberapa perbedaan akibat proses inferensi yang menyebabkan penurunan detail dan tekstur pada beberapa area gambar.

4.3.3 Evaluasi MSE

Berdasarkan hasil pengujian, diperoleh nilai Mean Squared Error (MSE) sebesar 525840,11. MSE digunakan untuk mengukur rata-rata kuadrat selisih antara nilai piksel citra hasil rekonstruksi dengan citra referensi. Semakin kecil nilai MSE yang diperoleh, semakin kecil tingkat kesalahan yang terjadi dan semakin baik kualitas citra hasil rekonstruksi. Nilai MSE sebesar 525840,11 menunjukkan masih terdapat perbedaan antara citra hasil dan citra referensi, namun model tetap mampu mempertahankan informasi visual utama pada citra. Nilai ini juga sejalan dengan hasil evaluasi PSNR dan SSIM yang menunjukkan bahwa kualitas citra hasil rekonstruksi masih berada pada tingkat yang dapat diterima. Oleh karena itu, model Real-ESRGAN yang diimplementasikan masih mampu menghasilkan citra dengan kualitas yang baik meskipun telah melalui proses optimasi dan inferensi pada perangkat yang digunakan.

4.3.4 Evaluasi RMSE

Berdasarkan hasil pengujian, diperoleh nilai Root Mean Squared Error (RMSE) sebesar 20449,206. RMSE digunakan untuk mengukur tingkat kesalahan antara citra hasil rekonstruksi dan citra referensi dalam satuan yang sama dengan data asli. Semakin kecil nilai RMSE, semakin kecil perbedaan yang terjadi antara kedua citra sehingga kualitas hasil rekonstruksi menjadi lebih baik. Nilai RMSE sebesar 20449,206 menunjukkan masih terdapat perbedaan antara citra hasil dan citra referensi. Namun demikian, hasil evaluasi menggunakan PSNR dan SSIM menunjukkan bahwa model masih mampu mempertahankan informasi visual utama pada citra. Oleh karena itu, nilai RMSE yang diperoleh perlu dianalisis

bersama metrik evaluasi lainnya untuk memberikan gambaran yang lebih komprehensif mengenai kualitas citra hasil image up-scale.

4.3.5 Hasil Evaluasi PSNR, SSIM, MSE, RMSE

Berdasarkan hasil pengujian yang ditunjukkan pada grafik, terdapat perbedaan performa antara CPU, GPU, dan NPU pada penggunaan precision FP16 dan INT8. Dari aspek efisiensi komputasi, penggunaan precision INT8 secara umum mampu menurunkan latency dibandingkan FP16 pada seluruh perangkat yang diuji. Latency terendah diperoleh pada GPU dengan precision INT8 sebesar 615,32 ms, diikuti oleh NPU dengan precision FP16 sebesar 788,33 ms dan NPU dengan precision INT8 sebesar 925,48 ms. Sementara itu, CPU menunjukkan waktu inferensi paling tinggi pada kedua precision, yaitu 8.087,26 ms untuk FP16 dan 5.441,56 ms untuk INT8. Hasil ini menunjukkan bahwa GPU dan NPU memiliki kemampuan akselerasi yang lebih baik dibandingkan CPU dalam menjalankan model Real-ESRGAN.



Gambar 4.4 *Bar chart* Metriks Hasil

Dari sisi kualitas citra, nilai PSNR tertinggi diperoleh pada GPU dengan precision INT8 sebesar 19,626 dB, sedangkan nilai terendah diperoleh pada CPU dengan precision INT8 sebesar 18,429 dB. Perbedaan nilai PSNR antar konfigurasi relatif kecil, yang menunjukkan bahwa penggunaan precision INT8 tidak menyebabkan penurunan kualitas citra yang signifikan. Hasil serupa juga terlihat pada metrik SSIM, di mana seluruh konfigurasi menghasilkan nilai di atas 0,64. Nilai SSIM tertinggi diperoleh pada GPU dengan precision INT8 sebesar 0,715140, yang menunjukkan tingkat kemiripan struktural terbaik terhadap citra referensi.

Evaluasi menggunakan metrik error menunjukkan bahwa GPU dengan precision INT8 menghasilkan nilai MSE dan RMSE terendah, masing-masing sebesar 71,129 dan 8,432. Nilai error yang lebih rendah menunjukkan bahwa citra hasil rekonstruksi memiliki perbedaan yang lebih kecil terhadap citra referensi. Sebaliknya, CPU dengan precision INT8 menghasilkan nilai MSE dan RMSE tertinggi, yaitu 110,579 dan 10,512, yang mengindikasikan tingkat kesalahan rekonstruksi yang lebih besar dibandingkan konfigurasi lainnya.

Secara keseluruhan, hasil pengujian menunjukkan bahwa penggunaan precision INT8 mampu meningkatkan efisiensi komputasi tanpa mengurangi kualitas citra secara signifikan. Selain itu, GPU dengan precision INT8 memberikan kombinasi terbaik antara kualitas citra dan efisiensi komputasi berdasarkan nilai PSNR, SSIM, MSE, dan RMSE yang diperoleh. Sementara itu, NPU menunjukkan performa yang kompetitif dengan latency yang rendah serta kualitas citra yang relatif mendekati GPU, sehingga berpotensi menjadi solusi yang efektif untuk implementasi model Real-ESRGAN pada perangkat edge dengan keterbatasan sumber daya komputasi.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa implementasi model Real-ESRGAN menggunakan precision FP16 dan INT8 pada CPU, GPU, dan NPU memberikan karakteristik performa yang berbeda baik dari aspek kualitas citra maupun efisiensi komputasi. Hasil pengujian menunjukkan bahwa penggunaan precision INT8 secara umum mampu meningkatkan efisiensi inferensi dengan menurunkan latency dibandingkan FP16 pada seluruh perangkat yang diuji. Latency terendah diperoleh pada GPU dengan precision INT8 sebesar 615,32 ms, sedangkan CPU menghasilkan waktu inferensi tertinggi pada kedua precision.

Dari aspek kualitas citra, perbedaan nilai PSNR dan SSIM antar konfigurasi relatif kecil. GPU dengan precision INT8 menghasilkan nilai PSNR tertinggi sebesar 19,626 dB dan nilai SSIM tertinggi sebesar 0,715140. Selain itu, GPU dengan precision INT8 juga menghasilkan nilai MSE dan RMSE terendah, masing-masing sebesar 71,129 dan 8,432, yang menunjukkan tingkat kesalahan rekonstruksi paling rendah dibandingkan konfigurasi lainnya. Hasil ini menunjukkan bahwa penggunaan precision INT8 tidak menyebabkan penurunan kualitas citra yang signifikan meskipun memberikan peningkatan efisiensi komputasi.

Berdasarkan keseluruhan hasil evaluasi, GPU dengan precision INT8 memberikan kombinasi terbaik antara kualitas citra dan efisiensi komputasi. Sementara itu, NPU menunjukkan performa yang kompetitif dengan latency yang

rendah serta kualitas citra yang mendekati GPU, sehingga berpotensi menjadi alternatif yang efektif untuk implementasi image super-resolution pada perangkat edge. Dengan demikian, penelitian ini membuktikan bahwa pemilihan precision dan perangkat pemrosesan memiliki pengaruh terhadap performa model Real-ESRGAN, serta menunjukkan adanya trade-off antara kualitas citra dan efisiensi komputasi yang perlu dipertimbangkan sesuai kebutuhan aplikasi.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa hal yang dapat menjadi perhatian untuk pengembangan penelitian selanjutnya. Penelitian ini masih menggunakan jumlah dan variasi data uji yang terbatas, sehingga penelitian berikutnya dapat menggunakan dataset yang lebih beragam dari segi resolusi, jenis objek, serta tingkat degradasi citra untuk memperoleh hasil evaluasi yang lebih representatif. Selain itu, penelitian selanjutnya dapat melakukan perbandingan dengan model super-resolution lainnya, seperti ESRGAN, EDSR, SRCNN, atau A-ESRGAN, sehingga diperoleh gambaran yang lebih komprehensif mengenai performa masing-masing model. Penelitian ini juga berfokus pada evaluasi kualitas citra menggunakan metrik PSNR, SSIM, MSE, dan RMSE serta efisiensi komputasi berdasarkan latency inferensi. Oleh karena itu, penelitian selanjutnya dapat menambahkan parameter evaluasi lain, seperti konsumsi daya, penggunaan memori, dan throughput untuk memberikan analisis yang lebih menyeluruh terhadap performa model pada berbagai perangkat pemrosesan. Selain itu, evaluasi kualitas citra dapat dilengkapi dengan pengujian berbasis persepsi manusia sehingga hasil yang diperoleh tidak hanya bergantung pada metrik numerik, tetapi juga mempertimbangkan kualitas visual yang

dirasakan oleh pengguna. Penelitian selanjutnya dapat mengeksplorasi teknik optimasi model lainnya, seperti pruning, quantization-aware training, maupun pendekatan optimasi yang lebih modern untuk meningkatkan efisiensi komputasi tanpa mengurangi kualitas citra secara signifikan. Implementasi model juga dapat dikembangkan pada berbagai perangkat edge computing dengan spesifikasi perangkat keras yang berbeda untuk mengetahui pengaruh arsitektur perangkat terhadap performa inferensi. Dengan demikian, hasil penelitian yang diperoleh diharapkan dapat memberikan kontribusi yang lebih luas terhadap pengembangan teknologi image super-resolution berbasis deep learning pada lingkungan komputasi modern.

Daftar pustaka

Agarwal, V., Lohani, M. C., Singh Bist, A., Rahardja, L., Hardini, M., & Mustika, G. (2022). Deep CNN-Real ESRGAN: An Innovative Framework for Lung Disease Prediction. *2022 IEEE Creative Communication and Innovative Technology, ICCIT 2022*. https://doi.org/10.1109/ICCIT55355.2022.10118586

Agirbas, A., Sabyrova, R., & Kirhalli, T. E. (2025). Landscape Design asArt: An Experimental Methodology with the Use of Neurographic Art and Metaball Geometry. *International Journal of Built Environment and Sustainability*, *12*(1), 1–8. https://doi.org/10.11113/IJBES.V12.N1.1214

Andriadi, K., Zarlis, M., Heryadi, Y., & Chowanda, A. (2024). Evaluation of REAL-ESRGAN Using Different Types of Image Degradation. *11th International Conference on ICT for Smart Society: Integrating Data and Artificial Intelligence for a Resilient and Sustainable Future Living, ICISS 2024 - Proceeding*. https://doi.org/10.1109/ICISS62896.2024.10751549

Bakurov, I., Buzzelli, M., Schettini, R., Castelli, M., & Vanneschi, L. (2022). Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Systems with Applications*, *189*, 116087. https://doi.org/10.1016/J.ESWA.2021.116087

Batarius, P., Sinlae, A. A., & Fahik, E. F. (2022). Analysis of the Quality of Natural Dyes in Weaving Exposed to Sunlight Using MSE and PSNR Parameters. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(5), 797–802. https://doi.org/10.29207/RESTI.V6I5.4339

Bernabe, S., Gonzalez, C., Fernandez, A., & Bhangale, U. (2021). Portability and Acceleration of Deep Learning Inferences to Detect Rapid Earthquake Damage From VHR Remote Sensing Images Using Intel OpenVINO Toolkit. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 6906–6915. https://doi.org/10.1109/JSTARS.2021.3075961

Chaves-González, E., & León-Vega, L. G. (2022). Benchmarking the NXP i.MX8M+ neural processing unit: smart parking case study. *Tecnología En Marcha*, *35*(9). https://doi.org/10.18845/TM.V35I9.6487

Chen, G., Chen, J., Liao, H., Liu, K., Tang, X., & Kong, Y. (2025). Super-resolution reconstruction of off-axis digital holography based on Real-ESRGAN. *Applied Optics, Vol. 64, Issue 9, Pp. 2245-2254*, *64*(9), 2245–2254. https://doi.org/10.1364/AO.545914

Chen, Y., & Lu, L. (2026). AscQLUT: A decode-fused INT4 GEMM Kernel for Ascend NPUs. *Journal of Systems Architecture*, 103839. https://doi.org/10.1016/J.SYSARC.2026.103839

Chung, J., Kim, H. M., Shin, K., Lyuh, C. G., Cho, Y. C. P., Han, J., Kwon, Y., Gong, Y. H., & Chung, S. W. (2022). A layer-wise frequency scaling for a neural processing unit. *ETRI Journal*, *44*(5), 849–858. https://doi.org/10.4218/ETRIJ.2022-0094

Enhance AI Upscaling with Intel AI Boost NPU. (n.d.). Retrieved June 2, 2026, from https://www.intel.com/content/www/us/en/developer/articles/technical/enhance-ai-upscaling-with-intel-ai-boost-npu.html

Fraga, L. G. de la. (2021). Differential Evolution under Fixed Point Arithmetic and FP16 Numbers. *Mathematical and Computational Applications*, *26*(1), 13. https://doi.org/10.3390/MCA26010013

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2023). A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191

Jeon, W., Lee, J., Kang, D., Kal, H., & Ro, W. W. (2021). PIMCaffe: Functional Evaluation of a Machine Learning Framework for In-Memory Neural Processing Unit. *IEEE Access*, *9*, 96629–96640. https://doi.org/10.1109/ACCESS.2021.3094043

Kim, S., Park, G., & Yi, Y. (2021). Performance Evaluation of INT8 Quantized Inference on Mobile GPUs. *IEEE Access*, *9*, 164245–164255. https://doi.org/10.1109/ACCESS.2021.3133100

Kononchuk, R., Cai, J., Ellis, F., Thevamaran, R., & Kottos, T. (2022). Exceptional-point-based accelerometers with enhanced signal-to-noise ratio. *Nature 2022 607:7920*, *607*(7920), 697–702. https://doi.org/10.1038/s41586-022-04904-w

Liu, J., Wang, Q., Zhang, D., & Shen, L. (2021). Super-Resolution Model Quantized in Multi-Precision. *Electronics 2021, Vol. 10, Page 2176*, *10*(17), 2176. https://doi.org/10.3390/ELECTRONICS10172176

Liu, K., & Cai, Y. (2025). A deep convolutional generative adversarial network for data imputation: Application to wind speed time series. *Advances in Wind Engineering*, *2*(2), 100054. https://doi.org/10.1016/j.awe.2025.100054

Ma, S., Liu, C., Li, Z., & Yang, W. (2022). Integrating Adversarial Generative Network with Variational Autoencoders towards Cross-Modal Alignment for Zero-Shot Remote Sensing Image Scene Classification. *Remote Sensing*, *14*(18), 4533. https://doi.org/10.3390/RS14184533

Maza, C. F., & Martínez, S. G. (2025). Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video.

Revista Tecnológica, *37*(E1), 56–76. https://doi.org/10.37815/RTE.V37NE1.1317

Mudeng, V., Kim, M., & Choe, S. W. (2022). Prospects of Structural Similarity Index for Medical Image Analysis. *Applied Sciences 2022, Vol. 12, Page 3754*, *12*(8), 3754. https://doi.org/10.3390/APP12083754

Muppalla, S. V., Deekonda, L., Deveragatla, S., & Harini, V. (2026). ESRGAN-Assisted Sequential Transfer Learning for Disaster-Oriented Object Detection in Remote Sensing Images. *2026 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)*, 1–6. https://doi.org/10.1109/ICISSGT67741.2026.11483865

N, S., T, R., & DS, M. (2024). Bunionectomy as an Acute Postoperative Pain Model: Overview of Common Experimental Methods, and Insights from Past Clinical Trials. *Journal of Pain Research*, *Volume 17*, 4399–4420. https://doaj.org/article/00a067bd2c4e4ab1981e4c37bb244704

Nandal, P., Pahal, S., Khanna, A., & Rogerio Pinheiro, P. (2024). Super-Resolution of Medical Images Using Real ESRGAN. *IEEE Access*, *12*, 176155–176170. https://doi.org/10.1109/ACCESS.2024.3497002

OpenVINO 2026.2 — OpenVINO™ documentation — Version(2026). (n.d.). Retrieved June 2, 2026, from https://docs.openvino.ai/2026/index.html

Praveen, B. C., & Sivasankaran, K. (2026a). Fast and Accurate Sentiment Analysis Using a Hybrid LLM Model With OpenVINO Optimization. *IEEE Access*, *14*, 16176–16187. https://doi.org/10.1109/ACCESS.2026.3658411

Praveen, B. C., & Sivasankaran, K. (2026b). Fast and Accurate Sentiment Analysis Using a Hybrid LLM Model With OpenVINO Optimization. *IEEE Access*, *14*, 16176–16187. https://doi.org/10.1109/ACCESS.2026.3658411

Raymond, C., Zhang, D., Cabello, J., Liu, L., Moyaert, P., Burneo, J. G., Dada, M. O., Hicks, J. W., Finger, E., Soddu, A., Andrade, A., Jurkiewicz, M. T., & Anazodo, U. C. (2024). SMART-PET: a Self-SiMilARiTy-aware generative adversarial framework for reconstructing low-count [18F]-FDG-PET brain imaging. *Frontiers in Nuclear Medicine*, *4*, 1469490. https://doi.org/10.3389/FNUME.2024.1469490/TEXT

Senthil, Mr. A., & Tomlinson, Mr. L. (2025). Enhancing super-resolution models: a comparative analysis of real-ESRGAN, AESRGAN, and ESRGAN. *Https://Doi.Org/10.1117/12.3062693*, *13636*, 116. https://doi.org/10.1117/12.3062693

Sivakoti, K. (2025). *GANCompress: GAN-Enhanced Neural Image Compression with Binary Spherical Quantization*. https://arxiv.org/pdf/2505.13542

Tan, T., & Cao, G. (2021). *Efficient Execution of Deep Neural Networks on Mobile Devices with NPU*. 283–298. https://doi.org/10.1145/3412382.3458272

Totad, K. S., Hanchinal, A. R., Shanbhog, N. R., Patgar, T. V., & Dhulavvagol, P. M. (2025). Quantization Techniques for Optimizing MobileNetV3Large in Yoga Pose Recognition on Edge Devices. *Procedia Computer Science*, *260*, 1000–1008. https://doi.org/10.1016/J.PROCS.2025.03.284

Usama, Z., Alavi, A., & Chan, J. (2025). A Review on the Applications of GANs for 3D Medical Image Analysis. *Applied Sciences (Switzerland)*, *15*(20), 11219. https://doi.org/10.3390/APP152011219/S1

Vivekananthan, S. (2024). *Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAEs, GANs, and Stable Diffusion*. http://arxiv.org/abs/2408.08751

Wang, J., & Zhao, W. (2025). Automatic cutting and suturing control system based on improved FP16 visual recognition algorithm. *International Journal of Advanced Robotic Systems*, *22*(4). https://doi.org/10.1177/17298806251348118

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. *Proceedings of the IEEE International Conference on Computer Vision*, *2021-October*, 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217

Wang, Z., She, Q., & Ward, T. E. (2022). Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Computing Surveys*, *54*(2), 37. https://doi.org/10.1145/3439723;CSUBTYPE:STRING:JOURNAL;SERIALTOPIC:TOPIC:ACM-PUBTYPE

Wei, Z., Huang, Y., Chen, Y., Zheng, C., & Gao, J. (2024). A-ESRGAN: Training Real-World Blind Super-Resolution with Attention U-Net Discriminators. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14327 LNAI*, 16–27. https://doi.org/10.1007/978-981-99-7025-4_2/SAVE-RESEARCH

Wenzel, M. (2023). Generative Adversarial Networks and Other Generative Models. *Neuromethods*, *197*, 139–192. https://doi.org/10.1007/978-1-0716-3195-9_5/FIGURES/27

Xue, W., Xu, B., Yang, K., Liu, Y., Fan, D., Xu, P., & Tian, Y. (2026). *SGEMM-cube: Precision-Recovery FP32 GEMM Approximation on Ascend NPUs with FP16 Matrix Engines*. http://arxiv.org/abs/2507.23387

Zhao, Z., Gao, Y., Min, B., Miao, X., Hu, J., Liu, Y., & Pharksuwan, K. (2025). PConvSRGAN: Real-world super-resolution reconstruction with pure

convolutional networks. *Computer Vision and Image Understanding*, *260*, 104465. https://doi.org/10.1016/J.CVIU.2025.104465

Zheng, X., Li, Y., Chu, H., Feng, Y., Ma, X., Wang, Z., Luo, J., Guo, J., Qin, H., Magno, M., & Liu, X. (2026). An empirical study of Qwen3 quantization. *Visual Intelligence 2026 4:1*, *4*(1), 11-. https://doi.org/10.1007/S44267-026-00114-4

Zhu, L., Zhong, B., & Ma, K. K. (2025). APSNR: Artifact Peak Signal-to-Noise Ratio for Image Quality Assessment. *IEEE Transactions on Image Processing*, *34*, 7180–7192. https://doi.org/10.1109/TIP.2025.3623864

Zhu, M., Zhang, J., Bu, L., Nie, S., Bai, Y., Zhao, Y., & Mei, N. (2024). Methodology and Experimental Verification for Predicting the Remaining Useful Life of Milling Cutters Based on Hybrid CNN-LSTM-Attention-PSA. *Machines*, *12*(11), 752. https://doi.org/10.3390/MACHINES12110752

Zhu, Z., Lei, Y., Qin, Y., Zhu, C., & Zhu, Y. (2023). IRE: Improved Image Super-Resolution Based on Real-ESRGAN. *IEEE Access*, *11*, 45334–45348. https://doi.org/10.1109/ACCESS.2023.3256086

Zongjin, J., Rui, Z., Caihang, L., & Yunsheng, Z. (2025). Experimental study of hollow fiber membrane cooler based on response surface methodology. *Xi'an Gongcheng Daxue Xuebao*, *39*(1), 52–61. https://doi.org/10.13338/J.ISSN.1674-649X.2025.01.007