DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS ATTENTION MECHANISM DAN DEEPSORT

TESIS

Oleh :

TEGUH BUDI SANTOSO

NIM 240605220002

PROGRAM STUDI MAGISTER INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM

MALANG

2026

DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS ATTENTION MECHANISM DAN DEEPSORT

TESIS

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh :
TEGUH BUDI SANTOSO
NIM 240605220002

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS ATTENTION MECHANISM DAN DEEPSORT

TESIS

Oleh:
TEGUH BUDI SANTOSO
NIM 240605220002

Telah Diperiksa dan Disetujui untuk di uji:
Tanggal 11 Mei 2026

Pembimbing I,	Pembimbing II,
Dr. Ir. Fachrul Kurniawan ST., M.MT ., IPU	Dr. M. Imamudin Lc, MA
NIP. 19771020200912 1 001	NIP. 19740602 200901 1 010

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Ir. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS ATTENTION MECHANISM DAN DEEPSORT

TESIS

Oleh:
TEGUH BUDI SANTOSO
NIM 240605220002

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal : 11 Mei 2026

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Dr. Ir. Fresy Nugroho, ST., MT, IPM, ASEAN Eng NIP. 19710722 201101 1 001	( )
Penguji II	:	Dr. Ir. Yunifa Miftachul Arif, M. T NIP. 19830316 201101 1 004	( )
Pembimbing I	:	Dr. Ir. Fachrul Kurniawan ST., M.MT ., IPU NIP. 19771020200912 1 001	( )
Pembimbing II	:	Dr. M. Imamudin Lc, MA NIP. 19740602 200901 1 010	

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Ir. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	:	Teguh Budi Santoso
NIM	:	240605220002
Program Studi	:	Magister Informatika
Fakultas	:	Sains dan Teknologi
Judul Tesis	:	Deteksi Dan Pelacakan Wajah Pada Lingkungan Keramaian Menggunakan YOLOv9 Berbasis *Attention Mechanism* Dan *Deepsort*

Menyatakan dengan sebenarnya bahwa Tesis yang saya tulisan ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 11 Mei 2026

Yang membuat pernyataan,

Teguh Budi Santoso

NIM. 240605220002

MOTTO

“Kerjakan yang bisa dikerjakan hari ini, dengan sepenuh hati”

KATA PENGANTAR

Puji syukur kepada Allah SWT yang telah melimpahkan karunia, rahmat, dan hidayah-Nya sehingga penulis dapat menyelesaikan laporan Tesis yang berjudul "**DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS *ATTENTION MECHANISM* DAN *DEEPSORT***" Laporan Tesis ini disusun untuk memenuhi sebagian persyaratan memperoleh gelar Magister Komputer di Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penulis menyadari bahwa selama menyelesaikan tesis ini, banyak sekali bantuan, dukungan, dan saran dari semua pihak, baik secara langsung maupun tidak langsung, selama proses penyusunan hingga selesai. Oleh karena itu, penulis ingin mengucapkan terima kasih dan hormat yang sebesar-besarnya kepada:

1. **Bapak Dr. Ir. Fachrul Kurniawan ST., M.MT ., IPU.** selaku Dosen Pembimbing I dan **Bapak Dr. M. Imamudin Lc, MA.** selaku Dosen Pembimbing II atas motivasi, arahan, serta pengalaman berharga yang diberikan kepada penulis.
2. Seluruh dosen Program Studi Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah dengan tiada lelah memberikan ilmunya kepada penulis.
3. **Kedua orang tua** serta keluarga besar yang telah memberikan dukungan, perhatian, nasihat, motivasi, dan doa yang tiada henti kepada penulis untuk menyelesaikan tesis ini.

4. Seluruh rekan mahasiswa angkatan ke-11 Program Studi Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan banyak pengalaman, saran, dan kritik yang membangun pada tesis ini.
5. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah membantu dalam menyelesaikan tesis ini.

Harapan dan doa penulis, semoga semua kebaikan dan jasa dari pihak-pihak yang telah membantu dalam penyelesaian tesis ini diterima oleh Allah SWT, serta mendapat balasan yang lebih baik dan berlipat ganda. Penulis menyadari bahwa tesis ini masih jauh dari sempurna, baik dari segi materi maupun penyajian. Oleh karena itu, saran dan kritik yang membangun sangat diharapkan demi penyempurnaan tesis ini. Penulis berharap tesis ini dapat memberikan manfaat dan menambah wawasan bagi semua pihak.

Malang, 11 Mei 2026
Penulis

Daftar Isi

Daftar Gambar

Daftar Tabel

ABSTRAK

Teguh Budi Santoso, 2026, **DETEKSI DAN PELACAKAN WAJAH PADA LINGKUNGAN KERAMAIAN MENGGUNAKAN YOLOV9 BERBASIS *ATTENTION MECHANISM* DAN *DEEPSORT*.** Tesis, Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. Fachrul Kurniawan ST., M.MT ., IPU (II) Dr. Mochamad Imamudin Lc, MA

Kata Kunci: Deteksi wajah, Pelacakan wajah, YOLOv9, *Attention mechanism*, *DeepSORT*, Keramaian padat, Multi-object tracking

Penelitian ini bertujuan mengembangkan dan mengevaluasi sistem deteksi dan pelacakan wajah pada lingkungan keramaian padat menggunakan integrasi teknik pra-pemrosesan citra, YOLOv9 dengan *Attention Mechanism*, *MobileFaceNet*, dan *DeepSORT*. Lingkungan keramaian menghadirkan tantangan berupa *occlusion*, pergerakan dinamis, perubahan pencahayaan, variasi skala dan sudut pandang wajah yang dapat menurunkan akurasi deteksi serta kestabilan identitas selama pelacakan. Tahap awal penelitian menerapkan pra-pemrosesan citra melalui normalisasi, penyesuaian resolusi, dan augmentasi untuk meningkatkan kualitas data masukan, yang kemudian dievaluasi menggunakan metrik *Detection Score* (DET), *Peak Signal-to-Noise Ratio* (PSNR), dan *Structural Similarity Index Measure* (SSIM). Model YOLOv9 diperkuat dengan *Attention Mechanism* untuk meningkatkan kemampuan ekstraksi fitur wajah pada kondisi kepadatan tinggi dan *occlusion*, sedangkan *MobileFaceNet* dan *DeepSORT* digunakan untuk mempertahankan konsistensi identitas antar-*frame*.

Eksperimen dilakukan melalui variasi rasio data *training*-*validation* (90:10, 50:50, dan 10:90) serta variasi parameter pelatihan berupa *epoch*, *batch size*, dan *learning rate* untuk menganalisis stabilitas dan kemampuan generalisasi model. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, mAP@0.5, dan *validation loss* pada tahap deteksi, serta MOTA, IDF1, *Identity Switches*, dan FPS pada tahap pelacakan. Hasil penelitian menunjukkan bahwa konfigurasi terbaik diperoleh pada rasio 90:10 dengan *epoch* 200, *batch size* 16, dan *learning rate* 0.0005, yang menghasilkan *precision* sebesar 97.20%, *recall* sebesar 97.04%, *F1-score* sebesar 97.12%, mAP@0.5 sebesar 97.33%, MOTA sebesar 83.9%, dan IDF1 sebesar 81.9%. Hasil tersebut menunjukkan bahwa integrasi *Attention Mechanism* pada YOLOv9 mampu meningkatkan kualitas deteksi wajah secara signifikan, sementara *MobileFaceNet* dan *DeepSORT* mampu mempertahankan konsistensi identitas objek secara efektif pada lingkungan keramaian padat.

ABSTRACT

Teguh Budi Santoso, 2026, **Face Detection and Tracking in Crowded Environments Using YOLOv9 Based *Attention Mechanism* and *DeepSORT***. Thesis, Master's Program in Informatics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisors: (I) Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU; (II) Dr. Mochamad Imamudin, Lc., M.A.

Keywords: Face detection, Face tracking, YOLOv9, *Attention mechanism*, *DeepSORT*, Dense crowds, Multi-object tracking.

This study aims to develop and evaluate a face detection and tracking system in dense crowd environments by integrating image preprocessing techniques, YOLOv9 with an Attention Mechanism, MobileFaceNet, and DeepSORT. Crowded environments present several challenges, including occlusion, dynamic movements, illumination variations, and differences in facial scale and viewing angles, which can reduce detection accuracy and identity consistency during tracking. The preprocessing stage applies image normalization, resolution adjustment, and data augmentation to improve input image quality, which is subsequently evaluated using Detection Score (DET), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). YOLOv9 is enhanced with an Attention Mechanism to improve facial feature extraction under high-density and occlusion conditions, while MobileFaceNet and DeepSORT are employed to maintain identity consistency across video frames.

Experiments were conducted using variations of training-validation data ratios (90:10, 50:50, and 10:90) as well as training parameters, including epoch, batch size, and learning rate, to analyze model stability and generalization capability. Performance evaluation was carried out using precision, recall, F1-score, mAP@0.5, and validation loss for the detection stage, as well as MOTA, IDF1, Identity Switches, and FPS for the tracking stage. The results show that the best configuration was achieved using a 90:10 ratio with 200 epochs, a batch size of 16, and a learning rate of 0.0005, resulting in a precision of 97.20%, recall of 97.04%, F1-score of 97.12%, mAP@0.5 of 97.33%, MOTA of 83.9%, and IDF1 of 81.9%. These findings demonstrate that the integration of an Attention Mechanism into YOLOv9 significantly improves face detection performance, while MobileFaceNet and DeepSORT effectively maintain object identity consistency in dense crowd environments.

الملخص

تيغوه بودي سانتوسو، 2026، **كشف الوجوه وتتبعها في البيئات المزدحمة باستخدام YOLOv9 المستند إلى آلية الانتباه وDeepSORT**. رسالة ماجستير، برنامج الماجستير في المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرفان: (الأول) الدكتور المهندس فخر الكرنياوان، S.T., M.MT., IPU؛ (الثاني) الدكتور محمد إمام الدين، ليسانس، ماجستير.

الكلمات المفتاحية: كشف الوجوه، تتبع الوجوه، YOLOv9، آلية الانتباه، DeepSORT، الحشود الكثيفة، تتبع الأجسام المتعددة.

تهدف هذه الدراسة إلى تطوير نظام كشف الوجوه وتتبعها في بيئات الحشود الكثيفة وتقييمه، من خلال دمج تقنيات المعالجة المسبقة للصور، ونموذج YOLOv9 المعزَّز بآلية الانتباه، وشبكة MobileFaceNet، وخوارزمية DeepSORT. تُفرز البيئات المزدحمة تحديات عدة، منها: الإخفاء، والحركة الديناميكية، وتباين الإضاءة، فضلاً عن الاختلافات في حجم الوجه وزوايا الرؤية، مما قد يُقلل من دقة الكشف واتساق الهوية أثناء التتبع. تُطبّق مرحلة المعالجة المسبقة تطبيعَ الصور وضبطَ الدقة وتوسيعَ البيانات، بهدف تحسين جودة صور المدخلات، ويُقيَّم ذلك لاحقاً باستخدام درجة الكشف (DET)، ونسبة الإشارة إلى الضوضاء القمية (PSNR)، ومؤشر التشابه الهيكلي (SSIM). يُعزَّز نموذج YOLOv9 بآلية الانتباه لتحسين استخلاص السمات الوجهية في ظروف الكثافة العالية والإخفاء، فيما تُوظَّف شبكة MobileFaceNet وخوارزمية DeepSORT للحفاظ على اتساق الهوية عبر إطارات الفيديو.

أُجريت التجارب باستخدام تباينات في نسب تقسيم بيانات التدريب والتحقق (90:10، و50:50، و10:90)، ومعاملات التدريب، شاملةً عدد الحقب وحجم الدُّفعة ومعدل التعلم، وذلك لتحليل استقرار النموذج وقدرته على التعميم. جرى تقييم الأداء باستخدام مقاييس الدقة والاستدعاء ودرجة F1 ومتوسط الدقة المتوسطة mAP@0.5 وخسارة التحقق في مرحلة الكشف، إلى جانب مقاييس MOTA و IDF1 وعدد تبديلات الهوية وإطارات الثانية (FPS) في مرحلة التتبع. تُظهر النتائج أن أفضل إعداد تحقَّق باستخدام نسبة 90:10 مع 200 حقبة وحجم دُفعة 16 ومعدل تعلم 0.0005، إذ بلغت الدقة 97.20%، والاستدعاء 97.04%، ودرجة F1 نسبة 97.12%، ومتوسط الدقة المتوسطة mAP@0.5 نسبة 97.33%، فضلاً عن بلوغ MOTA نسبة 83.9% و IDF1 نسبة 81.9%. تُثبت هذه النتائج أن دمج آلية الانتباه في نموذج YOLOv9 يُحسِّن أداء كشف الوجوه تحسيناً ملحوظاً، في حين تُحافظ شبكة MobileFaceNet وخوارزمية DeepSORT بفاعلية على اتساق هوية الأجسام في بيئات الحشود الكثيفة.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penerapan teknologi kecerdasan buatan dalam bidang pengolahan gambar digital mengalami perkembangan yang sangat pesat. Sistem yang mampu memahami objek visual melalui kamera kini menjadi pusat perhatian dalam penelitian komputer modern. Salah satu bidang yang berkembang sangat cepat adalah pengenalan wajah (*face recognition*) yang berperan penting dalam berbagai kebutuhan keamanan, pemantauan publik, dan analisis perilaku manusia di ruang terbuka. Kemampuan sistem untuk mengenali wajah manusia dalam skala besar menjadi tantangan tersendiri karena kondisi visual di dunia nyata terus berubah mengikuti dinamika lingkungan (Alharbey et al., 2022). Wajah dapat muncul dari berbagai arah, tertutup bagian tubuh atau pakaian, serta terpengaruh variasi cahaya sepanjang waktu, sehingga sistem yang dikembangkan dituntut memiliki ketahanan dan akurasi yang tinggi dalam kondisi yang tidak terkontrol.

Lingkungan ramai semakin meningkatkan kompleksitas proses deteksi dan pengenalan wajah. Pada situasi padat, wajah sering berada sangat dekat satu sama lain sehingga saling menutupi secara geometris. Kondisi *occlusion* ini menyulitkan model dalam mengekstraksi fitur visual secara utuh, terlebih ketika kamera berada pada jarak jauh yang menyebabkan ukuran wajah relatif kecil dan detail visual berkurang. Penelitian pada pengawasan jamaah Haji menunjukkan bahwa deteksi wajah pada kerumunan besar memerlukan pendekatan yang mampu menahan

gangguan visual dan menjaga stabilitas pemrosesan dalam durasi yang panjang (Alharbey et al., 2022). Sejumlah studi menyimpulkan bahwa keramaian merupakan salah satu kondisi paling menantang dalam pengenalan wajah karena sistem harus mengelola informasi visual dalam jumlah besar secara simultan.

Tantangan tersebut semakin meningkat ketika kualitas gambar menurun akibat variasi pencahayaan, *noise*, maupun penggunaan masker yang menutupi sebagian fitur wajah. Kondisi ini ditegaskan oleh Kurniawan et al. (2023) yang menunjukkan bahwa kualitas *input* gambar sangat memengaruhi keberhasilan model deteksi berbasis YOLO dalam mengenali wajah dan atribut pelindung. Perkembangan metode *deep learning* memberikan kontribusi signifikan dalam menjawab tantangan ini. Model *Convolutional Neural Networks* (CNN) mampu menghasilkan representasi fitur yang kaya dan adaptif terhadap variasi pose, pencahayaan, serta kondisi lingkungan. CNN menjadi fondasi utama berbagai model modern seperti *RetinaFace* dan *MobileFaceNet* yang banyak digunakan pada sistem pengenalan wajah berbasis video (Almuqren et al., 2023).

Model deteksi satu tahap seperti *You Only Look Once* (YOLO) memberikan keunggulan penting bagi sistem pengenalan wajah di keramaian karena kemampuannya memproses gambar secara cepat dan efisien. YOLOv9 sebagai versi terbaru menghadirkan penyempurnaan pada komponen backbone, neck, dan mekanisme penggabungan fitur, sehingga lebih efektif dalam mendeteksi objek berukuran kecil seperti wajah pada kamera pengawas (Narayanan et al., 2025). Kinerja ini semakin diperkuat melalui penerapan yang membantu model memusatkan fokus pada area wajah yang relevan di tengah latar belakang

kompleks. Pendekatan ini meningkatkan sensitivitas deteksi dan menjaga konsistensi hasil pada kondisi keramaian tinggi.

Setelah wajah terdeteksi, keberlanjutan identitas pada setiap *frame* menjadi aspek krusial. Algoritma pelacakan seperti mampu menjaga konsistensi identitas dengan mengombinasikan informasi gerak melalui *Kalman Filter* dan fitur visual berbasis CNN (Djarah et al., 2024). Pendekatan ini memungkinkan sistem tetap stabil meskipun terjadi perubahan pose, pencahayaan, atau gangguan sementara akibat *occlusion*. Stabilitas pelacakan sangat penting bagi sistem pengawasan yang bertujuan mengamati aktivitas individu secara berkelanjutan dalam lingkungan publik yang padat.

Urgensi pengembangan sistem yang akurat dan stabil ini juga dapat ditinjau dari perspektif nilai normatif Islam. Surah Ar-Rahman ayat 7–9 menegaskan prinsip keseimbangan dan keadilan sebagai fondasi keteraturan kehidupan:

وَالسَّمَآءَ رَفَعَهَا وَوَضَعَ الْمِيْزَانَۙ ۝٧ اَلَّا تَطْغَوْا فِى الْمِيْزَانِ ۝٨ وَاَقِيْمُوا الْوَزْنَ بِالْقِسْطِ وَلَا تُخْسِرُوا الْمِيْزَانَ ۝٩

Artinya: "*Langit telah Dia tinggikan dan Dia telah menciptakan timbangan (keadilan dan keseimbangan), agar kamu tidak melampaui batas dalam timbangan itu. Tegakkanlah timbangan itu dengan adil dan janganlah kamu mengurangi timbangan itu.*"

Dalam Tafsir Al-Tabari, konsep *al-mizan* dijelaskan sebagai prinsip keadilan yang bersifat menyeluruh, tidak terbatas pada timbangan fisik semata, tetapi mencakup seluruh bentuk pengukuran, penilaian, dan ketetapan yang digunakan manusia dalam menjalankan urusan kehidupan. Al-Tabari menegaskan bahwa perintah *"janganlah kamu melampaui batas dalam timbangan"* bermakna larangan terhadap segala bentuk ketidakadilan, penyimpangan, dan kekeliruan dalam menetapkan

ukuran dan keputusan. Sementara perintah *"tegakkanlah timbangan dengan adil"* mengandung kewajiban untuk memastikan ketepatan, kejujuran, dan proporsionalitas dalam setiap sistem yang digunakan manusia (Al-Tabari, 2007).

Dalam konteks teknologi pengenalan wajah, prinsip *al-mizan* sebagaimana dijelaskan oleh Al-Ṭabari tercermin pada tuntutan akan sistem yang presisi, terukur, dan bertanggung jawab (Al-Tabari, 2007). Ketepatan model dalam mendeteksi dan melacak wajah menjadi bentuk implementasi keadilan teknis, karena kesalahan pengukuran atau keputusan sistem dapat menimbulkan dampak nyata terhadap keselamatan, hak, dan kehormatan individu di ruang publik. Dengan demikian, akurasi dan stabilitas sistem bukan hanya persoalan teknis, tetapi juga cerminan penerapan nilai keseimbangan dan keadilan dalam pemanfaatan teknologi.

Urgensi pengembangan sistem pengenalan wajah pada lingkungan keramaian juga dapat dijelaskan melalui perspektif *maqaṣid al-syari'ah*. Dalam teori *maqasid* klasik yang dirumuskan oleh Al-Ghazali, tujuan utama syariat adalah menjaga lima kebutuhan dasar manusia, yaitu agama (*hifz al-din*), jiwa (*hifz al-nafs*), akal (*hifz al-'aql*), keturunan (*hifz al-nasl*), dan harta (*hifz al-mal*) (Al-Ghazali, 2008). Prinsip-prinsip tersebut menunjukkan bahwa pemanfaatan teknologi harus diarahkan untuk menghasilkan kemaslahatan serta mencegah potensi bahaya sosial (Kamali, 2008).

Dalam penelitian ini, prinsip *hifz al-din* tercermin melalui dukungan teknologi terhadap keamanan dan keteraturan pelaksanaan aktivitas keagamaan pada ruang publik yang padat, seperti ibadah berjamaah, Haji, maupun Umrah, sehingga masyarakat dapat menjalankan ibadah secara lebih aman dan tertib.

Prinsip *hifz al-nafs* diwujudkan melalui kemampuan sistem dalam membantu deteksi dini, pemantauan berkelanjutan, dan identifikasi individu secara real-time guna meminimalkan risiko kehilangan orang, kepanikan massal, maupun insiden keselamatan pada lingkungan keramaian (Auda, 2008).

Selanjutnya, *hifz al-'aql* tercermin pada pemanfaatan ilmu pengetahuan dan kecerdasan buatan secara rasional untuk mendukung pengambilan keputusan berbasis data dalam sistem keamanan publik. Adapun *hifz al-nasl* berkaitan dengan kemampuan sistem dalam membantu pelacakan individu rentan, seperti anak-anak atau anggota keluarga yang terpisah di tengah keramaian, sehingga mendukung perlindungan keluarga dan keselamatan kelompok sosial. Sementara itu, *hifz al-mal* diwujudkan melalui peningkatan keamanan lingkungan publik yang dapat membantu mengurangi tindak kriminal, kehilangan barang, maupun risiko kerugian material akibat kondisi keramaian yang tidak terkendali.

Selain perlindungan terhadap jiwa dan aspek sosial, pengembangan sistem pengenalan wajah juga berkaitan erat dengan prinsip *hifz al-'ird* (menjaga kehormatan manusia). Kehormatan manusia mencakup perlindungan identitas, martabat, dan hak individu agar tidak mengalami perlakuan tidak adil akibat kesalahan sistem. Dalam lingkungan keramaian, kesalahan deteksi atau pelacakan dapat menimbulkan salah identifikasi, stigma sosial, maupun tuduhan yang merugikan individu tertentu. Oleh karena itu, pengembangan model yang presisi, stabil, dan berimbang menjadi keharusan etis dan teknis. Sistem yang dirancang dengan akurasi tinggi mencerminkan upaya menjaga kehormatan manusia melalui penerapan prinsip keadilan, proporsionalitas, dan tanggung jawab dalam teknologi,

sehingga selaras dengan tujuan *hifz al-'ird* dalam *maqaṣid al-syari'ah* (Kamali, 2008).

Berdasarkan landasan teknis dan normatif tersebut, berbagai penelitian menunjukkan bahwa integrasi model deteksi cepat, , dan algoritma pelacakan mampu membentuk sistem pengenalan wajah yang tangguh untuk lingkungan keramaian. Pendekatan ini menjaga keseimbangan antara kecepatan, akurasi, dan stabilitas, sekaligus mendukung tujuan perlindungan jiwa dan kehormatan manusia. Oleh karena itu, penelitian ini merancang pendekatan yang mengintegrasikan YOLOv9, *Attention Mechanism*, dan untuk menghadapi tantangan pengenalan wajah di keramaian secara komprehensif dan berkeadilan.

1.2 Rumusan Masalah

Berdasarkan permasalahan deteksi dan pelacakan wajah pada lingkungan keramaian yang dipengaruhi oleh *occlusion*, variasi pencahayaan, pergerakan dinamis, serta kepadatan objek, maka rumusan masalah dalam penelitian ini adalah bagaimana merancang, mengimplementasikan, dan mengevaluasi sistem deteksi dan pelacakan wajah menggunakan integrasi YOLOv9 dengan , *MobileFaceNet*, dan *DeepSORT*, serta menganalisis pengaruh variasi rasio data pelatihan dan parameter pelatihan berupa *epoch*, *batch size*, dan *learning rate* terhadap performa deteksi dan pelacakan wajah pada lingkungan keramaian berdasarkan metrik evaluasi deteksi, pelacakan, dan performa *real-time*.

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi sistem deteksi dan pelacakan wajah pada lingkungan keramaian menggunakan integrasi YOLOv9 dengan , *MobileFaceNet*, dan DeepSORT, serta menganalisis pengaruh variasi rasio data pelatihan dan parameter pelatihan berupa *epoch*, *batch size*, dan *learning rate* terhadap performa sistem berdasarkan metrik deteksi, pelacakan, dan performa *real-time* guna memperoleh konfigurasi model yang optimal, stabil, dan robust pada kondisi keramaian padat.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan beberapa manfaat sebagai berikut:

1. Manfaat Keilmuan : Memberikan kontribusi terhadap pengembangan ilmu visi komputer dengan memperkaya kajian mengenai integrasi YOLOv9, *Attention Mechanism*, dan dalam sistem pengenalan wajah pada lingkungan keramaian.
2. Manfaat Teknis : Menyediakan rancangan sistem deteksi dan pelacakan wajah yang mampu bekerja secara konsisten pada video keramaian, khususnya dalam menghadapi *occlusion*, pergerakan cepat, dan variasi pencahayaan.
3. Manfaat Aplikatif : Menjadi dasar pengembangan sistem pengawasan cerdas yang dapat diterapkan pada lingkungan publik seperti fasilitas transportasi, ruang terbuka, dan pusat keramaian untuk mendukung kebutuhan keamanan dan pemantauan otomatis.

4. Manfaat Pengembangan Sistem : Memberikan referensi evaluasi performa terkait akurasi deteksi dan stabilitas pelacakan yang dapat dimanfaatkan dalam pengembangan sistem pengenalan wajah berbasis AI di masa mendatang.

BAB II

LANDASAN TEORI

2.1 Penelitian Terkait

Penelitian mengenai deteksi dan pelacakan wajah pada lingkungan keramaian telah mengalami perkembangan signifikan seiring meningkatnya kebutuhan sistem pemantauan cerdas pada area publik. Model pendeteksi objek berbasis *deep learning* terus dikembangkan untuk meningkatkan akurasi dalam kondisi visual yang kompleks, seperti perubahan pencahayaan, *occlusion*, dan jarak pandang kamera yang bervariasi. Berbagai studi yang berfokus pada penggunaan YOLOv9, mekanisme perhatian, ekstraksi fitur wajah, dan pelacakan multiobjek memberikan dasar ilmiah yang kuat bagi penelitian ini.

Penelitian oleh Narayanan et al. (2025) membahas penggunaan YOLOv9 untuk mendeteksi wajah dalam kondisi visual yang beragam, termasuk *noise*, *blur*, dan perubahan pencahayaan. Arsitektur YOLOv9 pada studi tersebut menunjukkan kemampuan ekstraksi fitur yang baik meskipun data mengalami gangguan visual. Temuan ini relevan karena deteksi wajah pada lingkungan ramai sering menghasilkan kualitas gambar yang tidak stabil dan membutuhkan model yang tangguh terhadap variasi tersebut.

Studi oleh Alsabei et al. (2025) menerapkan YOLOv9 pada pemantauan kerumunan jamaah Haji, lingkungan yang dipenuhi *occlusion*, pergerakan cepat, dan perubahan intensitas cahaya. Hasil penelitian ini memperlihatkan bahwa YOLOv9 mampu mempertahankan akurasi dan kecepatan pemrosesan secara real-

time pada kondisi padat dan dinamis. Hal ini memberikan dukungan kuat terhadap pemilihan YOLOv9 sebagai model deteksi wajah pada penelitian ini.

Masalah kualitas gambar dijelaskan oleh Singh dan Reibman (2024) melalui pengembangan *Task-Aware Image Quality Estimator*. Penelitian tersebut menunjukkan bagaimana *blur*, *noise*, dan distorsi pencahayaan dapat menurunkan kinerja detektor wajah. Pendekatan ini menguatkan pentingnya pra-pemrosesan dan peningkatan kualitas gambar untuk mendukung stabilitas performa sistem deteksi wajah pada kondisi nyata.

Penanganan occlusion berat dikaji oleh Thaher et al. (2025) melalui integrasi HOG dan *deep learning*. Penelitian ini menunjukkan bahwa metode yang mempertimbangkan pola visual lokal dapat tetap mendeteksi wajah meskipun sebagian area wajah tertutup. Temuan tersebut selaras dengan kebutuhan penelitian ini karena wajah pada lingkungan ramai sering terekam dalam kondisi tidak utuh akibat tumpang tindih antarindividu.

Dalam pengembangan fitur wajah, Xue (2025) menunjukkan bahwa *MobileFaceNet* mampu menghasilkan *embedding* wajah yang efisien dan stabil pada berbagai kondisi visual. Model ini menghasilkan representasi fitur berdimensi rendah yang cocok diterapkan pada sistem pelacakan *real-time*, sehingga relevan untuk digunakan pada tahap ekstraksi fitur dalam penelitian ini.

Penelitian oleh Djarah et al. (2024) mengintegrasikan YOLOv9 dan untuk pelacakan multiobjek dan menunjukkan bahwa mampu mempertahankan identitas objek secara konsisten melalui prediksi gerakan dan pembandingan *embedding*. Temuan ini mendukung penggunaan sebagai mekanisme pelacakan pada sistem

yang memerlukan stabilitas identitas antarframe, seperti pelacakan wajah di keramaian.

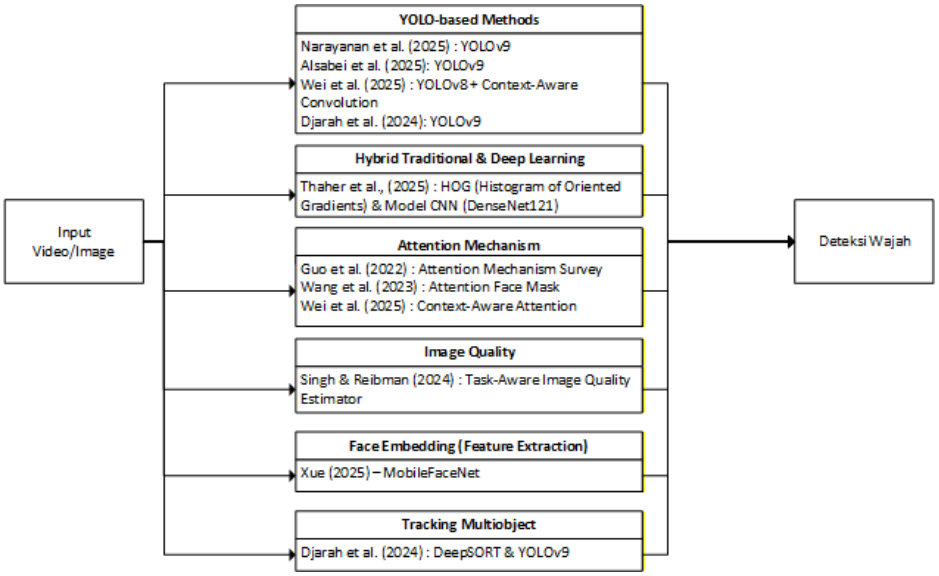
Dalam ranah teori *attention*, Guo et al. (2022) menyajikan survei komprehensif mengenai mekanisme *attention* pada *computer vision*. Studi ini menjelaskan bagaimana *attention* meningkatkan representasi fitur dengan menyoroti area penting dalam gambar dan mengurangi pengaruh bagian yang kurang relevan. Konsep ini relevan dengan kebutuhan penelitian ini karena wajah dalam keramaian sering mengalami *occlusion* atau hanya terlihat sebagian.

Penerapan *attention* untuk wajah yang mengalami *occlusion* juga ditunjukkan oleh Wang et al. (2023) melalui penelitian mengenai pengenalan wajah bermasker. Studi tersebut menunjukkan bahwa *attention mechanism* dapat memperkuat fitur wajah yang tetap terlihat meskipun sebagian area tertutup masker. Temuan ini relevan karena kondisi serupa juga muncul pada keramaian, di mana sebagian wajah sering terhalang oleh objek lain.

Pendekatan berbasis konteks dikembangkan oleh Wei et al. (2025) melalui *context-aware convolutions*, yang memanfaatkan informasi sekitar wajah untuk mempertahankan deteksi ketika sebagian fitur wajah tidak terlihat. Pendekatan ini memberikan dukungan bahwa penggunaan perhatian berbasis konteks dapat meningkatkan ketelitian deteksi dalam lingkungan visual yang kompleks dan padat.

2.2 Kerangka Teori

Kerangka teori dari penelitian ini didasarkan pada penelitian sebelumnya atau studi yang relevan yang di kutip sebagai titik acuan. Berikut kerangka teori dalam penelitian ini dapat dilihat pada gambar 2.1.



Gambar 2.1 Kerangka Teori

Gambar 2.1 menggambarkan hubungan antar komponen teori yang menjadi dasar penyusunan sistem deteksi dan pelacakan wajah pada penelitian ini. Diagram tersebut menunjukkan bagaimana berbagai pendekatan dalam bidang visi komputer digunakan untuk memproses *input* berupa video atau gambar hingga menghasilkan *output* berupa deteksi wajah yang stabil. Setiap kelompok metode merepresentasikan kontribusi penelitian terdahulu yang mendukung komponen sistem yang dikembangkan.

Bagian pertama pada diagram menunjukkan kelompok YOLO-*based Methods*, yang meliputi beberapa penelitian yang memanfaatkan YOLOv9 atau pendekatan YOLO lain dalam mendeteksi wajah pada kondisi visual yang

kompleks. Kelompok ini menjadi landasan utama bagi proses deteksi cepat dan akurat karena model YOLO memiliki kemampuan untuk memproses informasi secara real time. Penelitian oleh Narayanan et al. (2025), Alsabei et al. (2025), dan Djarah et al. (2024) menjadi dasar bagi elemen deteksi objek dalam sistem ini, sedangkan Wei et al. (2025) menambahkan konteks penggunaan YOLO dengan mekanisme berbasis konteks.

Kelompok berikutnya, *Hybrid Traditional* & *Deep Learning*, merujuk pada pendekatan yang menggabungkan metode klasik seperti *Histogram of Oriented Gradients* (HOG) dengan model *deep learning*. Penelitian Thaher et al. (2025) menegaskan bahwa kombinasi metode tradisional dan CNN dapat mempertahankan deteksi pada wajah yang mengalami *occlusion* berat. Pengetahuan ini mendukung pemahaman bahwa sistem deteksi wajah perlu mempertimbangkan kondisi wajah yang tidak selalu terlihat secara penuh dalam keramaian.

Selanjutnya, kelompok *Attention Mechanism* berisi penelitian yang menjelaskan bagaimana mekanisme perhatian dapat memperkuat representasi fitur wajah dalam berbagai kondisi visual. Studi oleh Guo et al. (2022), Wang et al. (2023), dan Wei et al. (2025) menunjukkan bahwa *attention* mampu menyoroti area wajah yang paling relevan, sehingga membantu model dalam mempertahankan ketelitian deteksi ketika sebagian wajah tertutup atau terlihat dalam kondisi yang kurang ideal. Konsep *attention* ini kemudian diintegrasikan dalam sistem untuk meningkatkan keandalan ekstraksi fitur.

Komponen *Image Quality* menunjukkan pentingnya kualitas *input* sebagai faktor yang memengaruhi akurasi deteksi. Penelitian Singh dan Reibman (2024)

menekankan bahwa penilaian kualitas gambar membantu memetakan kelayakan gambar sebelum diproses lebih lanjut. Hal ini relevan ketika sistem menerima input dari kamera pengawasan yang memiliki kualitas tidak stabil.

Kelompok *Face Embedding* (*Feature Extraction*) digambarkan melalui penelitian Xue (2025), yang menjelaskan bagaimana *MobileFaceNet* dapat menghasilkan embedding wajah yang efisien dan stabil. *Embedding* ini sangat penting untuk menjaga konsistensi identitas objek pada tahap pelacakan, terutama dalam video yang memiliki banyak individu.

Terakhir, kelompok *Tracking Multiobject* merujuk pada pendekatan pelacakan berbasis yang digunakan untuk mempertahankan identitas wajah dari satu *frame* ke *frame* berikutnya. Penelitian Djarah et al. (2024) menunjukkan bagaimana kombinasi prediksi gerakan dan pembandingan *embedding* membantu menghasilkan pelacakan yang stabil pada lingkungan padat.

Secara keseluruhan, diagram Kerangka Teori mengilustrasikan integrasi antara berbagai penelitian yang mendukung komponen sistem, mulai dari deteksi wajah, penguatan fitur, peningkatan kualitas visual, ekstraksi *embedding*, hingga pelacakan multiobjek. Semua elemen tersebut bekerja bersama untuk menghasilkan sistem yang mampu mendeteksi dan melacak wajah pada lingkungan keramaian secara akurat dan efisien.

No.	Judul	Penulis & Tahun	Yang Diteliti	Metode	Fitur	Hasil
1	YOLOv9-Based Human Face Detection and Counting Under Complex Image Qualities	Narayanan et al. (2025)	Deteksi wajah pada kondisi visual sulit (blur, noise, low-light)	YOLOv9	Backbone-Neck-Head YOLOv9	Deteksi stabil pada kualitas gambar tidak ideal
2	Enhancing Crowd Safety at Hajj Using YOLOv9	Alsabei et al. (2025)	Deteksi objek pada kerumunan besar dan dinamis	YOLOv9	Deteksi real-time	Akurasi tinggi pada crowd dengan occlusion berat
3	Task-Aware Image Quality Estimator for Face Detection	Singh & Reibman (2024)	Pengaruh kualitas gambar terhadap deteksi wajah	Image Quality Estimator (DET score)	Penilaian kualitas visual	Kualitas gambar menentukan performa detektor
4	A Hybrid Approach for Heavily Occluded Face Detection Using HOG and Deep Learning	Thaher et al. (2025)	Deteksi wajah dengan occlusion berat	HOG dan CNN	Fitur lokal wajah	Deteksi tetap muncul meskipun wajah tertutup sebagian besar

5	Facial Recognition for Surveillance Videos Based on RetinaFace and MobileFaceNet	Xue (2025)	Representasi fitur wajah untuk aplikasi video surveillance	MobileFaceNet dan RetinaFace	Embedding 128-dim	Embedding stabil untuk wajah dalam video
6	Online Multi-Object Tracking with YOLOv9 and Optimized by Optical Flow	Djarah et al. (2024)	Pelacakan multiobjek berbasis YOLOv9	YOLOv9 dan *DeepSORT*	Kalman Filter dan Cosine Metric	Pelacakan stabil pada objek bergerak cepat
7	Attention Mechanisms in Computer Vision: A Survey	Guo et al. (2022)	Mekanisme attention dalam visi komputer	Spatial dan Channel Attention	Penguatan area penting	Attention meningkatkan fokus model pada fitur relevan
8	Masked Face Recognition System Based on Attention Mechanism	Wang et al. (2023)	Pengenalan wajah bermasker	Attention-based Recognition	Penekanan fitur wajah terbuka	Identifikasi tetap stabil pada wajah bermasker
9	Robust Face Mask Detection Using Context-Aware Convolutions	Wei et al. (2025)	Deteksi wajah bermasker pada skenario kompleks	Context-Aware Attention dan YOLOv8	Penguatan fitur melalui konteks	Deteksi meningkat meskipun fitur wajah tidak lengkap

Tabel 2.1 Penelitian Terdahulu

2.3 YOLOv9

YOLOv9 (*You Only Look Once version* 9) merupakan pengembangan terbaru dari keluarga algoritma YOLO yang dirancang untuk meningkatkan kecepatan dan akurasi dalam deteksi objek secara *real-time*. Pada penelitian Alsabei et al. (2025) dan Narayanan et al. (2025), YOLOv9 digunakan dalam konteks deteksi perilaku abnormal pada kerumunan Haji serta deteksi dan perhitungan wajah pada lingkungan kompleks yang melibatkan wajah manusia, hewan, maupun kondisi gambar rendah. Model ini memperkenalkan beberapa inovasi arsitektural, seperti *RepNCSPELAN4 block*, *SPPELAN* module, serta *Programmable Gradient Information* (PGI), yang secara kolektif bertujuan untuk memperbaiki aliran informasi, mengurangi kompleksitas komputasi, dan meningkatkan generalisasi pada kondisi lingkungan nyata.

Secara garis besar, alur kerja YOLOv9 terdiri dari tiga tahap utama, yaitu *Backbone*, *Neck*, dan *Head*. Pada bagian *Backbone*, model mengekstraksi fitur visual dari gambar *input* menggunakan konvolusi berulang dan blok khusus (*RepNCSPELAN4*) yang memaksimalkan penggunaan kembali fitur tanpa menambah beban komputasi secara signifikan. Tahap *Neck* melakukan fusi multi-skala dengan teknik *upsampling*, *concatenation*, serta modul *SPPELAN* yang memperluas *receptive field*. Bagian Head bertugas menghasilkan prediksi akhir berupa bounding box $B = (x, y, w, h)$ skor *objectness*, dan probabilitas kelas. Proses ini dirancang agar mampu mendeteksi wajah atau perilaku abnormal meskipun objek dalam kondisi *occlusion*, *noise*, atau *low-light*.

Secara matematis, YOLOv9 memformulasikan masalah deteksi objek sebagai regresi langsung terhadap koordinat *bounding box* dan probabilitas kelas. Prediksi *bounding box* biasanya dinyatakan dalam bentuk:

$$\widehat{b_x} = \sigma(t_x) + c_x,\ \widehat{b_y} = \sigma(t_y) + c_y,\ \widehat{b_w} = p_w \cdot e^{t_w},\ \widehat{b_h} = p_h \cdot e^{t_h} \qquad (2.1)$$

dengan (c_x, c_y) adalah koordinat sel grid, (p_w, p_h) ukuran *anchor box*, dan (t_x, t_y, t_w, t_h) parameter hasil prediksi. Fungsi *sigmoid* σ*sigma* digunakan untuk menormalisasi prediksi koordinat relatif terhadap sel grid.

Selain koordinat, YOLOv9 juga menghasilkan objectness score dan probabilitas kelas. Probabilitas keberadaan kelas $P(class_i|object)$ dihitung menggunakan fungsi *Softmax*:

$$P(class_i|object) = \frac{e^{z_i}}{\sum_j e^{z_j}} \qquad (2.2)$$

dengan z_i adalah skor logit untuk kelas ke-i. Prediksi akhir berupa skor gabungan:

$$PP(object) \cdot P(class_i|object) \qquad (2.3)$$

yang menunjukkan keyakinan model bahwa sebuah objek dengan kelas tertentu benar-benar hadir di lokasi *bounding box*.

Untuk mengoptimalkan pelatihan, YOLOv9 menggunakan kombinasi beberapa fungsi kerugian (*loss function*). Kerugian koordinat *bounding box* dihitung dengan *Complete IoU* (CIoU) *Loss* atau EIoU *Loss*:

$$LCIoU = 1 - IoU + + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \qquad (2.4)$$

dengan IoU adalah *Intersection over Union*, ρ jarak Euclidean antara pusat prediksi dan pusat ground truth, c diagonal minimum *bounding box* yang mencakup keduanya, v perbedaan aspek rasio, dan α koefisien keseimbangan.

Sedangkan kerugian klasifikasi dan objectness dihitung dengan *Binary Cross-Entropy Loss* (BCE):

$$L_{BCE} = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \quad (2.5)$$

Fungsi kerugian total dalam YOLOv9 kemudian dinyatakan sebagai kombinasi berbobot:

$$L = \lambda_{box} L_{CIoU} + \lambda_{obj} L_{BCE}^{obj} + \lambda_{cls} L_{BCE}^{cls} \quad (2.6)$$

dengan $\lambda_{box}, \lambda_{obj}, \lambda_{cls}$ adalah bobot untuk setiap komponen.

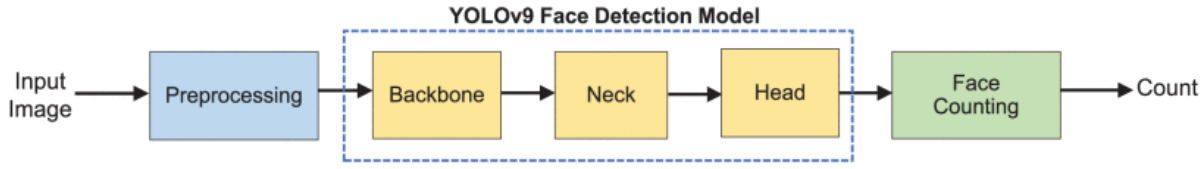
Dalam praktiknya, YOLOv9 juga memanfaatkan *Soft*-NMS (*Non-Maximum Suppression*) yang dimodifikasi untuk mempertahankan *bounding box* yang saling tumpang tindih namun tetap relevan. Hal ini penting pada kasus *crowd detection*, karena wajah sering saling menutupi. Dengan mekanisme ini, YOLOv9 terbukti mampu memberikan akurasi mAP lebih tinggi dibanding YOLOv4-YOLOv8, sekaligus menjaga kecepatan inferensi real-time di perangkat terbatas seperti Raspberry Pi.

Berdasarkan kajian dari jurnal-jurnal tersebut, dapat dipahami bahwa YOLOv9 tidak hanya sekadar penyempurnaan dari versi sebelumnya, tetapi juga sebuah framework adaptif yang dapat dioptimalkan untuk berbagai konteks: mulai dari deteksi wajah di *crowd* padat, penghitungan jumlah individu, hingga deteksi perilaku abnormal yang berpotensi menimbulkan bahaya. Studi terkait performa YOLO pada deteksi wajah dengan atribut tambahan seperti masker menunjukkan bahwa model YOLO memiliki stabilitas yang baik pada berbagai variasi dataset dan resolusi gambar (Kurniawan et al., 2023). Hal ini membuat YOLOv9 sangat

relevan sebagai pijakan dalam penelitian terkini terkait pengolahan gambar di lingkungan kompleks dan dinamis.

Sebagai pelengkap penjelasan mengenai arsitektur YOLOv9, Gambar berikut memperlihatkan alur lengkap dari proses deteksi wajah hingga perhitungan jumlah wajah. Model dimulai dengan preprocessing untuk menyesuaikan ukuran dan kualitas input gambar, kemudian diteruskan ke tiga komponen utama YOLOv9, yaitu *Backbone*, *Neck*, dan *Head*. *Backbone* berfungsi sebagai feature extractor yang mengekstraksi pola dasar dari gambar, *Neck* melakukan *multi-scale feature fusion* untuk menggabungkan informasi dari berbagai tingkat kedalaman, sedangkan Head menghasilkan *bounding box*, skor *objectness*, dan probabilitas kelas.

Setelah tahap *Head*, YOLOv9 dapat diperluas dengan modul tambahan, salah satunya adalah *Face Counting*, yang berfungsi menghitung jumlah wajah yang berhasil terdeteksi. Tahap ini penting dalam aplikasi kepadatan kerumunan atau crowd management, di mana bukan hanya keberadaan wajah yang ingin diketahui, tetapi juga jumlah individu dalam suatu adegan. Dengan demikian, model YOLOv9 tidak hanya menjadi detektor wajah, tetapi juga mendukung fungsi *crowd analytics* yang lebih luas.



Gambar 2.2 Diagram Arsitektur YOLOv9 *Face Detection Model* dengan *Face Counting* (Narayanan et al., 2025)

2.4 Attention Mechanism

Attention Mechanism merupakan komponen penting dalam pengembangan model *deep learning* modern yang berfungsi memberikan kemampuan adaptif bagi jaringan untuk fokus pada bagian informasi yang paling relevan dari suatu representasi. Dalam konteks visi komputer, mekanisme ini membantu model memprioritaskan area tertentu dalam suatu gambar dengan menimbang kontribusi fitur berdasarkan tingkat kepentingannya. Pendekatan ini dikembangkan pertama kali dalam domain *natural language processing* oleh Bahdanau et al. (2015) dan diaplikasikan secara luas pada *Convolutional Neural Networks* (CNN) seperti yang dikaji oleh Fabrianto, et al. (2025), yang memperkenalkan konsep *alignment* untuk menentukan relevansi suatu elemen input terhadap proses prediksi. Mekanisme tersebut kemudian berkembang lebih jauh melalui arsitektur *Transformer* yang dikemukakan oleh Vaswani et al. (2017) serta dieksplorasi penerapannya dalam model visi komputer oleh Chusna & Khumaidi (2024), yang mengadopsi struktur *self-attention* untuk menangkap ketergantungan antarfitur dalam satu urutan secara lebih efisien.

Mekanisme *self-attention* bekerja dengan memproyeksikan fitur input X ke dalam tiga representasi berbeda, yaitu *Query* (Q), *Key* (K), dan *Value* (V), melalui transformasi linear menggunakan matriks bobot $W_Q, W_K,$dan W_V. Representasi tersebut dirumuskan sebagai berikut:

$$Q = XW_Q, K = XW_K, V = XW_V. \tag{2.6}$$

Selanjutnya, hubungan antara *Query* dan *Key* dihitung menggunakan operasi dot-product:

$$Score(Q, K) = QK^T, \tag{2.7}$$

yang kemudian dinormalisasi melalui fungsi *softmax* menjadi:

$$\alpha = softmax(QK^T), \tag{2.8}$$

sehingga keluaran akhir *attention* diperoleh melalui:

$$Attention(Q, K, V) = \alpha V. \tag{2.9}$$

Formulasi matematis ini memungkinkan model untuk menilai tingkat relevansi setiap fitur terhadap fitur lainnya, sehingga model dapat memusatkan perhatian pada bagian informasi yang lebih bermakna. Pada visi komputer, kemampuan ini sangat penting karena gambar memiliki distribusi informasi yang tidak merata. Area tertentu seperti mata, hidung, atau kontur wajah memiliki pengaruh lebih besar dalam proses identifikasi dibandingkan bagian latar belakang atau area kosong.

Penerapan attention dalam visi komputer telah berkembang menjadi beberapa kategori. Menurut Guo et al. (2022), *attention* dapat dibagi menjadi attention spasial, kanal, gabungan keduanya, serta attention berbasis konteks. *Attention* spasial bekerja dengan memetakan lokasi penting dalam gambar sehingga model dapat fokus pada area yang relevan, misalnya bagian wajah yang terlihat ketika sebagian wajah mengalami *occlusion*. Untuk attention kanal, pendekatan ini menentukan kanal fitur mana yang memiliki kontribusi dominan dalam representasi. Salah satu implementasi populer dari attention kanal adalah *Squeeze-and-Excitation (SE) Block*, yang menghitung representasi kanal melalui *Global Average Pooling* dengan rumus:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j), \tag{2.10}$$

yang kemudian diproyeksikan ke dalam jaringan penuh untuk menghasilkan bobot perhatian kanal:

$$s = \sigma(W_2\, \delta(W_1\, z)), \hat{X}_c = s_c \cdot X_c. \tag{2.11}$$

Hasil ini memberikan penguatan terhadap kanal yang memiliki kontribusi terbesar sehingga fitur penting dapat tetap dipertahankan meskipun fitur pendukung mengalami gangguan visual.

Selain attention spasial dan kanal, perhatian berbasis konteks menjadi penting pada gambar yang mengalami penutupan sebagian. Pendekatan ini diperkuat oleh penelitian Wei et al. (2025) yang mendemonstrasikan bahwa memanfaatkan informasi visual dari area sekitar wajah mampu membantu model mendeteksi wajah meskipun sebagian besar fitur utama tertutup. Pendekatan context-aware ini menekankan bahwa pola visual di area non-wajah dapat membantu memprediksi struktur wajah secara lebih baik ketika fitur utama tidak tersedia secara lengkap.

Attention juga digunakan secara luas dalam pengenalan wajah pada kondisi terhalang. Wang et al. (2023) menunjukkan bahwa attention mampu memperkuat bagian wajah yang tetap terlihat ketika wajah tertutup masker, seperti mata dan dahi. Hal ini menghasilkan representasi fitur yang lebih stabil untuk keperluan deteksi dan pengenalan wajah pada lingkungan yang tidak terkendali. Temuan tersebut berdampak langsung pada penelitian ini, karena wajah dalam keramaian

sering mengalami occlusion yang disebabkan oleh tumpang tindih antarindividu atau pergerakan cepat yang menyebabkan wajah tidak terekam secara utuh.

Dalam konteks penelitian ini, *Attention Mechanism* digunakan untuk meningkatkan kemampuan YOLOv9 dalam mengekstraksi fitur wajah yang lebih kuat dan selektif. Lingkungan keramaian memiliki tingkat dinamika visual yang tinggi, mencakup perubahan *pose*, *occlusion*, perbedaan jarak kamera, serta perubahan intensitas cahaya yang memengaruhi kualitas fitur wajah. Dengan memanfaatkan attention, sistem deteksi dapat lebih fokus pada bagian wajah yang tetap terlihat dan mengurangi pengaruh gangguan visual, sehingga hasil deteksi menjadi lebih konsisten. Selain itu, peningkatan representasi fitur melalui attention mendukung keberlanjutan proses pelacakan menggunakan *DeepSORT*, karena embedding wajah yang lebih stabil akan mempermudah proses asosiasi identitas antarframe.

Secara keseluruhan, *Attention Mechanism* memberikan landasan penting dalam meningkatkan ketahanan model terhadap kondisi visual yang tidak ideal. Integrasinya ke dalam pipeline deteksi dalam penelitian ini berfungsi memperkuat fitur wajah, mengatasi kendala occlusion, dan meningkatkan akurasi deteksi secara keseluruhan. Teori dasar dari Bahdanau et al. (2015), konsep arsitektur self-attention dari Vaswani et al. (2017) serta Chusna & Khumaidi (2024), serta temuan empiris dari penelitian terbaru dalam bidang visi komputer membentuk fondasi yang kuat bagi pendekatan yang digunakan pada penelitian ini.

2.5 MobileFaceNet

MobileFaceNet merupakan arsitektur jaringan saraf konvolusional ringan yang dirancang untuk menghasilkan embedding wajah secara efisien dengan tetap mempertahankan akurasi representasi fitur. Model ini dikembangkan dengan mengadaptasi konsep dasar dari *MobileNet*, yaitu *depthwise separable convolution*, yang memisahkan operasi konvolusi menjadi dua tahap: *depthwise convolution* untuk memproses tiap kanal secara independen, dan *pointwise convolution* untuk menggabungkan informasi antar kanal. Metode ini mengurangi jumlah parameter dan kompleksitas komputasi secara signifikan sebagaimana dijelaskan oleh Howard et al. (2017) serta dievaluasi efisiensinya untuk pengenalan wajah oleh Xiao, et al. (2022), yang menjadi dasar arsitektur *MobileFaceNet*.

Selain itu, *MobileFaceNet* memanfaatkan struktur bottleneck residual seperti pada *MobileNetV2* yang dikembangkan oleh Sandler et al. (2018) dan dioptimasi lebih lanjut dalam pemrosesan citra medis dan visi komputer oleh Somoal & Dzikrillah (2025). Struktur ini terdiri dari tiga komponen: lapisan ekspansi untuk memperbesar dimensi fitur, *depthwise convolution* untuk melakukan ekstraksi fitur spasial, dan lapisan proyeksi untuk mengembalikan dimensi fitur agar tetap ringan. Penggunaan *inverted residuals* meningkatkan stabilitas aliran gradien dan mendukung pembelajaran fitur wajah yang tetap konsisten meskipun terdapat perubahan pose, ekspresi, atau pencahayaan. Dengan demikian, struktur dasar *MobileFaceNet* memungkinkan model menghasilkan fitur wajah secara efisien tanpa memerlukan sumber daya komputasi besar.

Proses ekstraksi fitur pada *MobileFaceNet* menghasilkan *embedding* wajah dalam bentuk vektor berdimensi rendah. *Embedding* tersebut merepresentasikan struktur numerik dari pola wajah yang dapat dibandingkan antar individu. Representasi ini digunakan secara luas dalam sistem pengenalan wajah modern, terutama setelah adanya pengembangan fungsi margin seperti *ArcFace* (Deng et al., 2019) serta pendekatan *transfer learning* seperti yang dikaji oleh Shukla & Tiwari (2023), yang menegaskan pentingnya embedding terdistribusi baik pada ruang vektor. Meskipun penelitian ini menggunakan *embedding* untuk keperluan pelacakan, bukan identifikasi, konsep representasi terstruktur ini tetap relevan karena embedding yang stabil berperan penting dalam mempertahankan identitas objek antarframe.

Pada implementasinya, *MobileFaceNet* menerapkan *global average pooling* pada bagian akhir jaringan untuk menghasilkan representasi global dari wajah. Tahap ini diikuti oleh proses normalisasi, seperti *L2-normalization*, agar *embedding* memiliki panjang vektor yang konsisten sehingga jarak antar embedding dapat dihitung menggunakan metrik seperti *cosine similarity*. Rumus dasarnya dituliskan sebagai berikut:

$$cosine(x, y) = \frac{x \cdot y}{\| x \| \ \| y \|}, \tag{2.12}$$

yang menunjukkan tingkat kesamaan arah *vektor* antara dua *embedding* wajah. Dalam konteks pelacakan wajah, embedding yang konsisten membantu menjaga keakuratan proses asosiasi identitas dalam algoritma pelacakan seperti *DeepSORT*.

Penelitian terbaru oleh Xue (2025), menunjukkan bahwa *MobileFaceNet* bekerja efektif dalam sistem pengawasan berbasis video. Model ini mampu menghasilkan embedding yang stabil pada wajah dengan variasi pose dan pencahayaan. Temuan ini memperkuat alasan pemilihan *MobileFaceNet* dalam penelitian ini karena lingkungan keramaian cenderung menghasilkan variasi tampilan wajah yang sangat dinamis. Selain itu, penelitian oleh Zhang et al. (2024) menunjukkan bahwa pengembangan lanjutan dari *MobileFaceNet* dapat meningkatkan ketahanan representasi fitur melalui optimasi arsitektur dan penyesuaian blok konvolusi. Hal tersebut menunjukkan fleksibilitas struktur *MobileFaceNet* dalam menyesuaikan kebutuhan sistem pengenalan atau pelacakan wajah.

Dalam penelitian ini, *MobileFaceNet* digunakan untuk menghasilkan embedding wajah yang menjadi dasar proses pelacakan identitas. Representasi fitur yang stabil dari model ini memungkinkan algoritma mempertahankan identitas wajah meskipun terjadi perubahan orientasi, jarak, atau *occlusion* parsial. Keunggulan *MobileFaceNet* dalam hal efisiensi komputasi juga menjadikannya ideal untuk diterapkan bersama detektor real-time seperti YOLOv9, yang memerlukan proses pasca-determinasi yang cepat dan akurat. Integrasi keduanya memberikan landasan kuat bagi sistem pelacakan wajah yang handal dalam lingkungan keramaian.

Secara keseluruhan, *MobileFaceNet* memberikan kontribusi signifikan sebagai model ekstraksi fitur wajah yang efisien, stabil, dan adaptif terhadap perubahan kondisi visual. Keunggulan dalam struktur ringan, kombinasi *depthwise*

separable convolution, serta mekanisme representasi fitur yang stabil menjadikannya sangat relevan untuk diterapkan dalam sistem pelacakan wajah berbasis video yang memerlukan kinerja real-time dan akurasi tinggi.

2.6 DeepSORT

merupakan algoritma pelacakan multiobjek yang dikembangkan sebagai pengembangan dari sistem SORT (*Simple Online and Real-time Tracking*). Tujuan utama algoritma ini adalah mempertahankan identitas objek pada setiap frame video sehingga pergerakan objek dapat diikuti secara berkesinambungan. SORT yang dikembangkan oleh Bewley et al. (2016) dan dikaji penerapannya bersama algoritma YOLO oleh Susetianingtias, et al. (2023) hanya mengandalkan informasi posisi objek yang diperoleh dari deteksi bounding box setiap frame menggunakan metode *tracking-by-detection*. Walaupun bekerja sangat cepat, SORT memiliki kelemahan pada kondisi *occlusion*, perubahan gerakan yang tidak terprediksi, dan interaksi antarobjek dalam lingkungan padat. Untuk mengatasi keterbatasan tersebut, Wojke et al. (2017) beserta penerapannya yang dievaluasi untuk perhitungan dan pelacakan objek oleh Cahyani et al. (2024) memperkenalkan dengan menambahkan komponen *appearance feature embedding* untuk memperkuat proses asosiasi identitas.

menggabungkan dua komponen penting, yaitu *motion model* dan *appearance model*. Motion model digunakan untuk memprediksi pergerakan objek antarframe menggunakan *Kalman Filter*. Pada pelacak multiobjek, *Kalman Filter* digunakan untuk memperkirakan posisi dan kecepatan objek berdasarkan observasi

sebelumnya melalui persamaan *state-space*. Proses prediksi dan pembaruan dilakukan melalui dua rumus dasar:

Prediksi state:

$$\hat{x}_{k|k-1} = F\hat{x}_{k-1|k-1} \tag{2.13}$$

Prediksi kovariansi:

$$P_{k|k-1} = FP_{k-1|k-1}F^T + Q \tag{2.14}$$

Di mana Fadalah matriks transisi, Padalah kovariansi *error*, dan Q adalah *noise* proses. Ketika deteksi baru muncul pada *frame* ke-k, *Kalman Filter* memperbarui prediksi melalui:

***Update state*:**

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K(z_k - H\hat{x}_{k|k-1}) \tag{2.15}$$

***Update* kovariansi:**

$$P_{k|k} = (I - KH)P_{k|k-1} \tag{2.16}$$

Pendekatan berbasis prediksi ini memungkinkan menghitung kemungkinan lokasi objek pada *frame* berikutnya meskipun terjadi pergerakan cepat atau hilangnya sebagian fitur akibat *occlusion*.

Sementara itu, komponen *appearance model* pada menggunakan representasi fitur (*embedding*) dari gambar objek yang diekstraksi menggunakan jaringan deep learning. *Appearance embedding* memungkinkan sistem tidak hanya mengandalkan informasi posisi, tetapi juga pola visual yang membedakan satu objek dari objek lain. Wojke et al. (2017) serta penelitian pendukung dari Cahyani et al. (2024) menggunakan CNN ringan untuk menghasilkan embedding berdimensi rendah, yang kemudian digunakan untuk menghitung kesamaan antar frame

menggunakan jarak *cosine*. Pendekatan ini membantu objek tetap dikenali meskipun bergerak secara cepat, berubah sudut pandang, atau mengalami tumpang tindih dengan objek lain.

Proses asosiasi identitas dalam dilakukan melalui algoritma *Hungarian Algorithm* yang mencari kecocokan optimal antara prediksi *Kalman Filter* dan hasil deteksi pada *frame* saat ini. Kombinasi antara motion model dan appearance model membuat lebih stabil dibandingkan *SORT* dan metode pelacakan sederhana lainnya. Ketika *occlusion* terjadi, motion model membantu memprediksi keberadaan objek berdasarkan pergerakan sebelumnya, sedangkan *appearance model* membantu memastikan bahwa objek tetap dikenali setelah keluar dari area tertutup.

Penelitian oleh Djarah et al. (2024) menunjukkan bahwa integrasi dan YOLOv9 menghasilkan performa pelacakan yang stabil dalam lingkungan dinamis dengan kepadatan objek tinggi. Detektor YOLOv9 memberikan *bounding box* yang akurat pada setiap frame, sementara *DeepSORT* memanfaatkan informasi tersebut untuk mempertahankan identitas objek secara konsisten. Pendekatan gabungan ini terbukti efektif pada kondisi visual kompleks, termasuk perubahan arah gerakan, perbedaan skala objek, hingga tumpang tindih antarindividu. Dengan dukungan embedding, sistem tetap mampu melakukan asosiasi objek meskipun wajah terlihat dengan sudut berbeda atau mengalami occlusion sebagian.

DeepSORT juga memiliki keunggulan penting lainnya, yaitu kemampuannya untuk menghapus *lost tracks* dan menambah *new tracks* secara adaptif. Ketika objek menghilang dalam beberapa *frame* (misalnya tertutup objek

lain), pelacak masih dapat memprediksi kemungkinan posisinya dalam jangka waktu tertentu. Jika objek tidak muncul kembali setelah batas tertentu, sistem akan menghapus identitas tersebut untuk menghindari akumulasi *error*. Sebaliknya, ketika objek baru muncul, *DeepSORT* membuat track baru secara otomatis. Mekanisme ini menjadikan *DeepSORT* sesuai untuk skenario keramaian, di mana jumlah objek dalam frame dapat berubah-ubah.

Dalam konteks penelitian ini, *DeepSORT* digunakan untuk menjaga kesinambungan identitas wajah antarframe. *Embedding* wajah diperoleh dari *MobileFaceNet*, yang memberikan representasi fitur berukuran kecil namun cukup stabil untuk mengatasi perubahan pose, pencahayaan, dan sudut pandang. Dengan demikian, hasil deteksi dari YOLOv9 tidak hanya menghasilkan posisi wajah, tetapi juga dapat dilacak secara konsisten sehingga sistem memberikan keluaran berupa rangkaian pelacakan wajah yang berkelanjutan. Integrasi antara *DeepSORT* dan MobileFaceNet memastikan bahwa proses pelacakan tetap akurat meskipun objek saling menutupi atau bergerak tidak teratur, menjadikan sistem lebih andal untuk lingkungan keramaian.

Secara keseluruhan, *DeepSORT* menyediakan landasan kuat bagi sistem pelacakan multiobjek dalam penelitian ini. Kombinasi antara model gerakan prediktif dan model penampilan berbasis deep learning memberikan kemampuan bagi sistem untuk mempertahankan identitas wajah dalam kondisi yang berubah-ubah. Dukungan teori dari SORT, perkembangan *DeepSORT* oleh Wojke et al. (2017) dan Susetianingtias, et al. (2023), serta evaluasi integrasi dengan YOLOv9

pada studi terbaru memperkuat relevansi algoritma ini dalam skenario pelacakan wajah pada lingkungan padat.

BAB III

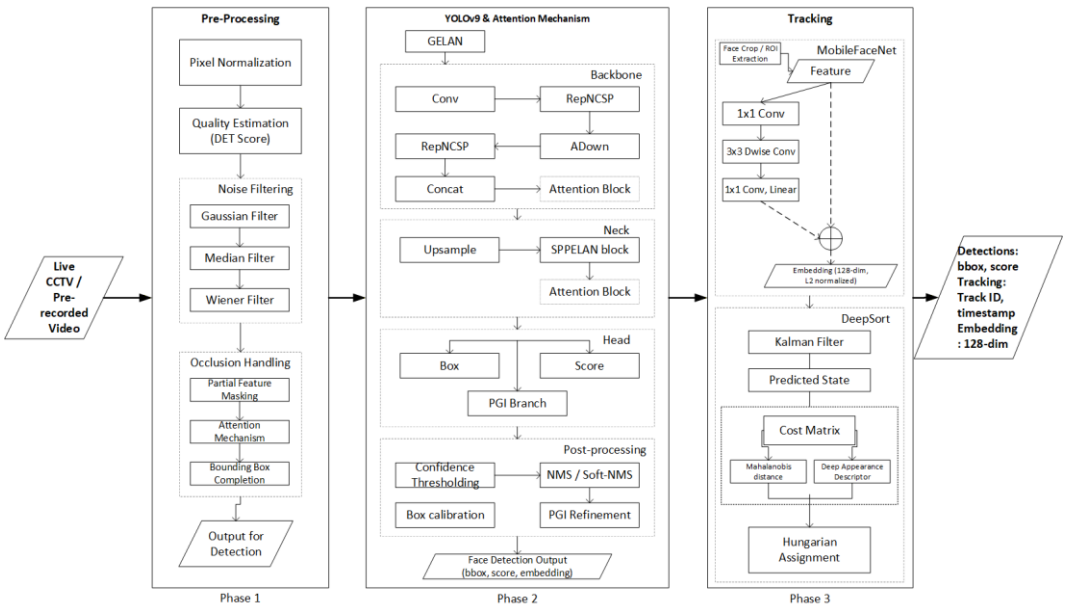
METODOLOGI

Bab ini akan membahas tentang analisis sistem dan perancangan penelitian ini. Beberapa tahapan yang akan di bahas termasuk penelitian yang dilakukan, persyaratan yang akan dibuat , dan penyelesaian masalah untuk pengembangan Deteksi dan Pelacakan Wajah pada Lingkungan Keramaian.

3.1 Desain Sistem

Desain sistem pada Gambar 3.1 menunjukkan alur lengkap sistem yang dibangun untuk mendeteksi dan melacak wajah pada lingkungan dengan tingkat keramaian tinggi. Alur tersebut dibagi ke dalam tiga tahap utama, yaitu Pre-Processing, Deteksi Wajah dengan YOLOv9 dan *Attention Mechanism*, serta *Tracking* yang menggabungkan MobileFaceNet dan *DeepSORT*. Ketiga tahap tersebut bekerja secara berurutan dan saling terhubung untuk menghasilkan sistem yang mampu mendeteksi, mengekstraksi fitur, serta melacak wajah secara konsisten pada setiap *frame* video.

Diagram ini dirancang untuk menunjukkan bagaimana setiap komponen dalam *pipeline* memiliki peran spesifik, mulai dari peningkatan kualitas visual hingga proses asosiasi identitas dalam pelacakan. Struktur desain mengikuti praktik penelitian modern pada deteksi dan pelacakan objek, sebagaimana dijelaskan dalam studi oleh Narayanan et al. (2025), Alsabei et al. (2025), dan Djarah et al. (2024).



Gambar 3.1 Desain Sistem

Tahap 1: *Pre-Processing* (Pra-Pemrosesan dan Penanganan Kualitas Visual)

Tahap ini mempersiapkan data video mentah sebelum diproses oleh model deteksi. Tahap pra-pemrosesan mencakup *pixel normalization*, *quality estimation*, *noise filtering*, serta *occlusion handling*. Seluruh langkah ini bertujuan mempertahankan integritas visual wajah agar detektor dapat bekerja lebih stabil.

Beberapa langkah penting dalam tahap ini meliputi:

1. ***Pixel Normalization*** Dilakukan untuk menstabilkan distribusi nilai piksel antar-frame sehingga model lebih konsisten dalam mempelajari representasi visual.
2. ***Quality Estimation* (*DET Score*)** Mengacu pada pendekatan Singh & Reibman (2024), kualitas gambar diukur berdasarkan tingkat ketajaman, pencahayaan, dan tingkat *occlusion* untuk menentukan kelayakan frame masuk ke tahap deteksi.

3. ***Noise Filtering*** Diagram menunjukkan penggunaan tiga tahap *filtering*: *Gaussian Filter, Median Filter*, dan *Wiener Filter*. Pendekatan berlapis ini diterapkan untuk menyesuaikan jenis *noise* yang ditemukan seperti *Gaussian noise*, *salt-and-pepper*, atau *motion blur*.
4. **Occlusion Handling** Dilakukan melalui *partial feature masking*, *attention-based enhancement*, dan *bounding box completion* sebagaimana didukung oleh Alsabei et al. (2025) dan Alharbey et al. (2022).

Fase ini menjadi fondasi dari pipeline karena deteksi wajah pada kondisi keramaian sangat sensitif terhadap kualitas visual input.

Tahap 2: Deteksi Wajah dengan YOLOv9 dan *Attention Mechanism*

Pada tahap kedua, data hasil pra-pemrosesan dimasukkan ke dalam sistem deteksi berbasis YOLOv9. Diagram menunjukkan struktur lengkap YOLOv9 yang digunakan dalam penelitian, terdiri dari bagian:

- ***Backbone* (*GELAN* dan *Attention Block*)** Berfungsi mengekstraksi fitur dasar dari gambar. *Attention Block* pertama ditempatkan pada bagian akhir Backbone untuk memperkuat fitur penting.
- ***Neck* (*SPPELAN*)** Melakukan *multi-scale feature fusion*. Attention Block kedua ditempatkan di bagian akhir *Neck* untuk mempertajam informasi konteks multi-skala.
- ***Head* (*Box, Score, PGI Branch*)** Memproduksi *bounding box*, skor deteksi, dan mengaktifkan PGI *Branch* (*Programmable Gradient Information*) selama pelatihan.

Penempatan *attention mechanism* pada dua titik penting (*Backbone* dan *Neck*) sejalan dengan temuan Guo et al. (2022) dan Wei et al. (2025) mengenai efektivitas *attention* dalam memperkuat fitur relevan pada pengolahan gambar yang kompleks.

Tahap ini menghasilkan output berupa *face detection output* yang mencakup *bounding box*, *score*, dan *embedding* awal yang akan digunakan pada fase pelacakan.

Tahap 3: Pelacakan (MobileFaceNet dan *DeepSORT*)

Tahap ketiga dalam desain mencakup proses ekstraksi fitur dan pelacakan identitas wajah antar-frame.

A. Ekstraksi Fitur Menggunakan *MobileFaceNet*

MobileFaceNet mengambil *cropped face* berdasarkan bounding box dari YOLOv9. Diagram memperlihatkan struktur *convolutional block MobileFaceNet* yang menyertakan:

- 1×1 Conv
- 3×3 Depthwise Conv
- 1×1 Linear Projection
- L2 Normalization pada embedding

Proses ini menghasilkan embedding berdimensi 128 yang stabil, sesuai temuan Xue (2025).

B. Pelacakan Wajah Menggunakan *DeepSORT*

DeepSORT menerima *bounding box* dan *embedding* dari *MobileFaceNet* untuk melakukan asosiasi identitas. Diagram fase pelacakan menunjukkan alur:

1. ***Kalman Filter***

   Memprediksi *predicted state* wajah berdasarkan frame sebelumnya.

2. ***Cost Matrix***

   Menggabungkan dua metrik:

   - *Mahalanobis distance* (gerakan)
   - *Cosine appearance distance* (embedding wajah)

3. ***Hungarian Assignment***

   Menyelesaikan proses pencocokan *track* dengan deteksi baru.

Fase ini menghasilkan keluaran berupa *Track ID*, lintasan (*trajectory*), dan *bounding box* yang diperbarui setiap *frame*. Pendekatan ini didukung literatur dari Wojke et al. (2017) serta Djarah et al. (2024) yang menunjukkan efektivitas *DeepSORT* pada pelacakan dalam kerumunan.

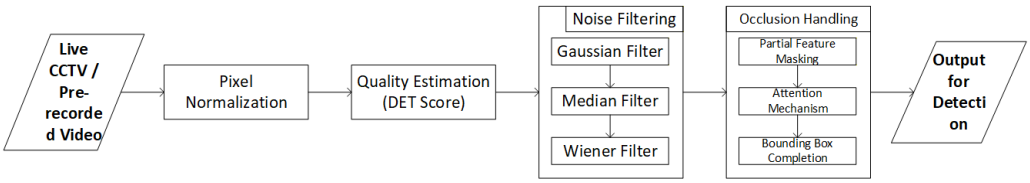
3.1.1 Pra-pemrosesan dan Penanganan Kualitas Gambar & Occlusion

Sebelum data video diproses oleh algoritma deteksi wajah berbasis YOLOv9, diperlukan tahap awal yang berfungsi untuk meningkatkan kualitas input serta menangani permasalahan visual seperti noise, pencahayaan yang buruk, atau wajah yang tertutup sebagian (*occlusion*). Proses ini menjadi sangat penting karena performa model deteksi wajah sangat dipengaruhi oleh kualitas gambar yang masuk. Jika input video memiliki kualitas rendah, maka deteksi akan menurun drastis meskipun model yang digunakan sudah canggih (Narayanan et al., 2025).

Pada kondisi keramaian, kamera pemantau sering merekam wajah dalam situasi visual yang kompleks. Variasi pencahayaan, tingginya pergerakan orang,

serta posisi kepala yang saling menutupi menyebabkan sebagian wajah tidak terlihat secara utuh. Penggunaan masker, topi, atau objek lain di sekitar area pengambilan gambar juga meningkatkan potensi *occlusion*. Pra-pemrosesan diperlukan untuk menstabilkan kualitas gambar sebelum memasuki proses deteksi wajah menggunakan YOLOv9. Langkah ini membantu mempertahankan informasi visual yang relevan pada area wajah sehingga model dapat bekerja dengan input yang lebih konsisten dan sesuai kebutuhan proses pelacakan.

Penelitian-penelitian terdahulu menekankan pentingnya proses ini. Misalnya, Alsabei et al. (2025) menggunakan berbagai teknik augmentasi seperti *rotasi, flipping, scaling*, dan *blurring* untuk meningkatkan representasi data sehingga YOLOv9 dapat bekerja lebih baik di lingkungan nyata. Sementara itu, Singh & Reibman (2024) memperkenalkan konsep *Detectability Score (DET)* untuk mengukur kelayakan suatu gambar diproses lebih lanjut dalam pipeline deteksi wajah. Hasil penelitian Kurniawan et al. (2023) memperlihatkan bahwa peningkatan resolusi gambar dapat meningkatkan ketelitian deteksi tetapi juga menambah waktu proses pelatihan, sehingga tahap normalisasi dan peningkatan kualitas visual perlu dilakukan secara efisien sebelum gambar diproses lebih lanjut oleh model YOLO. Penelitian ini relevan dalam konteks keramaian, di mana sistem dapat mengabaikan gambar yang kualitasnya terlalu rendah sehingga sumber daya komputasi lebih efisien.



Gambar 3.2 Pra-pemrosesan dan Penanganan Kualitas Gambar dan *Occlusion*

Secara umum, proses ini terbagi ke dalam dua bagian utama: (1) pra-pemrosesan video yang mencakup normalisasi piksel, serta (2) penanganan kualitas & *occlusion* yang mencakup estimasi kualitas gambar, filtering noise, dan penanganan wajah tertutup sebagian. Berikut penjelasan teknisnya.

1. Pra-Pemrosesan Video (Normalisasi)

Normalisasi Piksel digunakan untuk menyamakan distribusi nilai intensitas piksel sehingga model lebih stabil saat dilatih atau digunakan.

$$I_{norm}(x,y) = \frac{I(x,y) - \mu}{\sigma} \quad (3.1)$$

Keterangan:

- $I(x,y)$ = nilai piksel asli pada koordinat (x, y).
- μ = rata-rata intensitas gambar.
- σ = standar deviasi intensitas.

Metode ini membantu mengurangi perbedaan pencahayaan antar frame video dan mempercepat konvergensi model.

2. Penanganan Kualitas Gambar & Occlusion

A. *Quality Estimation* (QE)

Untuk mengukur apakah sebuah gambar layak diproses lebih lanjut, Singh & Reibman (2024) memperkenalkan *Detectability Score (DET)*:

$$DET(I) = \alpha \cdot Q_{sharpness}(I) + \beta \cdot Q_{illumination}(I) + \gamma \cdot Q_{occlusion}(I)DET(I) \quad (3.2)$$

- $Q_{sharpness}(I)$ = tingkat ketajaman (misalnya *Laplacian variance*).
- $Q_{illumination}(I)$ = distribusi pencahayaan.
- $Q_{occlusion}(I)$ = proporsi wajah yang tertutup.
- $\alpha, \beta, \gamma\alpha$ = bobot yang ditentukan secara empiris.

Jika $DET(I) < T$ (*threshold*), gambar akan dibuang karena dianggap tidak layak.

B. *Noise Filtering*

Berbagai teknik filtering digunakan untuk meningkatkan kualitas input:

1. ***Gaussian Filter*** untuk Gaussian Noise:

$$I'(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i, y+j) \cdot G(i,j) \tag{3.3}$$

dengan *kernel Gaussian*:

$$G(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}} \tag{3.4}$$

2. ***Median Filter*** untuk *Salt* & *Pepper Noise*:

$$I'(x,y) = \text{median}\{I(x+i, y+j) \mid (i,j) \in \Omega\} \tag{3.5}$$

3. ***Wiener Filter*** untuk *Motion Blur*:

$$I'(u,v) = \frac{H^*(u,v)}{|H(u,v)|^2 + K} \cdot I(u,v) \tag{3.6}$$

dengan $H(u,v)$ = fungsi transfer degradasi, KKK = rasio *noise-to-signal*.

Narayanan et al. (2025) menegaskan bahwa *noise-specific filtering* sangat krusial untuk menjaga performa YOLOv9 di gambar *low-quality*.

C. Occlusion Handling

Wajah di keramaian sering tertutup masker atau objek lain. Beberapa strategi:

- *Partial Feature Masking*: hanya fitur tertentu (mata, hidung) yang digunakan.

- *Attention Mechanism*: memberi bobot lebih pada bagian wajah yang terbuka.
- *Bounding Box Completion*: prediksi bagian wajah hilang dengan CNN.

Alharbey et al. (2022) menambahkan *adaptive* untuk meningkatkan akurasi pada wajah yang terhalang di kerumunan.

3.1.2 Deteksi Wajah Dengan YOLOv9 Dan *Attention Mechanism*

Setelah input video melalui tahap pra-pemrosesan dan penanganan kualitas, gambar yang dihasilkan sudah memiliki kualitas lebih baik serta relatif bebas dari gangguan noise dan masalah occlusion berat. Gambar ini kemudian masuk ke proses selanjutnya, yaitu deteksi wajah.

Dalam sistem ini digunakan YOLOv9, model deteksi objek mutakhir dari keluarga YOLO (*You Only Look Once*), yang memiliki keunggulan dalam hal kecepatan real-time sekaligus akurasi tinggi (Wang et al., 2024; Alsabei et al., 2025). YOLOv9 bekerja dengan cara memproses gambar dalam sekali jalan (*single-stage detector*) menggunakan *backbone* CNN untuk ekstraksi fitur, *neck* untuk penggabungan fitur multi-skala, dan head untuk prediksi bounding box wajah. Studi Kurniawan et al. (2023) menunjukkan bahwa keluarga YOLO memiliki ketahanan yang baik terhadap variasi tampilan wajah dan kondisi lingkungan, sehingga penggunaan YOLOv9 dalam penelitian ini didasarkan pada konsistensi performa YOLO pada berbagai aplikasi deteksi wajah

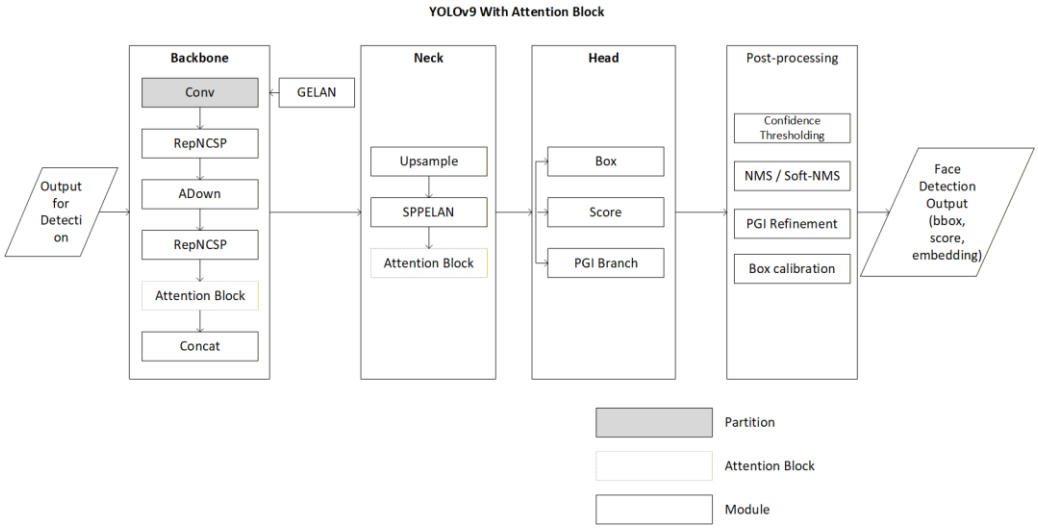
Pada kondisi keramaian, wajah sering terekam dalam keadaan tertutup sebagian, berada pada sudut pandang yang kurang menguntungkan, atau tampak

dengan pencahayaan yang berubah-ubah. *Attention Mechanism* ditambahkan untuk memperkuat proses deteksi pada kondisi seperti ini. Mekanisme perhatian bekerja dengan menekankan area fitur yang paling informatif dan membantu model fokus pada bagian wajah yang masih terlihat dengan jelas. Pendekatan ini mendukung peningkatan ketelitian deteksi karena model dapat mempertahankan sensitivitas terhadap pola wajah meskipun terdapat *occlusion* atau gangguan visual lainnya. Penelitian sebelumnya menunjukkan bahwa penguatan fitur berbasis perhatian efektif pada skenario visual yang kompleks dan dinamis (Alharbey et al., 2022).

1. Desain Arsitektur YOLOv9 dengan *Attention Mechanism*

Arsitektur YOLOv9 (*You Only Look Once Version 9*) merupakan pengembangan model deteksi objek yang menekankan keseimbangan antara kecepatan inferensi dan akurasi deteksi. Dalam penelitian ini, YOLOv9 diperluas dengan penambahan *Attention Block* pada titik-titik strategis jaringan untuk memperkuat kemampuan model dalam mengekstraksi fitur penting serta menjaga kestabilan informasi pada kondisi visual yang kompleks seperti kepadatan tinggi, pencahayaan tidak merata, dan occlusion parsial.

Penambahan berfungsi mengarahkan jaringan agar lebih fokus pada area gambar yang signifikan, sekaligus mengurangi gangguan dari latar belakang yang tidak relevan (Guo et al., 2022). Struktur utama YOLOv9 dengan terdiri atas tiga bagian utama: *Backbone* (*GELAN*), *Neck* (*SPPELAN*), dan *Head* (*Decoupled Head* dan *PGI Branch*).



Gambar 3.3 Arsitektur YOLOv9 dengan *Attention Mechanism*

A. *Backbone* (*GELAN*)

Backbone berfungsi sebagai ekstraktor fitur utama dari gambar masukan. YOLOv9 menggunakan *Generalized Efficient Layer Aggregation Network* (*GELAN*), pengembangan dari *ELAN* yang digunakan pada YOLOv7. GELAN memperkuat efisiensi representasi fitur dengan menggabungkan beberapa jalur konvolusi secara paralel, menjaga kesinambungan informasi, dan meminimalkan kehilangan gradien (Djarah et al., 2024).

Proses *Backbone* dimulai dari *Convolution Layer* untuk ekstraksi fitur dasar, diikuti oleh *RepNCSP Block* yang berfungsi menjaga kontinuitas informasi melalui koneksi residual. Tahap berikutnya adalah *ADown* (*Adaptive Downsampling*) untuk memperluas *receptive field* tanpa kehilangan detail penting. Setelah melalui blok kedua *RepNCSP*, ditambahkan *Attention Block* untuk memperkuat penonjolan fitur yang penting secara kanal dan spasial. Mekanisme ini dapat diimplementasikan dengan modul seperti SE (*Squeeze-and-Excitation*), CBAM (*Convolutional Block Attention Module*), atau ECA (*Efficient Channel Attention*) (Guo et al., 2022).

Tujuan penambahan *Attention Block* di akhir *Backbone* adalah untuk meningkatkan selektivitas fitur sehingga jaringan mampu fokus pada area penting gambar dan menekan informasi yang tidak relevan, yang terbukti meningkatkan performa deteksi pada gambar padat atau dengan noise tinggi (Wang, Li, & Zou, 2023; Thaher et al., 2025).

B. *Neck* (*SPPELAN*)

Neck YOLOv9 berfungsi menggabungkan fitur dari berbagai kedalaman jaringan (*multi-scale feature fusion*). Arsitektur ini menggunakan *SPPELAN* (*Spatial Pyramid Pooling ELAN*), kombinasi antara *Spatial Pyramid Pooling* dan *ELAN structure*. SPPELAN memperluas *receptive field* dengan melakukan *pooling* paralel pada beberapa ukuran kernel, menjaga agar model dapat menangkap konteks spasial yang luas tanpa kehilangan detail penting (Alsabei et al., 2025).

Tahapan *Neck* dimulai dari *Upsampling* atau *Concatenation* untuk menyatukan peta fitur dari *Backbone* dengan resolusi berbeda. Kemudian hasil fusi diproses melalui SPPELAN, menghasilkan fitur multi-skala yang kaya. Di akhir *Neck*, ditambahkan satu *Attention Block* untuk memperkuat bagian gambar yang relevan dan mengurangi gangguan visual.

C. *Head* (*Decoupled Head dan PGI Branch*)

Bagian *Head* berfungsi untuk menghasilkan prediksi akhir berupa posisi, ukuran, dan kelas objek. YOLOv9 menggunakan pendekatan *Decoupled Head*, di mana jalur regresi (*Box Head*) dan klasifikasi (*Score Head*) dipisahkan agar pembelajaran lebih efisien dan stabil. Selain itu, YOLOv9 memiliki cabang

tambahan, yaitu *Programmable Gradient Information* (PGI), yang berperan memperkuat aliran *gradien* selama pelatihan (Narayanan et al., 2025).

Komponen utama *Head* terdiri atas:

1. *Box Head*: memprediksi koordinat dan ukuran *bounding box* (x, y, w, h).
2. *Score Head*: menghitung probabilitas keberadaan objek dan klasifikasi kelas.
3. *PGI Branch*: digunakan hanya saat pelatihan untuk memperkuat stabilitas pembelajaran dan mempercepat konvergensi jaringan.

PGI membantu mengurangi *information bottle*neck dan memperbaiki generalisasi model, sebagaimana ditunjukkan oleh peningkatan *mean Average Precision (mAP)* pada dataset dengan kondisi kompleks (Djarah et al., 2024; Alsabei et al., 2025).

D. Integrasi *Attention Mechanism*

Penambahan *Attention Mechanism* pada YOLOv9 dilakukan di dua titik strategis di akhir *Backbone* dan akhir *Neck* untuk meningkatkan fokus jaringan terhadap fitur penting. Mekanisme ini memberikan tiga dampak utama:

- Peningkatan selektivitas fitur, karena jaringan dapat memprioritaskan area gambar dengan informasi penting.
- Reduksi *noise* visual, dengan menekan respon latar belakang yang tidak relevan.
- Peningkatan pemahaman konteks multi-skala, karena *Attention* membantu menghubungkan objek dari berbagai ukuran dalam satu representasi spasial (Guo et al., 2022; Wang et al., 2023).

Penelitian oleh Gong et al. (2022) menunjukkan bahwa integrasi attention berbasis *Swin Transformer* dan *Normalization-based Attention Module (NAM)* pada deteksi objek kecil menghasilkan peningkatan *mAP* hingga 7.1%. Dengan prinsip serupa, penempatan *Attention Block* pada YOLOv9 terbukti meningkatkan akurasi deteksi tanpa menambah beban komputasi secara signifikan.

Secara umum, proses deteksi pada YOLOv9 dengan *Attention Block* dapat dijabarkan sebagai berikut:

1. Gambar masukan diproses oleh *Backbone* (*GELAN*) untuk menghasilkan peta fitur utama.
2. Fitur hasil ekstraksi difusikan melalui *Neck* (*SPPELAN*) untuk menggabungkan informasi dari berbagai skala.
3. *Attention Block* pada *Neck* mempertegas area penting sebelum diteruskan ke *Head*.
4. Head melakukan prediksi posisi (*Box*), kategori objek (*Score*), serta aktivasi PGI *Branch* selama pelatihan untuk memperbaiki aliran gradien dan kestabilan pembelajaran.

Desain arsitektur YOLOv9 dengan *Attention Block* menghasilkan sistem deteksi yang efisien, adaptif, dan memiliki akurasi tinggi dalam kondisi kompleks. Kombinasi *GELAN*, *SPPELAN*, *Attention Mechanism*, dan PGI memberikan keseimbangan optimal antara efisiensi komputasi dan presisi deteksi. Integrasi ini sangat relevan untuk aplikasi real-time seperti pengenalan wajah di area padat, pemantauan kerumunan, dan pengawasan video (Santoso, 2025; Alsabei et al., 2025).

2. YOLOv9 untuk Deteksi Wajah

YOLOv9 membagi gambar menjadi grid dan memprediksi *bounding box* serta confidence score pada tiap grid. Proses prediksi *bounding box* dilakukan dengan regresi langsung.

A. Prediksi *Bounding Box*

YOLO menggunakan regresi untuk memprediksi koordinat kotak:

$$b_x = \sigma(t_x) + c_x, \quad b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{t_w}, \quad b_h = p_h e^{t_h} \tag{3.7}$$

Keterangan:

- t_x, t_y, t_w, t_h = parameter hasil prediksi jaringan.
- c_x, c_y = koordinat *grid cell*.
- p_w, p_h = *prior anchor box*.
- b_x, b_y, b_w, b_h = koordinat pusat dan ukuran *bounding box* prediksi.

B. Confidence Score

Confidence YOLOv9 menghitung probabilitas adanya wajah dalam bounding box:

$$P_{conf} = P(object) \times IoU_{pred}^{truth} \tag{3.8}$$

dengan IoU (*Intersection over Union*) dihitung:

$$IoU = \frac{|B_{pred} \cap B_{gt}|}{|B_{pred} \cup B_{gt}|} \tag{3.9}$$

YOLOv9 meningkatkan proses ini dengan *hybrid task backbone* (HTB) dan *programmable gradient information* (PGI) untuk mengatasi masalah *overfitting* dan meningkatkan generalisasi.

3. *Attention Mechanism*

Untuk menghadapi occlusion dan kondisi *crowd*, *Attention Mechanism* digunakan bersamaan dengan YOLOv9. Mekanisme ini memungkinkan model memberikan bobot lebih tinggi pada bagian gambar yang paling relevan.

A. Rumus *Self-Attention*

Self-Attention bekerja dengan menghitung *attention score* antar fitur:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.10}$$

Keterangan:

- Q = *Query matrix* (representasi fitur yang dicari).
- K = *Key matrix* (representasi semua fitur).
- V = *Value matrix* (informasi aktual fitur).
- d_k = dimensi dari *vektor Key* (untuk normalisasi).

Dengan mekanisme ini, area wajah yang terlihat jelas (misalnya mata) akan mendapatkan bobot lebih tinggi, sementara bagian yang tertutup (masker, rambut, dll.) diberi bobot rendah.

B. Integrasi *Attention* ke YOLOv9

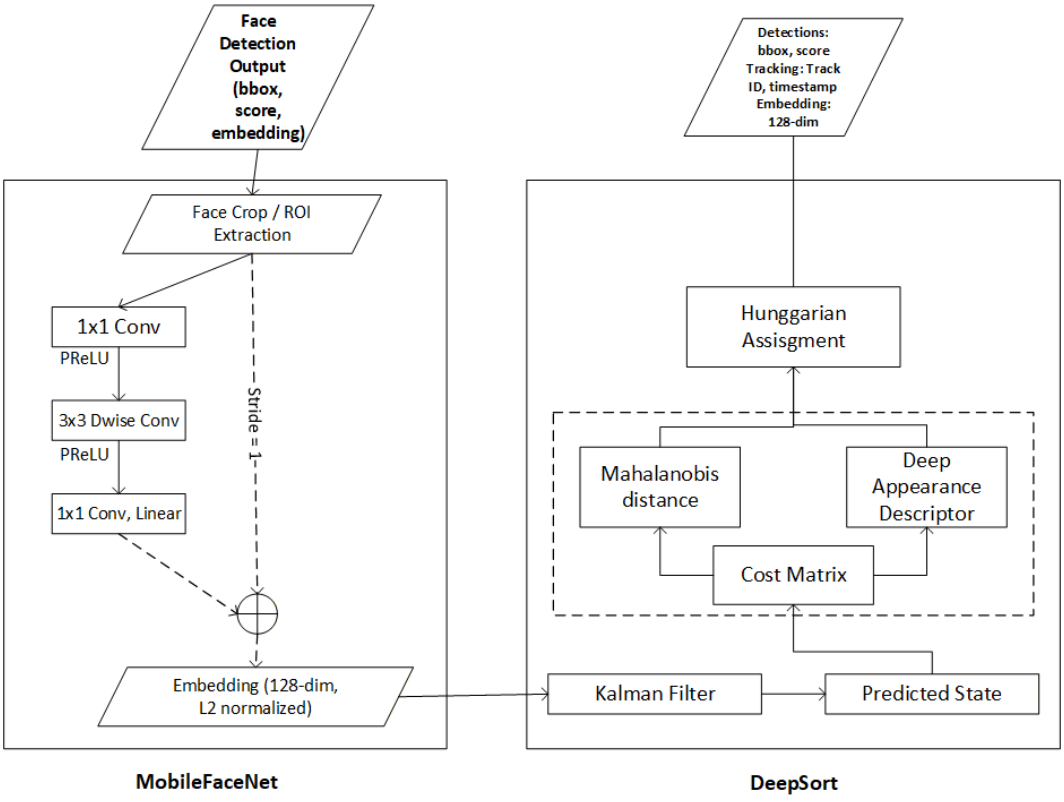
Attention dapat ditambahkan pada:

1. *Backbone* : untuk memperkuat fitur awal.
2. *Neck* (FPN/PAN) : untuk memperbaiki penggabungan multi-skala.
3. *Head* : untuk mengurangi *false detection*.

Penelitian Alharbey et al. (2022) menggunakan *adaptive attention* untuk meningkatkan fokus model pada area target sehingga hasil deteksi di crowd lebih akurat.

3.1.3 Pelacakan (MobileFaceNet dan *DeepSORT*)

Tahap ini bertujuan menghasilkan representasi fitur wajah yang stabil dan melakukan pelacakan identitas secara berkelanjutan pada rangkaian video. Proses dimulai dari hasil deteksi wajah berupa *bounding box* dan nilai kepercayaan, kemudian dilanjutkan dengan dua langkah utama: ekstraksi fitur menggunakan *MobileFaceNet* dan pelacakan identitas menggunakan *DeepSORT*. Diagram kerja lengkap sistem ditampilkan pada Gambar 3.4, yang menunjukkan keterkaitan antara kedua komponen secara terstruktur.



Gambar 3.4 Ekstraksi Fitur dan Pelacakan

Pada tahap ekstraksi fitur, setiap *bounding box* wajah diproses oleh *MobileFaceNet* untuk menghasilkan embedding berdimensi 128 yang telah

dinormalisasi dengan L2. Arsitektur *MobileFaceNet* dibangun dari kombinasi *1×1 convolution*, *depthwise separable convolution*, dan *linear projection*, yang merupakan perluasan dari konsep *MobileNet* dan *MobileNetV2* yang menekankan efisiensi komputasi melalui pengurangan parameter dan operasi konvolusi (Howard et al., 2017; Sandler et al., 2018; Somoal & Dzikrillah, 2025). Pendekatan ini memungkinkan model bekerja secara real-time dengan tetap mempertahankan ketelitian ekstraksi fitur wajah. *MobileFaceNet* telah terbukti efektif dalam sistem pengawasan berbasis video dengan tingkat dinamika visual tinggi, sebagaimana ditunjukkan dalam penelitian pengenalan wajah pada rekaman CCTV oleh Xue (2025) .

Embedding wajah yang dihasilkan *MobileFaceNet* selanjutnya dikirimkan menuju modul *DeepSORT* untuk proses pelacakan. Tahap awal *DeepSORT* memanfaatkan *Kalman Filter* untuk memprediksi posisi wajah pada frame berikutnya berdasarkan pergerakan sebelumnya. Prediksi ini menghasilkan *predicted state* yang mencerminkan koordinat wajah yang diperkirakan, sehingga sistem tetap mampu mempertahankan pelacakan meskipun terjadi perubahan posisi cepat atau penutupan sebagian wajah.

DeepSORT kemudian membentuk sebuah *cost matrix* untuk menentukan kecocokan antara deteksi baru dan prediksi *Kalman Filter*. *Cost matrix* dihasilkan dari penggabungan dua komponen utama:(1) *Mahalanobis distance*, yang menilai kesesuaian hasil deteksi dengan prediksi gerakan, dan (2) *appearance distance*, yang dihitung dari embedding wajah menggunakan *cosine metric*. Kedua metrik ini digabungkan untuk menghasilkan nilai asosiasi yang lebih reliabel, terutama pada

kondisi keramaian di mana objek dapat mengalami overlap. Proses asosiasi ini kemudian diselesaikan melalui *Hungarian Assignment Algorithm*, yang memastikan setiap wajah mendapatkan *Track ID* yang konsisten.

Pendekatan telah digunakan secara luas dalam pelacakan multiobjek, termasuk pada integrasi dengan YOLOv9 untuk lingkungan padat seperti lalu lintas dan kerumunan, sebagaimana ditunjukkan oleh Djarah et al. (2024) . Kinerja *DeepSORT* dalam mempertahankan identitas objek juga menjadi alasan pemilihannya dalam penelitian ini, karena metode ini mampu mengatasi *identity switches* dan kehilangan jejak (*track fragmentation*) yang sering terjadi ketika hanya mengandalkan informasi posisi.

Tahap ekstraksi fitur dan pelacakan ini menghasilkan keluaran berupa *Track ID, trajectory* wajah, dan koordinat *bounding box* yang diperbarui pada setiap frame. Keluaran tersebut menjadi dasar untuk evaluasi sistem pada tahap berikutnya, termasuk pengukuran konsistensi pelacakan, stabilitas identitas, serta ketahanan sistem terhadap kondisi visual yang kompleks pada lingkungan keramaian.

1. Ekstraksi Fitur (*MobileFaceNet*)

Wajah yang telah terdeteksi akan dipotong sesuai dengan *bounding box*, kemudian dimasukkan ke dalam model *MobileFaceNet* untuk diekstraksi menjadi *embedding vektor*. *Embedding* ini adalah representasi numerik berdimensi rendah (misalnya 128 atau 512 dimensi) yang mampu membedakan identitas wajah satu dengan lainnya.

Proses ini dapat dirumuskan sebagai berikut:

$$e = f_\theta(I_{crop}), \quad \hat{e} = \frac{e}{|e|_2} \tag{3.11}$$

dengan f_θ adalah model *MobileFaceNet*, I_{crop} adalah gambar wajah hasil *cropping*, e adalah *embedding*, dan $\hat{e}$ adalah *embedding* ter-normalisasi L2.

Menurut Zhang et al. (2024), *MobileFaceNet* dipilih karena ringan, cepat, dan cocok untuk *deployment real-time*, terutama pada sistem pengawasan publik seperti stasiun. Xue (2025) juga menunjukkan bahwa *MobileFaceNet*, ketika dipadukan dengan RetinaFace, mampu mencapai akurasi tinggi pada video surveillance dengan rata-rata F1-score 91%.

2. Pelacakan Wajah (*DeepSORT*)

Agar identitas wajah tetap konsisten antar-frame, digunakan algoritma pelacakan *DeepSORT*. *DeepSORT* mengombinasikan *Kalman filter* untuk memprediksi posisi wajah di frame berikutnya dengan *Hungarian algorithm* untuk mengasosiasikan deteksi baru dengan *track* yang sudah ada. Selain itu, *DeepSORT* memanfaatkan kemiripan *appearance* (*cosine similarity*) dari *embedding MobileFaceNet* sebagai salah satu metrik asosiasi.

Persamaan utama *Kalman filter* dalam prediksi adalah:

$$x_{k|k-1} = Fx_{k-1|k-1}, \qquad P_{k|k-1} = FP_{k-1|k-1}F^T + Q \tag{3.12}$$

sedangkan proses *update* adalah:

$$x_{k|k} = x_{k|k-1} + K_k(z_k - Hx_{k|k-1}) \tag{3.12}$$

DeepSORT terbukti efektif menjaga identitas objek/wajah pada kondisi occlusion maupun gerakan cepat. Djarah et al. (2024) menunjukkan bahwa

kombinasi YOLOv9, *DeepSORT* dan *optical flow* mampu meningkatkan *robustnes tracking* dengan kenaikan skor MOTA dan IDF1 yang signifikan.

3. *Output*

Tahap ini menghasilkan keluaran yang menggambarkan bagaimana sistem mendeteksi, mengekstraksi fitur, dan melacak wajah pada setiap *frame* video. Setiap wajah yang terdeteksi menerima *Track ID*, yang digunakan oleh algoritma untuk mempertahankan konsistensi pelacakan pada rangkaian *frame*. *DeepSORT* memberikan mekanisme pelacakan yang stabil melalui kombinasi estimasi gerakan dan pemanfaatan fitur visual, sehingga *Track ID* dapat mengikuti objek secara berkelanjutan pada video (Djarah et al., 2024).

Selain *Track ID*, sistem menghasilkan koordinat *bounding box* untuk setiap wajah. Informasi ini berisi posisi kiri atas, kanan bawah, serta nilai *confidence* dari hasil deteksi YOLOv9. Deteksi wajah berbasis YOLOv9 memberikan bounding box dengan ketepatan tinggi pada berbagai kondisi visual melalui struktur arsitektur yang mampu memproses gambar secara efisien (Narayanan et al., 2025).

Tahap ini juga menghasilkan *trajectory* untuk setiap wajah. *Trajectory* menggambarkan rangkaian posisi wajah dari awal hingga akhir kemunculannya dalam video. *DeepSORT* menggunakan prediksi pergerakan melalui *Kalman Filter* untuk menjaga kontinuitas titik posisi, sehingga *trajectory* wajah tersusun secara rapi meskipun terjadi occlusion atau perubahan posisi (Ngeni et al., 2024).

Sistem mencatat durasi kemunculan setiap wajah berdasarkan jumlah *frame* yang dilewati. Informasi ini membantu memahami perilaku pergerakan objek pada

video dan memberikan gambaran tentang dinamika wajah yang tertangkap oleh sistem selama proses pemantauan.

Tahap ini menghasilkan visualisasi beranotasi berupa video yang menampilkan *bounding box* dan *Track ID* pada setiap wajah. Anotasi visual ini mempermudah proses interpretasi hasil serta mendukung analisis kuantitatif maupun kualitatif pada eksperimen. Teknik visualisasi ini umum digunakan pada penelitian pelacakan objek dan membantu mengevaluasi performa sistem dalam kondisi nyata (Alsabei et al., 2025).

Secara keseluruhan, tahap ini menghasilkan data berupa *Track ID*, posisi wajah, *confidence* deteksi, *trajectory*, durasi kemunculan, dan video beranotasi yang memuat proses pelacakan secara menyeluruh.

3.2 Pengumpulan Data

Penelitian ini menggunakan dataset *Head Tracking* 21 (HT21) yang merupakan bagian dari *Crowd of Heads Dataset* (CroHD). Dataset ini diperkenalkan oleh Sundararaman, Braga, Marchand, dan Pettré pada konferensi CVPR 2021 sebagai dataset yang dirancang untuk pelacakan kepala manusia pada lingkungan dengan kepadatan tinggi. CroHD menyajikan anotasi kepala dalam jumlah besar yang direkam dalam berbagai kondisi lingkungan dan tingkat keramaian yang tinggi, sehingga cocok untuk pengembangan dan evaluasi sistem deteksi serta pelacakan wajah dalam kerumunan.

CroHD menghasilkan 2.276.838 anotasi kepala manusia yang tersebar dalam 11.463 frame pada sembilan sekuens video beresolusi Full-HD. Seluruh

sekuens direkam pada laju 25 fps dengan sudut pandang kamera yang lebih tinggi (*elevated viewpoint*). Pendekatan ini memberikan cakupan area yang lebih luas dan memudahkan proses pelacakan objek pada lingkungan padat. Situs perekaman mencakup area dalam ruangan seperti stasiun kereta serta lokasi luar ruangan dengan tingkat keramaian ekstrem seperti *Shibuya Crossing*. Sundararaman et al. (2021) menyatakan bahwa CroHD dibangun untuk memberikan anotasi yang konsisten dan berkualitas tinggi agar dapat digunakan sebagai dasar penelitian pelacakan pada kerumunan berukuran besar.

Dataset HT21 berada dalam kerangka *Multiple Object Tracking* (MOT). MOT merupakan bidang penelitian dalam visi komputer yang berfokus pada pelacakan banyak objek secara simultan dalam urutan video. Sistem MOT bekerja dengan dua komponen utama, yaitu deteksi objek dan asosiasi antar-frame untuk menjaga identitas pelacakan. *Benchmark* MOT sebelumnya berfokus pada pelacakan tubuh penuh, tetapi CroHD memperkenalkan paradigma baru yang berfokus pada pelacakan kepala agar lebih efisien di lingkungan yang memiliki tingkat occlusion tinggi. Fokus pada kepala memberikan stabilitas pelacakan lebih baik pada kondisi padat, sebagaimana disampaikan oleh Sundararaman et al. (2021) dalam pengembangan dataset ini.

Dataset HT21 terdiri atas dua bagian, yaitu training set dan test set. Keduanya menyediakan keragaman visual yang luas, termasuk perbedaan kondisi pencahayaan, waktu perekaman, dan keramaian. Rincian lengkap setiap sekuens ditunjukkan pada tabel berikut.

A. Traning Set

Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Deskripsi
	HT21-04	25	1920×1080	997 (00:40)	580	175,479	176.0	Stasiun kereta luar ruangan
	HT21-03	25	1920×1080	1000 (00:40)	811	257,939	257.9	Persimpangan pejalan kaki
	HT21-02	25	1920×1080	3315 (02:13)	1,276	733,622	221.3	Pintu keluar stadion pada malam hari
	HT21-01	25	1920×1080	429 (00:17)	85	21,456	50.0	Area stasiun kereta dalam ruangan
Total				5741 frm. (230 det)	2752	1188496	207.0	

Tabel 3.1 Datase Head Tracking 21 Traning Set

B Test Set

Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Deskripsi
	HT21-15	25	1920×734	1008 (00:40)	321	149,821	148.6	Jalan pejalan kaki
	HT21-14	25	1920×1080	1050 (00:42)	1,040	258,227	245.9	Stasiun kereta luar ruangan
	HT21-13	25	1920×1080	1000 (00:40)	734	259,603	259.6	Persimpangan pejalan kaki
	HT21-12	25	1920×1080	2080 (01:23)	737	380,647	183.0	Pintu keluar stadion pada malam hari
	HT21-11	25	1920×1080	585 (00:23)	133	38,492	65.8	Area stasiun kereta dalam ruangan
Total				5723 frm. (228 det.)	2965	1086790	189.9	

Tabel 3.2 Datase Head Tracking 21 Test Set

3.3 Desain Eksperimen

Desain eksperimen pada penelitian ini disusun berdasarkan alur data dan proses pemodelan yang ditunjukkan pada diagram sebelumnya, dengan pengujian dilakukan secara bertahap pada setiap komponen utama sistem. Eksperimen dimulai dengan pemisahan dataset *Head Tracking 21* (HT21) ke dalam dua

kelompok utama, yaitu *training set* dan *test set*. Pembagian ini mengikuti struktur asli dataset HT21 yang disusun oleh Sundararaman et al. (2021) pada CroHD, sehingga proses pengujian model tetap sejalan dengan standar evaluasi yang digunakan dalam penelitian pelacakan kepala pada lingkungan dengan kepadatan tinggi.

Selain pembagian data utama tersebut, desain eksperimen ini menerapkan dua skenario pengujian utama, yaitu pengujian berdasarkan variasi rasio data dan pengujian berdasarkan variasi parameter pelatihan model. Pada skenario pertama, pengujian dilakukan menggunakan beberapa konfigurasi rasio data *training* dan *validation*, yaitu 90%-10%, 50%-50%, dan 10%-90%. Setiap konfigurasi diuji secara terpisah dengan alur eksperimen yang sama untuk mengevaluasi pengaruh proporsi data terhadap performa sistem deteksi wajah pada kondisi keramaian tinggi. Pendekatan ini bertujuan untuk mengukur kemampuan generalisasi model ketika jumlah data pelatihan berubah.

Pada skenario kedua, eksperimen dilakukan melalui variasi parameter pelatihan model untuk menganalisis sensitivitas YOLOv9 dengan terhadap perubahan konfigurasi pelatihan. Parameter yang diuji meliputi jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan 16), serta *learning rate* (0.001 dan 0.0005). Variasi parameter dilakukan untuk mengevaluasi pengaruh proses pembelajaran terhadap kestabilan konvergensi model, kemampuan ekstraksi fitur wajah, serta performa deteksi pada lingkungan dengan tingkat *occlusion* dan kepadatan tinggi.

Training set digunakan untuk membangun model deteksi dan pelacakan, sedangkan *test set* digunakan untuk mengukur performa akhir model pada setiap

konfigurasi eksperimen. Pada tahap pelatihan, data kembali dibagi menjadi data *training* dan *validation* sesuai skenario rasio yang diterapkan. Proses validasi digunakan untuk memantau kestabilan model selama pelatihan, mengidentifikasi potensi *overfitting*, serta menentukan kualitas generalisasi model sebelum dilakukan pengujian akhir. Pendekatan validasi seperti ini juga diterapkan pada penelitian pelacakan berbasis YOLO dan *DeepSORT* sebagaimana dijelaskan oleh Djarah et al. (2024).

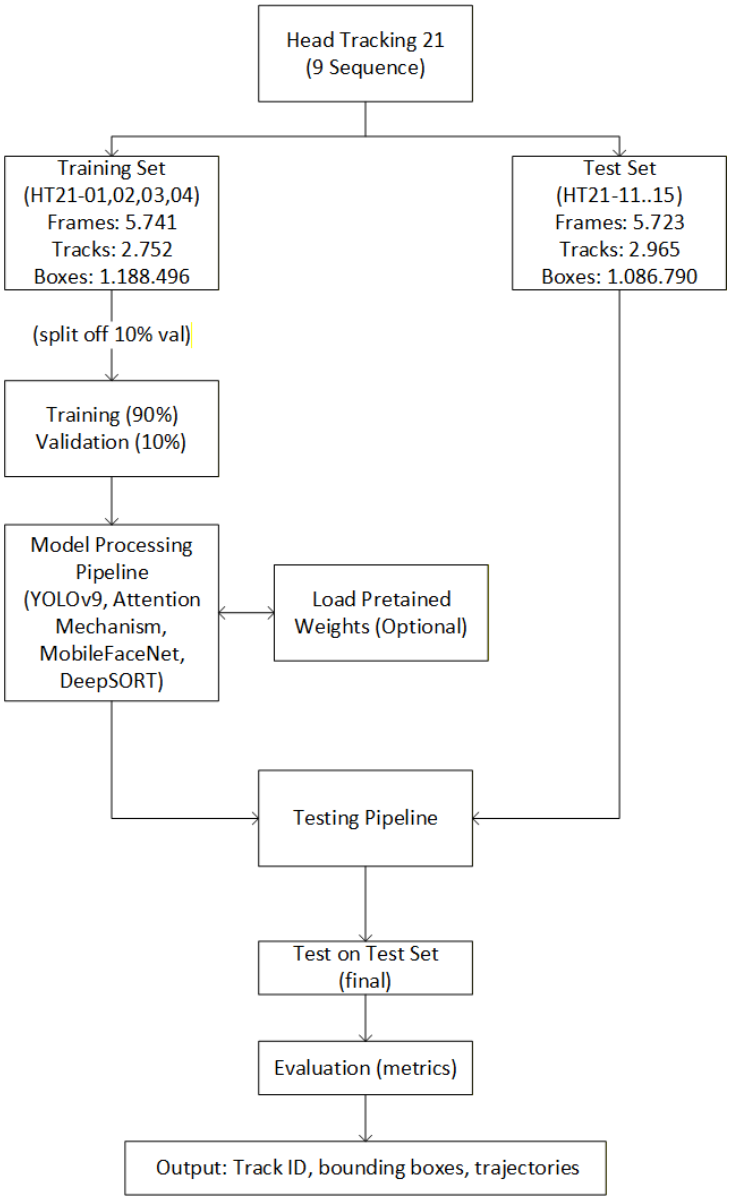
Tahap pertama pengujian dilakukan pada proses pra-pemrosesan dan kualitas masukan gambar. Pada tahap ini, data *training* dan *validation* digunakan untuk memastikan bahwa gambar dengan variasi pencahayaan, kepadatan tinggi, serta kondisi *occlusion* tetap dapat diproses secara stabil sebelum masuk ke tahap deteksi. Hasil pengujian tahap ini digunakan untuk memastikan bahwa informasi visual penting tetap terjaga dan layak digunakan pada tahap pemodelan berikutnya.

Tahap kedua mencakup proses pelatihan dan pengujian deteksi wajah melalui *Model Processing Pipeline* yang terdiri atas komponen deteksi berbasis YOLOv9 dan . *Pipeline* ini menerima input dari data *training* dan *validation* untuk menguji kemampuan model dalam mendeteksi wajah secara akurat pada kondisi keramaian. Pada tahap ini juga diterapkan penggunaan *pretrained weights* sebagai proses opsional untuk mempercepat konvergensi dan meningkatkan stabilitas pelatihan. Selain itu, beberapa konfigurasi parameter pelatihan diterapkan untuk memperoleh model terbaik berdasarkan metrik *precision*, *recall*, *F1-score*, mAP@0.5, serta *validation loss*. Strategi ini sesuai dengan karakteristik YOLOv9

yang mampu memanfaatkan bobot awal dari dataset berskala besar sebagaimana dijelaskan oleh Narayanan et al. (2025).

Tahap ketiga merupakan pengujian pelacakan identitas wajah menggunakan *DeepSORT*. Pada tahap ini, hasil deteksi wajah dari YOLOv9 digunakan sebagai masukan untuk proses pelacakan. Pengujian dilakukan pada skenario pergerakan individu yang dinamis, termasuk kondisi wajah saling menutupi serta keluar-masuk bidang pandang kamera. Evaluasi difokuskan pada kemampuan sistem dalam mempertahankan identitas wajah secara konsisten antar-*frame* video serta meminimalkan terjadinya *identity switching*.

Setelah seluruh tahap diuji secara terpisah, dilakukan *Testing Pipeline* pada *test set* untuk mengevaluasi kinerja sistem secara menyeluruh pada setiap konfigurasi eksperimen. Model dijalankan pada data uji yang tidak terlibat dalam proses pelatihan untuk menghasilkan keluaran berupa *Track ID*, *bounding box*, dan *trajectory* untuk setiap wajah yang terdeteksi. Ketiga keluaran ini menjadi dasar evaluasi kualitas deteksi dan pelacakan wajah pada kondisi nyata dengan tingkat keramaian tinggi sesuai karakteristik utama dataset HT21.



Gambar 3.5 Skema Alur Eksperimen

3.3.1 Dataset dan Pembagian Data

Dataset yang digunakan pada eksperimen ini berasal dari *Head Tracking 21* (HT21), yaitu kumpulan sembilan sekuens video yang disediakan dalam *Crowd of Heads Dataset* (CroHD) oleh Sundararaman et al. (2021). Dataset ini digunakan sesuai dengan struktur pembagian aslinya, yaitu empat sekuens untuk pelatihan dan

lima sekuens untuk pengujian. Pembagian tersebut ditampilkan pada diagram sebelumnya dan menjadi dasar alur eksperimen dalam penelitian ini.

Empat sekuens pada bagian pelatihan, yaitu HT21-01 hingga HT21-04, memiliki total 5.741 *frame*, 2.752 *track*, dan lebih dari satu juta anotasi kepala. Seluruh data pada bagian ini digunakan sebagai sumber utama pembentukan model deteksi wajah berbasis YOLOv9 dengan . Untuk mendukung proses pelatihan dan validasi model, data pada bagian ini dibagi kembali sesuai dengan skenario eksperimen yang diterapkan, yaitu rasio 90%-10%, 50%-50%, dan 10%-90% antara data *training* dan *validation*. Variasi pembagian ini digunakan untuk mengevaluasi pengaruh proporsi data pelatihan terhadap kestabilan dan kemampuan generalisasi model pada lingkungan keramaian dengan tingkat kepadatan tinggi.

Selain variasi rasio data, eksperimen juga dilakukan melalui perubahan parameter pelatihan model, meliputi jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan 16), serta *learning rate* (0.001 dan 0.0005). Pendekatan ini diterapkan untuk mengamati sensitivitas model terhadap perubahan konfigurasi pelatihan serta menentukan parameter yang menghasilkan performa deteksi wajah paling optimal. Proses validasi dilakukan pada setiap konfigurasi untuk memantau kestabilan model selama pelatihan, mengevaluasi proses konvergensi, dan meminimalkan potensi *overfitting*. Pendekatan validasi seperti ini juga diterapkan pada penelitian pelacakan multiobjek berbasis YOLO dan *DeepSORT* sebagaimana dijelaskan oleh Djarah et al. (2024).

Lima sekuens lainnya, yaitu HT21-11 hingga HT21-15, digunakan sebagai *test set*. Bagian ini terdiri dari 5.723 *frame* dan 2.965 *track* yang digunakan secara

eksklusif untuk proses evaluasi akhir. Data pada *test set* tidak mengalami augmentasi maupun perubahan struktur agar tetap merepresentasikan kondisi asli lingkungan keramaian sebagaimana didefinisikan dalam publikasi HT21. Dataset CroHD dirancang untuk menampilkan variasi kepadatan tinggi, perubahan pencahayaan, tingkat *occlusion*, serta dinamika pergerakan pada lingkungan urban, sehingga sesuai digunakan sebagai skenario uji untuk sistem deteksi dan pelacakan wajah.

Dalam skema eksperimen penelitian ini, struktur dasar pembagian sekuens HT21 tetap dipertahankan untuk menjaga konsistensi evaluasi terhadap penelitian asli. Variasi rasio data diterapkan pada pembagian internal data pelatihan dan validasi, sedangkan *test set* tetap digunakan secara independen untuk seluruh skenario pengujian. Dengan pendekatan ini, pengaruh jumlah data pelatihan dan konfigurasi parameter model terhadap performa sistem dapat dianalisis secara sistematis tanpa mengorbankan objektivitas evaluasi.

Secara keseluruhan, pembagian dataset pada penelitian ini memastikan bahwa data *training* dan *validation* digunakan dalam *Model Processing Pipeline*, sedangkan *test set* digunakan secara terpisah dalam *Testing Pipeline*. Rancangan ini menjamin bahwa model dievaluasi pada data yang benar-benar independen dari proses pelatihan, sehingga hasil pengujian dapat memberikan gambaran performa sistem deteksi dan pelacakan wajah pada kondisi keramaian yang kompleks dan mendekati skenario nyata.

3.3.2 Proses Pelatihan Model

Proses pelatihan model pada penelitian ini mengikuti alur *Model Processing Pipeline* sebagaimana ditunjukkan pada diagram eksperimen. Pelatihan dilakukan setelah data pada *training* dan *validation set* melalui tahap pra-pemrosesan. Seluruh proses disusun secara berurutan agar model mampu mempelajari pola deteksi wajah serta mempertahankan kestabilan pelacakan pada setiap *frame* video dalam kondisi keramaian padat.

Pada penelitian ini, proses pelatihan dilakukan melalui dua skenario eksperimen, yaitu pengujian multi-rasio data serta pengujian variasi parameter pelatihan model. Pada skenario pertama, pelatihan dilakukan secara berulang untuk setiap konfigurasi pembagian data *training*-*validation*, yaitu 90%-10%, 50%-50%, dan 10%-90%. Variasi rasio ini diterapkan untuk mengevaluasi pengaruh jumlah data pelatihan terhadap kemampuan generalisasi model pada kondisi lingkungan dengan tingkat kepadatan tinggi dan *occlusion* yang kompleks.

Selain variasi rasio data, eksperimen juga dilakukan dengan mengubah beberapa parameter pelatihan model, meliputi jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan 16), serta *learning rate* (0.001 dan 0.0005). Variasi parameter dilakukan untuk menganalisis sensitivitas YOLOv9 dengan terhadap perubahan konfigurasi pelatihan, serta menentukan kombinasi parameter yang menghasilkan performa deteksi paling optimal. Pada setiap konfigurasi eksperimen, proses pelatihan dilakukan menggunakan alur yang sama agar perbedaan performa yang dihasilkan dapat dianalisis secara objektif berdasarkan perubahan parameter yang diterapkan.

Tahap awal pelatihan dimulai dengan inisialisasi model deteksi wajah berbasis YOLOv9 yang dilengkapi dengan . YOLOv9 digunakan sebagai komponen utama deteksi karena struktur backbone, neck, dan *head* yang dimilikinya mampu menangani variasi visual yang kompleks, seperti kepadatan tinggi, perubahan sudut pandang, kondisi *occlusion*, serta variasi pencahayaan. Kemampuan tersebut didukung oleh penelitian Narayanan et al. (2025) yang menunjukkan bahwa YOLOv9 memiliki performa deteksi yang kuat pada lingkungan dengan dinamika visual tinggi.

Pada tahap inisialisasi, model dimuat menggunakan *pretrained weights* sebagai bobot awal. Selanjutnya, proses *fine-tuning* diterapkan untuk menyesuaikan model dengan karakteristik dataset HT21 yang memiliki tingkat kepadatan tinggi dan ukuran wajah relatif kecil. *Fine-tuning* dilakukan dengan melatih ulang sebagian parameter model, khususnya pada bagian neck, *head*, dan modul *attention*, sementara lapisan awal pada backbone dapat dipertahankan untuk menjaga representasi fitur dasar. Strategi ini memungkinkan model beradaptasi secara spesifik terhadap data target tanpa kehilangan kemampuan generalisasi dari pelatihan sebelumnya.

Setelah proses inisialisasi dan pengaturan *fine-tuning*, pelatihan dilanjutkan melalui *training loop*. Setiap *batch* data diproses melalui *forward pass* untuk menghasilkan prediksi awal berupa *bounding box* wajah dan nilai *confidence*. Prediksi tersebut dievaluasi menggunakan *loss function* untuk mengukur selisih antara keluaran model dengan anotasi *ground truth*. Nilai *loss* selanjutnya digunakan oleh *optimizer* untuk memperbarui parameter model secara iteratif

hingga seluruh data pelatihan diproses pada setiap *epoch*. Evaluasi pada *validation set* dilakukan secara periodik untuk memantau kestabilan konvergensi model, mengidentifikasi potensi *overfitting*, serta mengevaluasi perubahan performa pada setiap konfigurasi parameter pelatihan.

Tahap selanjutnya adalah ekstraksi fitur wajah menggunakan MobileFaceNet. Setiap wajah yang terdeteksi diproses untuk menghasilkan *embedding* sebagai representasi numerik fitur wajah. *Embedding* ini berfungsi sebagai dasar asosiasi identitas antar-*frame* dan menjadi masukan utama bagi modul pelacakan.

Tahap terakhir dalam proses pelatihan melibatkan *DeepSORT* sebagai modul pelacakan identitas. *DeepSORT* menggabungkan informasi *embedding* wajah dengan prediksi pergerakan objek yang dihasilkan oleh *Kalman Filter* untuk membentuk identitas pelacakan yang stabil. Pendekatan ini memungkinkan sistem mempertahankan identitas wajah meskipun terjadi pergerakan dinamis, *occlusion*, maupun perubahan posisi antar-*frame*. Strategi pelacakan ini telah digunakan secara luas dalam penelitian pelacakan multiobjek, termasuk studi oleh Djarah et al. (2024).

Secara keseluruhan, proses pelatihan menghasilkan beberapa model hasil *fine-tuning* berdasarkan variasi rasio data dan konfigurasi parameter pelatihan. Setiap model selanjutnya digunakan pada tahap pengujian untuk mengevaluasi performa sistem pada *test set* dan menentukan konfigurasi terbaik berdasarkan metrik *precision*, *recall*, *F1-score*, mAP@0.5, serta kestabilan *validation loss*.

3.3.3 Strategi Pengujian dan Evaluasi

Strategi pengujian dan evaluasi pada penelitian ini dirancang untuk menilai kinerja sistem deteksi dan pelacakan wajah secara menyeluruh pada setiap tahap pemrosesan, yaitu tahap pra-pemrosesan gambar, tahap deteksi wajah, dan tahap pelacakan identitas. Proses pengujian dilakukan setelah model menyelesaikan tahap pelatihan dan validasi. Seluruh tahapan pengujian mengikuti alur pada *Testing Pipeline* sebagaimana ditunjukkan pada diagram eksperimen, yang dimulai dari pemrosesan *test set* hingga menghasilkan metrik evaluasi dan keluaran akhir sistem.

Pengujian dalam penelitian ini dilakukan melalui dua skenario eksperimen utama, yaitu pengujian berbasis variasi rasio data dan pengujian berbasis variasi parameter pelatihan model. Pada skenario pertama, proses pengujian dilakukan secara berulang untuk setiap konfigurasi pembagian data *training-validation*, yaitu 90%-10%, 50%-50%, dan 10%-90%. Tujuan pengujian ini adalah untuk mengevaluasi pengaruh jumlah data pelatihan terhadap kemampuan generalisasi model pada kondisi keramaian dengan tingkat kepadatan tinggi.

Pada skenario kedua, pengujian dilakukan terhadap beberapa konfigurasi parameter pelatihan model, meliputi jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan 16), serta *learning rate* (0.001 dan 0.0005). Setiap model hasil pelatihan diuji menggunakan skenario evaluasi yang sama untuk menganalisis pengaruh konfigurasi pelatihan terhadap kestabilan konvergensi model, kemampuan deteksi wajah, serta performa pelacakan identitas pada lingkungan yang kompleks. Dengan pendekatan ini, perbedaan performa sistem dapat dianalisis secara lebih objektif berdasarkan perubahan parameter yang diterapkan.

Pengujian dilakukan dengan menjalankan model pada seluruh sekuens *test set* HT21 yang memuat variasi kondisi lingkungan, tingkat kepadatan tinggi, perubahan pencahayaan, serta dinamika pergerakan individu yang beragam. Dataset HT21 yang diperkenalkan oleh Sundararaman et al. (2021) dirancang secara khusus untuk skenario keramaian, sehingga sesuai digunakan sebagai dasar evaluasi sistem deteksi dan pelacakan wajah pada kondisi yang mendekati situasi nyata.

Pada tahap pra-pemrosesan, evaluasi difokuskan pada kualitas gambar masukan sebelum dan sesudah proses pra-pemrosesan. Penilaian dilakukan menggunakan metrik kualitas gambar *Peak Signal-to-Noise Ratio* (PSNR) dan *Structural Similarity Index Measure* (SSIM) untuk mengukur peningkatan kualitas visual serta menjaga kesesuaian struktur wajah. Penggunaan PSNR dan SSIM sebagai indikator kualitas gambar hasil pra-pemrosesan telah banyak diterapkan dalam penelitian *image enhancement* dan rekonstruksi wajah karena kedua metrik tersebut saling melengkapi dalam merepresentasikan kesamaan piksel dan preservasi struktur visual gambar (Salman & Kalakech, 2024; Tommy et al., 2025). Selain itu, stabilitas hasil pra-pemrosesan diamati melalui konsistensi gambar wajah yang dapat diproses dengan baik sebelum memasuki tahap deteksi.

Pada tahap deteksi wajah, proses pengujian menghasilkan keluaran berupa *bounding box* wajah beserta *confidence score*. Kualitas deteksi dievaluasi menggunakan metrik *Precision*, *Recall*, *F1-score*, dan *mean Average Precision* (mAP@0.5) untuk menilai ketepatan, kelengkapan, dan kestabilan deteksi wajah pada setiap *frame* video. Selain itu, evaluasi juga dilakukan terhadap *validation loss*

dan *confusion matrix* untuk menganalisis tingkat kesalahan klasifikasi serta kemampuan model dalam membedakan objek wajah dan non-wajah. Penelitian Narayanan et al. (2025) menunjukkan bahwa akurasi lokasi *bounding box*, nilai *confidence*, serta kestabilan mAP merupakan indikator penting dalam mengevaluasi performa YOLOv9 pada lingkungan visual yang kompleks.

Pada tahap pelacakan identitas, sistem menghasilkan keluaran berupa *embedding* wajah, *Track ID*, dan *trajectory* pada setiap objek. Evaluasi pelacakan difokuskan pada konsistensi identitas antar-*frame*, stabilitas lintasan, serta kemampuan sistem mempertahankan identitas objek meskipun terjadi *occlusion* atau perubahan pergerakan. Penilaian dilakukan menggunakan metrik pelacakan standar seperti IDF1, MOTA, dan jumlah *ID Switches*, sebagaimana lazim digunakan dalam penelitian pelacakan multiobjek berbasis *DeepSORT* (Djarah et al., 2024).

Hasil pengujian dari setiap konfigurasi rasio data dan parameter pelatihan selanjutnya digunakan untuk menilai kestabilan kinerja sistem secara keseluruhan. Evaluasi dilakukan terhadap seluruh rangkaian proses pada *test set*, sehingga hasil yang diperoleh mencerminkan performa akhir sistem dalam mendeteksi dan melacak wajah pada situasi keramaian padat. Seluruh metrik evaluasi yang digunakan pada setiap tahap dirangkum pada Tabel 3.3 sebagai dasar analisis performa sistem pada bab berikutnya.

Tahap	Kategori	Nama Metrik	Deskripsi	Kegunaan
Tahap 1	Pra-pemrosesan & Kualitas Gambar	PSNR	Mengukur rasio sinyal terhadap noise pada gambar hasil pra-pemrosesan.	Menilai peningkatan kualitas gambar setelah reduksi noise.
		SSIM	Mengukur kesamaan struktur gambar sebelum dan sesudah pra-pemrosesan.	Menilai sejauh mana struktur wajah tetap terjaga.
Tahap 2	Deteksi Wajah	Precision	Rasio deteksi benar terhadap seluruh prediksi.	Mengukur ketepatan bounding box wajah.
		Recall	Rasio deteksi benar terhadap seluruh wajah yang seharusnya terdeteksi.	Menilai kelengkapan deteksi wajah.

Tahap	Kategori	Nama Metrik	Deskripsi	Kegunaan
		F1-Score	Harmonisasi precision dan recall	Menilai keseimbangan performa
		Confidence Score	Nilai keyakinan model terhadap hasil deteksi.	Menyaring deteksi wajah yang relevan.
		mAP@0.5	Rata-rata akurasi deteksi objek model	Mengukur performa model deteksi objek
		Validation Loss	Tingkat kesalahan model selama validasi	Menilai kestabilan konvergensi
Tahap 3	Pelacakan Wajah	IDF1	Skor F1 untuk konsistensi identitas pelacakan.	Menilai kestabilan identitas wajah.
		MOTA	Akurasi pelacakan multiobjek.	Menilai performa pelacakan secara umum.

Tahap	Kategori	Nama Metrik	Deskripsi	Kegunaan
		ID Switches	Jumlah pergantian identitas.	Mengukur kesalahan pelacakan identitas.
Runtime	Efisiensi Sistem	FPS	Kecepatan pemrosesan per frame.	Menilai kelayakan real-time.

Tabel 3.3 Metrik Evaluasi

3.3.4 Parameter Eksperimen dan Lingkungan Komputasi

Konfigurasi parameter eksperimen dan lingkungan komputasi memegang peranan penting dalam keberhasilan proses pelatihan dan pengujian sistem pada penelitian ini. Seluruh eksperimen disusun secara terstruktur agar setiap komponen model, mulai dari YOLOv9, , MobileFaceNet, hingga *DeepSORT*, dapat berjalan secara stabil pada kondisi visual yang kompleks. Arsitektur YOLOv9 digunakan sebagai inti proses deteksi wajah karena kemampuannya dalam menangani variasi kepadatan, perubahan pencahayaan, serta kondisi *occlusion*, sebagaimana dijelaskan oleh Narayanan et al. (2025).

Berbeda dengan pendekatan eksperimen tetap (*fixed parameter setting*), penelitian ini menerapkan dua skenario konfigurasi eksperimen, yaitu variasi rasio data pelatihan dan variasi parameter pelatihan model. Variasi rasio dilakukan pada

konfigurasi 90%-10%, 50%-50%, dan 10%-90% antara data *training* dan *validation* untuk mengevaluasi pengaruh jumlah data terhadap kemampuan generalisasi model. Selain itu, dilakukan pula pengujian sensitivitas parameter pelatihan yang meliputi jumlah *epoch*, *batch size*, dan *learning rate* untuk menentukan konfigurasi yang menghasilkan performa deteksi wajah paling optimal.

Parameter pelatihan dirancang dengan mempertimbangkan kestabilan pembaruan bobot model, kapasitas memori GPU, serta efisiensi proses pelatihan pada lingkungan komputasi berbasis *cloud*. Variasi *epoch* diterapkan pada nilai 50, 100, 200 dan 300 untuk mengevaluasi tingkat konvergensi model. *Batch size* diuji pada nilai 8 dan 16 untuk menilai pengaruh ukuran *batch* terhadap kestabilan gradien dan performa deteksi. Sementara itu, *learning rate* diuji pada nilai 0.001 dan 0.0005 guna menganalisis sensitivitas pembaruan parameter model selama proses *fine-tuning*. Penentuan *confidence threshold*, *IoU threshold*, *optimizer*, dan *weight decay* mengacu pada konfigurasi umum yang digunakan pada model YOLO-series.

Mekanisme perhatian () diterapkan pada bagian backbone dan neck model untuk memperkuat fitur spasial penting dan menekan gangguan visual dari latar belakang. Pendekatan ini digunakan untuk meningkatkan kemampuan deteksi wajah pada kondisi kepadatan tinggi, perubahan skala, dan *occlusion*, sebagaimana ditunjukkan oleh Wei et al. (2025) dalam penelitian penguatan fitur berbasis konteks.

Tahap ekstraksi fitur wajah dilakukan menggunakan MobileFaceNet yang menghasilkan *embedding* berdimensi rendah sebagai representasi numerik wajah.

MobileFaceNet dipilih karena kemampuannya menghasilkan representasi wajah yang efisien dengan kompleksitas komputasi rendah sehingga sesuai untuk aplikasi *real-time*, sebagaimana dilaporkan oleh Xue (2025) serta Zhang et al. (2024).

Proses pelacakan identitas dilakukan menggunakan algoritma *DeepSORT* yang memadukan prediksi pergerakan berbasis *Kalman Filter* dan pengukuran kesamaan fitur menggunakan *cosine distance*. Pendekatan ini terbukti efektif dalam mempertahankan identitas objek pada kondisi keramaian dan pergerakan dinamis sebagaimana dijelaskan oleh Djarah et al. (2024).

Seluruh eksperimen dijalankan pada platform Google Colab yang menyediakan dukungan GPU berbasis CUDA. Lingkungan ini mendukung penggunaan GPU seperti NVIDIA T4 dan A100 sehingga memungkinkan pemrosesan dataset HT21 beresolusi tinggi dilakukan secara efisien. Ketersediaan pustaka seperti PyTorch, OpenCV, dan CUDA Toolkit pada lingkungan Colab mendukung proses pelatihan, visualisasi hasil, serta pencatatan eksperimen secara sistematis. Rangkuman parameter eksperimen dan konfigurasi lingkungan komputasi disajikan pada Tabel 3.4 untuk memastikan reprodusibilitas penelitian.

Kategori	**Parameter**	**Konfigurasi**	**Sumber / Referensi**
Parameter Pelatihan	Batch Size	8, 16	Disesuaikan kebutuhan GPU
	Epoch	50, 100, 200, 300	Eksperimen YOLOv9 (Narayanan et al., 2025)
	Learning Rate	0.001, 0,0005	Konfigurasi umum YOLO-series

Kategori	Parameter	Konfigurasi	Sumber / Referensi
	Optimizer	Adam	Literatur CV pelatihan deteksi
	Loss Function	YOLO Loss	Arsitektur YOLOv9 (Narayanan et al., 2025)
	Confidence Threshold	0.25	Praktik inferensi YOLO-series
	IoU Threshold	0.45	Praktik evaluasi deteksi objek
YOLOv9 Configuration	Input Size	640 × 640	Digunakan pada eksperimen YOLOv9 (Narayanan et al., 2025)
	Attention Mechanism	Aktif	Sesuai rancangan penelitian
	Pretrained Weights	Aktif (*Fine-tuning*)	*Transfer learning* umum
MobileFaceNet	Embedding Size	128-dim	Xue (2025) - MobileFaceNet pada pengenalan wajah
DeepSORT	Metric	Cosine Distance	Konfigurasi *DeepSORT* (Djarah et al., 2024)
	Motion Model	Kalman Filter	Standar *DeepSORT*

Kategori	Parameter	Konfigurasi	Sumber / Referensi
	Max Age	30	Praktik pelacakan
	Max Cosine Distance	0.2	*DeepSORT* (Djarah et al., 2024)
Lingkungan Komputasi	Platform	Google Colab	Lingkungan eksperimen
	GPU	NVIDIA A100 VRAM 40GB	Konfigurasi Colab
	Framework	PyTorch	Library pendukung YOLOv9
	CUDA	Versi otomatis sesuai GPU	Lingkungan Colab
	RAM	80 GB	Lingkungan Colab
	OS	Linux backend	Default Google Colab

Tabel 3.4 Parameter Eksperimen dan Lingkungan Komputasi

3.4 Tahapan Implementasi Sistem

Tahapan implementasi sistem menggambarkan proses teknis yang dilakukan untuk menjalankan keseluruhan *pipeline* deteksi dan pelacakan wajah sesuai dengan rancangan metodologi penelitian. Implementasi ini disusun mengikuti tiga tahap utama sistem, yaitu pra-pemrosesan data, deteksi dan ekstraksi fitur wajah, serta pelacakan identitas, hingga menghasilkan keluaran berupa *bounding box*, *Track ID*, dan *trajectory* wajah pada setiap *frame* video.

Tahap awal implementasi dimulai dari persiapan data. Setiap sekuens video pada *training set*, *validation set*, dan *test set* diekstraksi menjadi rangkaian *frame* agar dapat diproses oleh model. Struktur folder dan format data disusun agar kompatibel dengan kebutuhan *pipeline* pelatihan dan pengujian. Pada tahap ini dilakukan pembacaan file anotasi, pemetaan identitas awal, serta penyesuaian format data sehingga seluruh *frame* dan anotasi siap digunakan pada proses pemodelan. Implementasi dilakukan berdasarkan skenario pembagian data yang telah ditentukan, yaitu rasio 90%-10%, 50%-50%, dan 10%-90% antara data *training* dan *validation*.

Tahap pertama implementasi adalah pra-pemrosesan data, yang bertujuan untuk meningkatkan kualitas dan konsistensi gambar masukan. Setiap gambar mengalami proses penyesuaian ukuran (*resizing*), normalisasi nilai piksel, serta penerapan augmentasi ringan pada data pelatihan. Proses ini dilakukan untuk menjaga keseragaman ukuran masukan model dan mendukung kemampuan sistem dalam menghadapi variasi kondisi visual, seperti perubahan pencahayaan, tingkat kepadatan tinggi, dan kondisi *occlusion*. Data hasil pra-pemrosesan kemudian dikelompokkan ke dalam *batch* sesuai konfigurasi eksperimen agar proses pelatihan dan inferensi dapat berjalan secara efisien.

Tahap kedua implementasi mencakup proses deteksi wajah menggunakan YOLOv9 yang dilengkapi dengan . Model deteksi dimuat ke dalam memori menggunakan *pretrained weights* sebagai bobot awal untuk mendukung proses *fine-tuning*. Pada tahap pelatihan, eksperimen dilakukan melalui beberapa konfigurasi parameter, yaitu jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan

16), serta *learning rate* (0.001 dan 0.0005). Setiap konfigurasi dilatih secara terpisah untuk mengevaluasi pengaruh parameter pelatihan terhadap kemampuan model dalam mendeteksi wajah pada kondisi keramaian padat.

Pada proses pelatihan, setiap *batch* data diproses melalui *forward pass* untuk menghasilkan prediksi awal berupa *bounding box* wajah dan nilai *confidence*. Selanjutnya, *loss function* digunakan untuk mengukur selisih antara prediksi model dengan anotasi *ground truth*, kemudian hasilnya digunakan oleh *optimizer* untuk memperbarui bobot model secara iteratif hingga mencapai kondisi konvergen. Pada tahap inferensi, model menghasilkan deteksi wajah yang digunakan sebagai masukan untuk tahap berikutnya. Setiap konfigurasi model dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, mAP@0.5, serta *validation loss* untuk menentukan model terbaik.

Tahap ketiga implementasi adalah ekstraksi fitur wajah dan pelacakan identitas. Setiap wajah yang terdeteksi diproses menggunakan MobileFaceNet untuk menghasilkan *embedding* sebagai representasi numerik fitur wajah. *Embedding* ini digunakan sebagai dasar asosiasi identitas antar-*frame*. Proses pelacakan kemudian dilakukan menggunakan *DeepSORT*, yang menggabungkan prediksi pergerakan objek melalui *Kalman Filter* dan pengukuran kesamaan fitur menggunakan *cosine distance*. Evaluasi pelacakan dilakukan menggunakan model terbaik hasil tahap deteksi untuk menjaga konsistensi performa sistem pada seluruh skenario pengujian.

Berdasarkan hasil pelacakan, *trajectory* wajah disusun dari urutan posisi *bounding box* pada setiap *frame*. Lintasan ini memberikan gambaran pergerakan

wajah sepanjang durasi video dan digunakan sebagai salah satu indikator kestabilan pelacakan pada lingkungan keramaian dengan dinamika objek tinggi.

Tahap akhir implementasi adalah penyusunan dan penyimpanan keluaran sistem. Seluruh hasil deteksi dan pelacakan disimpan dalam format terstruktur yang mencakup *bounding box*, nilai *confidence*, *Track ID*, dan *trajectory*. Keluaran ini digunakan sebagai masukan utama pada tahap pengujian dan evaluasi performa sistem berdasarkan skenario variasi rasio data dan parameter pelatihan yang telah ditetapkan sebelumnya.

Secara keseluruhan, tahapan implementasi ini memastikan bahwa seluruh komponen sistem dijalankan sesuai urutan rancangan dan menghasilkan keluaran yang dibutuhkan untuk analisis performa deteksi dan pelacakan wajah pada lingkungan keramaian yang kompleks.

3.5 Kriteria Evaluasi Sistem

Kriteria evaluasi sistem disusun untuk memastikan bahwa model mampu menjalankan seluruh tahapan pemrosesan, mulai dari pra-pemrosesan gambar, deteksi wajah, hingga pelacakan identitas wajah secara akurat, stabil, dan efisien pada lingkungan dengan tingkat keramaian tinggi. Evaluasi dilakukan berdasarkan keluaran sistem yang meliputi kualitas gambar hasil pra-pemrosesan, *bounding box* wajah, *embedding*, *Track ID*, serta lintasan (*trajectory*) pada setiap rangkaian *frame*. Selain mengevaluasi performa deteksi dan pelacakan, penelitian ini juga menganalisis pengaruh variasi parameter pelatihan model terhadap kestabilan dan kemampuan generalisasi sistem. Dengan menggunakan kriteria evaluasi yang

terukur, penelitian ini dapat menilai kinerja setiap tahap sistem secara terpisah maupun menyeluruh pada data uji HT21.

1. Evaluasi Pra-pemrosesan dan Kualitas Gambar

Tahap pertama evaluasi difokuskan pada efektivitas proses pra-pemrosesan dalam meningkatkan kualitas gambar masukan sebelum dilakukan deteksi wajah. Evaluasi dilakukan dengan membandingkan kualitas gambar sebelum dan sesudah pra-pemrosesan menggunakan metrik *Peak Signal-to-Noise Ratio* (PSNR) dan *Structural Similarity Index Measure* (SSIM), yang umum digunakan dalam evaluasi kualitas gambar dan pra-pemrosesan wajah pada sistem visi komputer. PSNR digunakan untuk mengukur tingkat gangguan (*noise*) pada gambar hasil pemrosesan, sedangkan SSIM digunakan untuk menilai kesamaan struktur visual antara gambar asli dan gambar hasil pra-pemrosesan.

Selain itu, stabilitas hasil pra-pemrosesan diamati melalui konsistensi gambar masukan yang dapat diproses secara baik pada tahap deteksi. Pra-pemrosesan yang efektif diharapkan mampu mempertahankan informasi visual penting dan mendukung performa deteksi wajah pada kondisi pencahayaan tidak stabil, tingkat kepadatan tinggi, dan *occlusion*.

2. Evaluasi Deteksi Wajah

Tahap kedua evaluasi sistem bertujuan untuk mengukur kemampuan model dalam menghasilkan deteksi wajah yang tepat dan stabil pada lingkungan keramaian. Evaluasi dilakukan dengan membandingkan *bounding box* hasil prediksi dengan anotasi referensi (*ground truth*) pada *frame* uji. Metrik evaluasi

yang digunakan meliputi *Precision*, *Recall*, *F1-score*, *confidence score*, dan *mean Average Precision* (mAP@0.5).

Precision digunakan untuk menggambarkan tingkat ketepatan hasil deteksi wajah, sedangkan *Recall* menunjukkan kemampuan model dalam mendeteksi seluruh wajah yang muncul pada *frame*. *F1-score* digunakan sebagai indikator harmonisasi antara *precision* dan *recall* sehingga dapat memberikan gambaran keseimbangan performa model secara keseluruhan. Sementara itu, mAP@0.5 digunakan untuk mengukur akurasi deteksi model berdasarkan kesesuaian lokasi *bounding box* dengan anotasi referensi menggunakan ambang *Intersection over Union* (IoU) sebesar 0.5.

Selain evaluasi hasil deteksi, kestabilan proses pelatihan juga dianalisis melalui *validation loss* dan *confusion matrix*. *Validation loss* digunakan untuk memantau konvergensi model selama proses pelatihan dan mengidentifikasi potensi *overfitting*, sedangkan *confusion matrix* digunakan untuk mengevaluasi tingkat kesalahan klasifikasi antara objek wajah dan non-wajah. Evaluasi ini dilakukan pada setiap konfigurasi rasio data dan parameter pelatihan untuk menentukan model terbaik berdasarkan performa deteksi yang dihasilkan.

3. Evaluasi Pengaruh Parameter Pelatihan

Penelitian ini juga melakukan evaluasi terhadap pengaruh parameter pelatihan model terhadap performa sistem deteksi wajah. Parameter yang dianalisis meliputi jumlah *epoch*, *batch size*, dan *learning rate*. Evaluasi dilakukan dengan membandingkan perubahan nilai *precision*, *recall*, *F1-score*, mAP@0.5, serta *validation loss* pada setiap konfigurasi eksperimen.

Peningkatan jumlah *epoch* digunakan untuk mengevaluasi pengaruh lamanya proses pembelajaran terhadap kemampuan model membangun representasi fitur wajah. Variasi *batch size* dianalisis untuk menilai kestabilan pembaruan gradien selama proses pelatihan, sedangkan variasi *learning rate* digunakan untuk mengukur sensitivitas model terhadap perubahan proses optimasi bobot. Melalui evaluasi ini, penelitian dapat menentukan konfigurasi parameter yang menghasilkan performa deteksi paling optimal pada lingkungan keramaian.

4. Evaluasi Pelacakan Wajah

Tahap berikutnya adalah mengevaluasi kestabilan sistem dalam mempertahankan identitas wajah pada rangkaian *frame* video. Proses pelacakan menggunakan *DeepSORT* menghasilkan *Track ID* yang mengikuti objek wajah sepanjang pergerakannya. Evaluasi pelacakan dilakukan menggunakan metrik standar pada penelitian *Multiple Object Tracking* (MOT), seperti IDF1, MOTA, dan jumlah *ID Switches*.

IDF1 digunakan untuk mengukur keberhasilan sistem dalam mempertahankan identitas objek secara konsisten. MOTA menilai akurasi pelacakan berdasarkan kesalahan deteksi, kehilangan objek, dan pergantian identitas. Jumlah *ID Switches* menjadi indikator penting untuk menilai kesinambungan pelacakan pada kondisi pergerakan dinamis dan *occlusion*. Evaluasi pelacakan dilakukan menggunakan konfigurasi model terbaik hasil tahap deteksi untuk memastikan kestabilan identitas wajah pada seluruh skenario pengujian.

5. Evaluasi Efisiensi Sistem

Efisiensi sistem menjadi aspek penting dalam evaluasi, terutama untuk menilai kelayakan penerapan sistem pada skenario *real-time*. Parameter efisiensi yang digunakan meliputi waktu pemrosesan per *frame* dan *Frame Per Second* (FPS). FPS memberikan gambaran kemampuan sistem dalam memproses video secara kontinu tanpa keterlambatan signifikan.

Eksperimen dijalankan pada lingkungan Google Colab dengan dukungan GPU berbasis CUDA, sehingga kecepatan pemrosesan dapat bervariasi tergantung konfigurasi GPU yang digunakan. Evaluasi efisiensi dilakukan dengan mencatat nilai FPS selama pengujian pada *test set* HT21.

6. Evaluasi Konsistensi Lintasan

Konsistensi lintasan (*trajectory*) digunakan sebagai indikator tambahan dalam menilai kualitas pelacakan wajah. *Trajectory* menggambarkan alur pergerakan wajah dari satu *frame* ke *frame* berikutnya. Evaluasi dilakukan dengan mengamati kelancaran lintasan, kesinambungan posisi *bounding box*, serta kestabilan identitas wajah sepanjang video. Lintasan yang terputus atau tidak konsisten mengindikasikan adanya gangguan pada proses deteksi maupun pelacakan.

Karakteristik dataset HT21 yang memiliki pergerakan kerumunan dinamis dan tingkat *occlusion* tinggi menjadikan evaluasi lintasan penting untuk menilai kemampuan sistem dalam menghadapi kondisi tersebut.

7. Relevansi Kriteria dengan Tujuan Penelitian

Seluruh kriteria evaluasi yang diterapkan dirancang untuk mendukung tujuan utama penelitian, yaitu menghasilkan sistem deteksi dan pelacakan wajah

yang stabil, akurat, dan efisien pada lingkungan berkeramaian tinggi. Dataset HT21 yang diperkenalkan oleh Sundararaman et al. (2021) digunakan sebagai acuan karena merepresentasikan tantangan nyata pada ruang publik. Dengan menerapkan kriteria evaluasi yang terstruktur pada setiap tahap sistem, penelitian ini mampu memberikan penilaian performa yang komprehensif serta mendukung analisis pengaruh variasi rasio data dan parameter pelatihan terhadap kualitas deteksi dan pelacakan wajah.

BAB IV

IMPLEMENTASI DAN ANALISIS PRA-PEMROSESAN GAMBAR

Bab ini membahas pra-pemrosesan gambar sebagai tahap kunci sebelum deteksi dan pelacakan wajah pada lingkungan keramaian padat. Gambar dengan variasi pencahayaan, skala, dan oklusi diproses melalui penyesuaian resolusi, normalisasi, konversi format, dan augmentasi untuk meningkatkan kualitas serta konsistensi data. Tahap ini bertujuan memperjelas fitur wajah dan menstabilkan kinerja model YOLOv9. Analisis kuantitatif dan visual menunjukkan bahwa pra-pemrosesan berpengaruh langsung terhadap akurasi dan stabilitas sistem secara keseluruhan.

4.1 Hasil Kuantitatif dan Analisis Pra-Pemrosesan Data

Evaluasi kuantitatif pra-pemrosesan data dilakukan terhadap 10 sampel gambar lingkungan keramaian menggunakan tiga parameter utama, yaitu *Detection Score* (DET), *Peak Signal-to-Noise Ratio* (PSNR), dan *Structural Similarity Index Measure* (SSIM). Parameter DET digunakan untuk menggambarkan tingkat keterdeteksian wajah dalam suatu citra, khususnya pada kondisi visual kompleks seperti kerumunan padat dan occlusion. Sementara itu, PSNR dan SSIM digunakan untuk menilai kualitas visual hasil pemrosesan serta tingkat preservasi struktur spasial gambar wajah. Hasil evaluasi kuantitatif disajikan pada Tabel 4.1.

No	Dataset	File	DET Score	PSNR	SSIM
1	HT21-14	000806.jpg	118.56	36.69	0.9764
2	HT21-12	002029.jpg	45.92	41.00	0.9880
3	HT21-13	000124.jpg	183.06	33.86	0.9742
4	HT21-13	000819.jpg	207.59	33.72	0.9719
5	HT21-15	000785.jpg	142.12	37.47	0.9847
6	HT21-12	000469.jpg	78.85	38.78	0.9851
7	HT21-11	000379.jpg	111.22	40.23	0.9888
8	HT21-13	000685.jpg	169.47	34.27	0.9744
9	HT21-14	000277.jpg	119.95	36.82	0.9768
10	HT21-13	000387.jpg	181.43	33.80	0.9724

Tabel 4.1 Hasil Evaluasi Pra-Pemrosesan pada 10 Sampel Gambar

Berdasarkan Tabel 4.1, nilai DET berada pada rentang 45,92 hingga 207,59. Nilai terendah diperoleh pada sampel HT21-12/002029.jpg atau tabel nomor 2 yang merepresentasikan kondisi kepadatan wajah relatif rendah, sedangkan nilai tertinggi diperoleh pada sampel HT21-13/000819.jpg atau tabel nomor nomor 4 yang menunjukkan kondisi kerumunan sangat padat. Variasi nilai tersebut mencerminkan perbedaan tingkat kompleksitas visual antar citra dan menunjukkan bahwa pra-pemrosesan tidak menghilangkan informasi wajah yang penting bagi sistem deteksi. DET yang tetap tinggi pada kondisi padat mengindikasikan bahwa fitur wajah masih dapat dipertahankan sebagai kandidat objek deteksi.

Dari sisi kualitas visual, nilai PSNR berada pada rentang 33,72 dB hingga 41,00 dB. Seluruh nilai berada di atas ambang 30 dB yang secara umum dikategorikan sebagai kualitas citra baik dalam pengolahan gambar wajah. Hal ini menunjukkan bahwa proses reduksi noise dan peningkatan kontras dilakukan tanpa menyebabkan degradasi visual yang signifikan. Sementara itu, nilai SSIM berada

pada rentang 0,9719 hingga 0,9888, yang mendekati nilai maksimum 1. Konsistensi nilai SSIM tersebut mengindikasikan bahwa struktur spasial wajah tetap terjaga dengan sangat baik setelah pra-pemrosesan. Preservasi struktur ini menjadi aspek penting dalam menjaga stabilitas *bounding box* dan *confidence score* pada tahap deteksi berbasis YOLOv9.

Untuk melengkapi analisis kuantitatif, dilakukan evaluasi visual terhadap beberapa sampel representatif berdasarkan variasi tingkat kepadatan wajah. Pada sampel dengan kepadatan rendah yang ditunjukkan pada Gambar 4.1, pra-pemrosesan menunjukkan peningkatan kejelasan tekstur dan stabilitas pencahayaan tanpa distorsi struktural.

Gambar 4.1 Perbandingan gambar asli, gambar setelah penyaringan noise, dan gambar akhir (Sampel 2)

Pada sampel dengan kepadatan sedang seperti yang terlihat pada Gambar 4.2, terlihat pengurangan noise latar belakang secara signifikan tanpa menghilangkan detail penting pada area wajah. Struktur wajah tetap terdefinisi dengan baik sehingga tetap layak sebagai masukan sistem deteksi.

Gambar 4.2 Perbandingan gambar asli, gambar setelah penyaringan noise, dan gambar akhir (Sampel 1)

Pada kondisi kerumunan sangat padat seperti pada Gambar 4.3, meskipun kompleksitas visual meningkat dan nilai PSNR sedikit menurun, struktur wajah tetap dapat dikenali dengan baik. Penurunan ringan tersebut masih berada dalam batas toleransi dan tidak mengindikasikan degradasi visual yang signifikan.

Gambar 4.3 Perbandingan gambar asli, gambar setelah penyaringan noise, dan gambar akhir (Sampel 4)

Secara keseluruhan, hasil kuantitatif dan visual menunjukkan bahwa tahapan pra-pemrosesan berhasil menjaga keseimbangan antara peningkatan kualitas gambar dan preservasi informasi visual wajah. Nilai PSNR dan SSIM yang tinggi menunjukkan stabilitas kualitas citra, sedangkan variasi DET yang selaras dengan tingkat kepadatan objek menunjukkan bahwa informasi wajah tetap terjaga. Dengan demikian, hasil pra-pemrosesan dinilai layak dan stabil untuk digunakan sebagai masukan pada tahap deteksi wajah menggunakan YOLOv9 dengan *attention mechanism* pada lingkungan keramaian.

4.2 Pembahasan Hasil Pra-Pemrosesan Data

Berdasarkan hasil evaluasi kuantitatif dan analisis visual, tahapan pra-pemrosesan menunjukkan kontribusi yang signifikan terhadap peningkatan kualitas citra sekaligus menjaga kesiapan data sebagai masukan sistem deteksi wajah. Pembahasan ini menitikberatkan pada hubungan antara tiga parameter evaluasi utama, yaitu *Detection Score* (DET), *Peak Signal-to-Noise Ratio* (PSNR), dan *Structural Similarity Index Measure* (SSIM), sebagaimana lazim digunakan dalam evaluasi sistem pengolahan citra dan deteksi wajah berbasis visi komputer (Singh & Reibman, 2024).

Nilai PSNR berada pada rentang 33,72-41,00 dB. Nilai tertinggi diperoleh pada citra dengan DET rendah (DET = 45,92), yang merepresentasikan kondisi kerumunan relatif jarang dan gangguan visual minimal. Sebaliknya, nilai PSNR terendah diperoleh pada gambar dengan DET tertinggi (DET = 207,59), yang menunjukkan kompleksitas visual dan tingkat occlusion yang lebih tinggi. Pola ini menunjukkan bahwa peningkatan kepadatan objek secara alami memengaruhi kualitas sinyal visual. Meskipun demikian, seluruh nilai PSNR tetap berada di atas 30 dB, yang secara umum dikategorikan sebagai kualitas gambar baik dan memadai untuk proses ekstraksi fitur visual (Singh & Reibman, 2024; Salman & Kalakech, 2024). Hal ini mengindikasikan bahwa proses pra-pemrosesan mampu mereduksi noise tanpa menyebabkan degradasi visual yang signifikan.

Dari sisi kesetiaan struktur spasial, nilai SSIM berada pada rentang 0,9719-0,9888, yang menunjukkan tingkat kemiripan struktur yang sangat tinggi antara

citra sebelum dan sesudah pra-pemrosesan. Meskipun terjadi penurunan ringan pada citra dengan DET tinggi, nilainya tetap mendekati 1 dan berada dalam batas toleransi yang dapat diterima. Preservasi struktur ini penting karena kestabilan geometri wajah secara langsung memengaruhi ketepatan bounding box dan confidence score pada tahap deteksi (Singh & Reibman, 2024; Guo et al., 2022). Dengan demikian, hasil SSIM menguatkan bahwa proses peningkatan kualitas tidak mengganggu informasi spasial yang krusial bagi sistem deteksi.

Nilai DET yang berkisar antara 45,92 hingga 207,59 merefleksikan variasi tingkat kepadatan wajah pada setiap sampel uji. Peningkatan nilai DET seiring dengan bertambahnya jumlah wajah dalam citra tidak diikuti oleh penurunan kualitas visual yang signifikan, sebagaimana dibuktikan oleh stabilitas PSNR dan SSIM. Hal ini menunjukkan bahwa tahapan pra-pemrosesan berhasil menjaga keseimbangan antara reduksi noise dan preservasi informasi wajah, yang merupakan prasyarat penting dalam sistem deteksi wajah pada lingkungan keramaian padat (Singh & Reibman, 2024).

Secara keseluruhan, hubungan antara ketiga parameter menunjukkan pola yang konsisten. Pada citra dengan DET rendah hingga sedang (DET < 120), nilai PSNR dan SSIM cenderung lebih tinggi dan stabil. Pada citra dengan DET tinggi (DET > 170), terjadi penurunan ringan akibat kompleksitas visual dan occlusion yang meningkat, namun masih berada pada rentang yang mendukung proses ekstraksi fitur oleh model deteksi wajah (Guo et al., 2022). Dengan demikian, tahapan pra-pemrosesan dalam penelitian ini dapat dinyatakan efektif karena mampu meningkatkan kualitas citra tanpa mengorbankan informasi visual yang

esensial. Hasil ini menjadi landasan yang kuat untuk tahap implementasi deteksi wajah menggunakan YOLOv9 dengan *attention mechanism* serta integrasi pelacakan menggunakan *DeepSORT* pada lingkungan keramaian padat.

BAB V

IMPLEMENTASI DAN EVALUASI DETEKSI WAJAH MENGGUNAKAN YOLOv9 DENGAN *ATTENTION MECHANISM*

Bab ini melanjutkan hasil pra-pemrosesan pada Bab IV dengan mengimplementasikan gambar yang telah dinormalisasi dan diaugmentasi ke dalam model deteksi wajah berbasis YOLOv9 dengan . Proses evaluasi dilakukan melalui beberapa skenario eksperimen yang meliputi variasi rasio data *training-validation* sebesar 90:10, 50:50, dan 10:90, serta variasi parameter pelatihan berupa jumlah *epoch* (50, 100, 200 dan 300), *batch size* (8 dan 16), dan *learning rate* (0.001 dan 0.0005) untuk menganalisis pengaruh distribusi data dan konfigurasi pelatihan terhadap performa deteksi wajah pada lingkungan keramaian padat. Analisis dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, *mean Average Precision* (mAP@0.5), *validation loss*, *confusion matrix*, serta kurva *confidence* untuk mengevaluasi kestabilan generalisasi model dan menentukan konfigurasi terbaik yang selanjutnya digunakan pada tahap pelacakan identitas menggunakan *DeepSORT*.

5.1 Implementasi Deteksi Wajah YOLOv9 dengan *Attention Mechanism*

Implementasi deteksi wajah pada penelitian ini menggunakan arsitektur YOLOv9 yang diperkaya dengan sebagai kelanjutan dari tahap pra-pemrosesan pada Bab IV. Seluruh citra yang digunakan telah melalui proses normalisasi piksel, penyesuaian resolusi, dan augmentasi sehingga berada dalam kondisi terstandarisasi sebelum diproses oleh jaringan. Standarisasi ini berperan penting

dalam menjaga kestabilan konvergensi selama proses pelatihan serta meningkatkan konsistensi ekstraksi fitur pada kondisi lingkungan keramaian yang kompleks (Wang et al., 2022; Gong et al., 2022). Integrasi diterapkan untuk memperkuat representasi fitur yang relevan pada area wajah melalui proses pembobotan ulang (*re-weighting*) terhadap fitur spasial dan kanal, sehingga model menjadi lebih selektif terhadap informasi penting dan lebih tahan terhadap gangguan latar belakang (*background noise*) serta kondisi *occlusion* (Guo et al., 2022; Wang et al., 2023). Dalam implementasinya, citra hasil pra-pemrosesan dimasukkan ke jaringan YOLOv9 untuk menjalani proses ekstraksi fitur, penguatan fitur melalui , dan prediksi *bounding box* beserta *confidence score* wajah. Proses implementasi dilakukan melalui beberapa konfigurasi eksperimen yang meliputi variasi rasio data *training-validation* (90:10, 50:50, dan 10:90) serta variasi parameter pelatihan berupa jumlah *epoch* (50 dan 100), *batch size* (8 dan 16), dan *learning rate* (0.001 dan 0.0005). Setiap konfigurasi dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, mAP@0.5, *validation loss*, *confusion matrix*, dan kurva *confidence* untuk menganalisis pengaruh distribusi data dan konfigurasi pelatihan terhadap performa deteksi wajah pada lingkungan keramaian padat.

5.2 Evaluasi Kinerja Model YOLOv9 dengan *Attention Mechanism* Berdasarkan Variasi Rasio Data dan Parameter Pelatihan

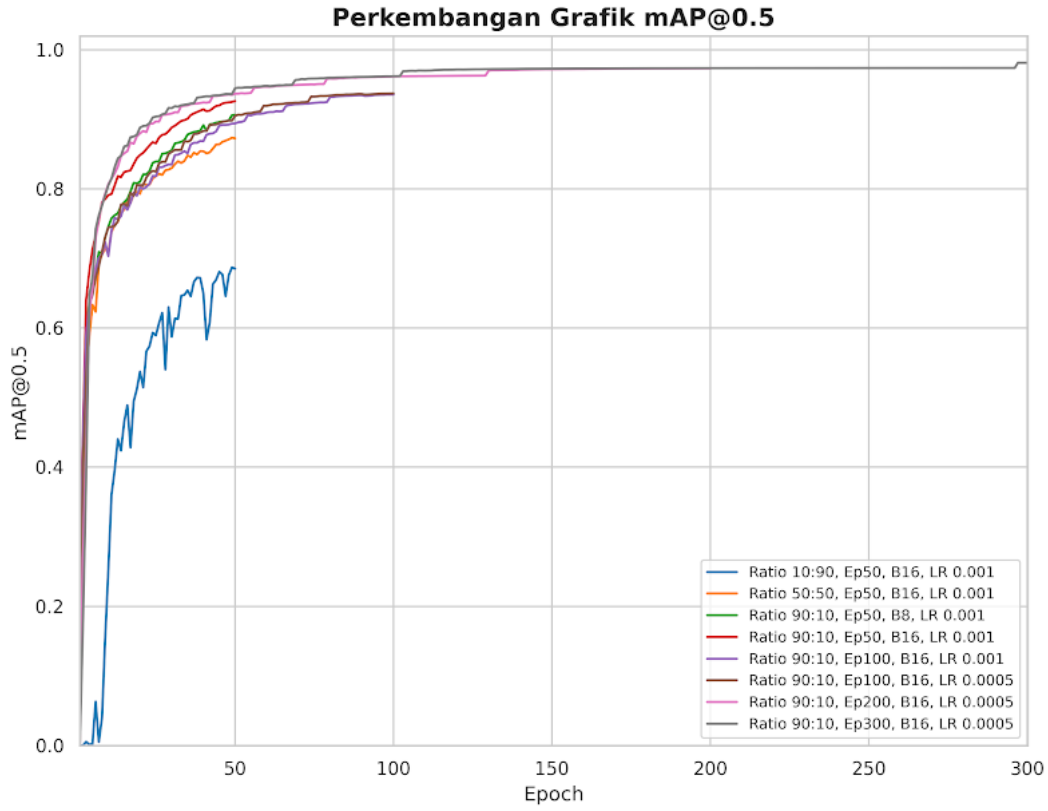
Evaluasi kinerja model YOLOv9 dengan *Attention Mechanism* dilakukan melalui delapan skenario eksperimen yang mencakup variasi rasio data *training-validation*, jumlah *epoch*, *batch size*, dan *learning rate*. Analisis difokuskan pada

perkembangan nilai *precision*, *recall*, *F1-score*, dan mAP@0.5 selama proses pelatihan untuk mengevaluasi kualitas konvergensi, kemampuan generalisasi, serta pengaruh masing-masing parameter terhadap performa deteksi wajah pada lingkungan keramaian padat.

Berdasarkan perkembangan mAP@0.5 pada Gambar 5.1, terlihat bahwa distribusi data pelatihan memberikan pengaruh paling besar terhadap kualitas deteksi model. Kurva biru yang merepresentasikan konfigurasi rasio 10:90, epoch 50, batch size 16, learning rate 0.001 menunjukkan performa terendah dengan mAP@0.5 hanya mencapai sekitar 0.687 pada akhir pelatihan. Kurva oranye yang merepresentasikan rasio 50:50, epoch 50, batch size 16, learning rate 0.001 menunjukkan peningkatan yang lebih baik dengan mAP@0.5 mencapai sekitar 0.874. Sementara itu, kurva merah yang mewakili rasio 90:10, epoch 50, batch size 16, learning rate 0.001 mencapai sekitar 0.926, menunjukkan bahwa peningkatan jumlah data latih menghasilkan representasi fitur wajah yang lebih baik terhadap variasi pencahayaan, ukuran objek, dan kondisi *occlusion*.

Selain rasio data, pengaruh parameter pelatihan juga terlihat jelas. Kurva ungu yang mewakili rasio 90:10, epoch 100, batch size 16, learning rate 0.001 menunjukkan peningkatan performa hingga mAP@0.5 sekitar 0.936. Kurva coklat yang merepresentasikan rasio 90:10, epoch 100, batch size 16, learning rate 0.0005 menghasilkan performa sedikit lebih tinggi dengan mAP@0.5 sekitar 0.937. Peningkatan lebih lanjut terlihat pada kurva merah muda (pink) untuk konfigurasi epoch 200, batch size 16, learning rate 0.0005, yang mencapai mAP@0.5 sebesar 0.973. Sementara itu, kurva abu-abu yang merepresentasikan epoch 300, batch size

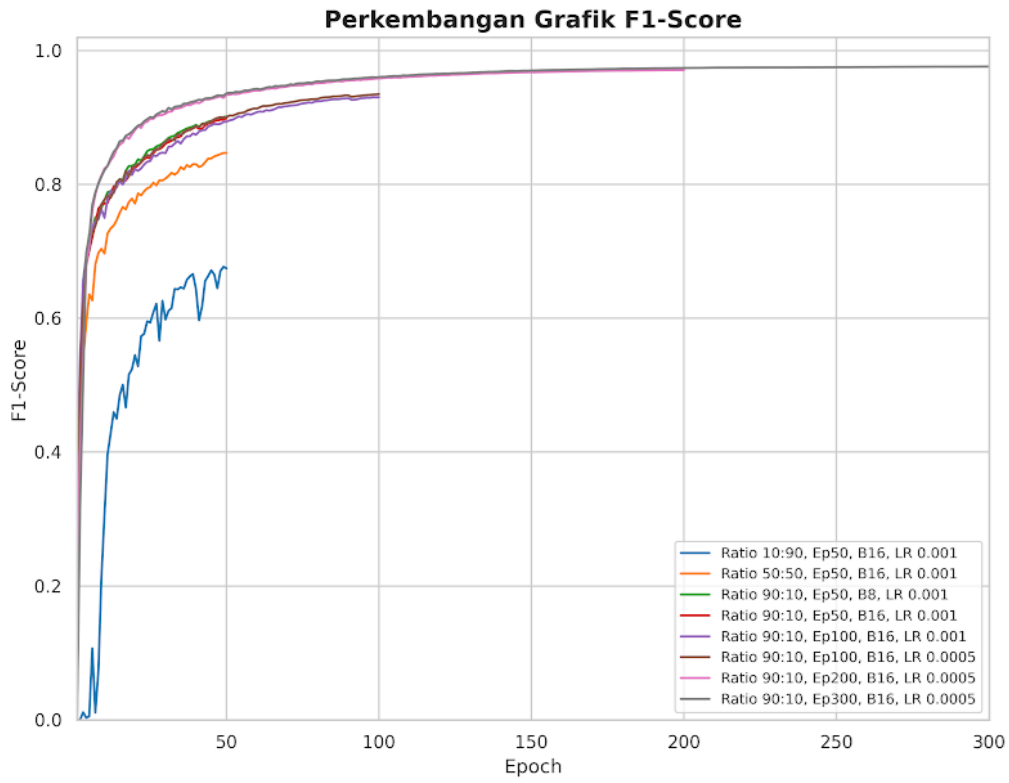
16, learning rate 0.0005 mencapai nilai tertinggi sekitar 0.981, namun menunjukkan gejala kejenuhan (*plateau*) pada fase akhir pelatihan.



Gambar 5.1 Perkembangan mAP@0.5 Per Epoch

Perkembangan F1-score pada Gambar 5.2 memperlihatkan pola yang konsisten dengan hasil mAP@0.5. Kurva biru (rasio 10:90) menunjukkan peningkatan lambat dengan nilai akhir sekitar 0.677, sedangkan kurva oranye (rasio 50:50) mencapai sekitar 0.847. Kurva merah (rasio 90:10, epoch 50) mencapai sekitar 0.899, kemudian meningkat pada kurva ungu (epoch 100, learning rate 0.001) menjadi sekitar 0.931 dan kurva coklat (epoch 100, learning rate 0.0005) menjadi sekitar 0.935. Peningkatan paling signifikan terlihat pada kurva pink (epoch 200) yang mencapai 0.971, sedangkan kurva abu-abu (epoch 300) mencapai sekitar 0.976. Pola ini menunjukkan bahwa penambahan *epoch* memberikan

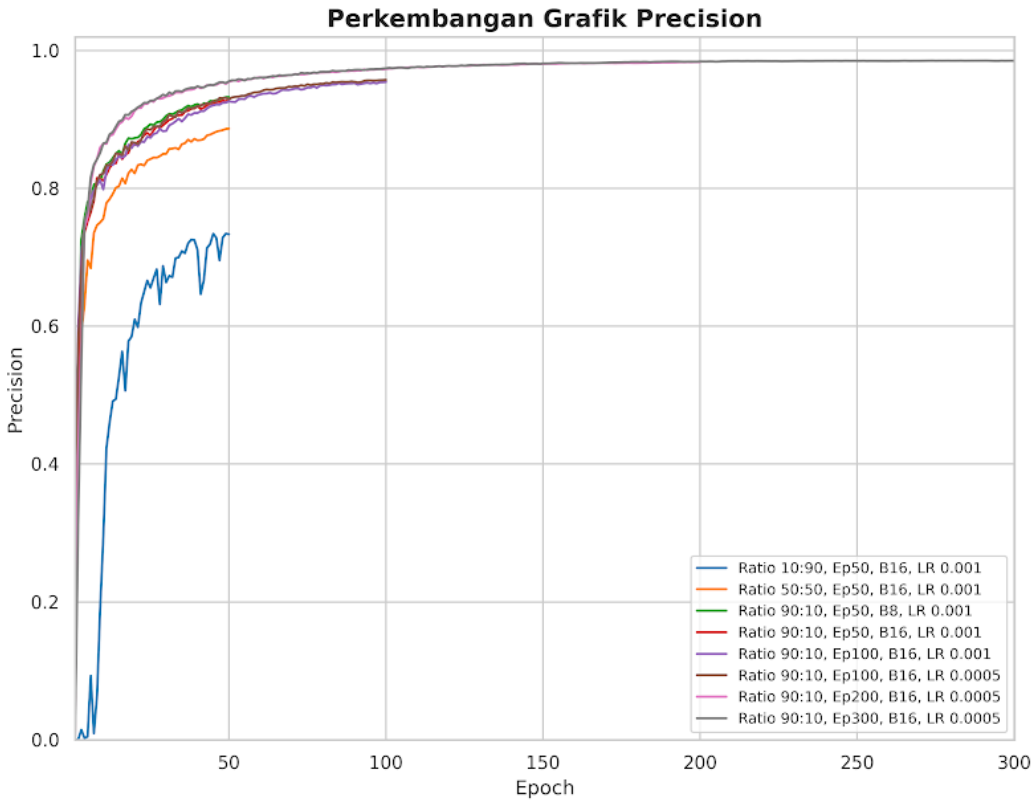
kesempatan lebih besar bagi model untuk menyempurnakan representasi fitur wajah, terutama ketika dikombinasikan dengan *learning rate* yang lebih rendah.



Gambar 5.2 Perkembangan F1-Score Per Epoch

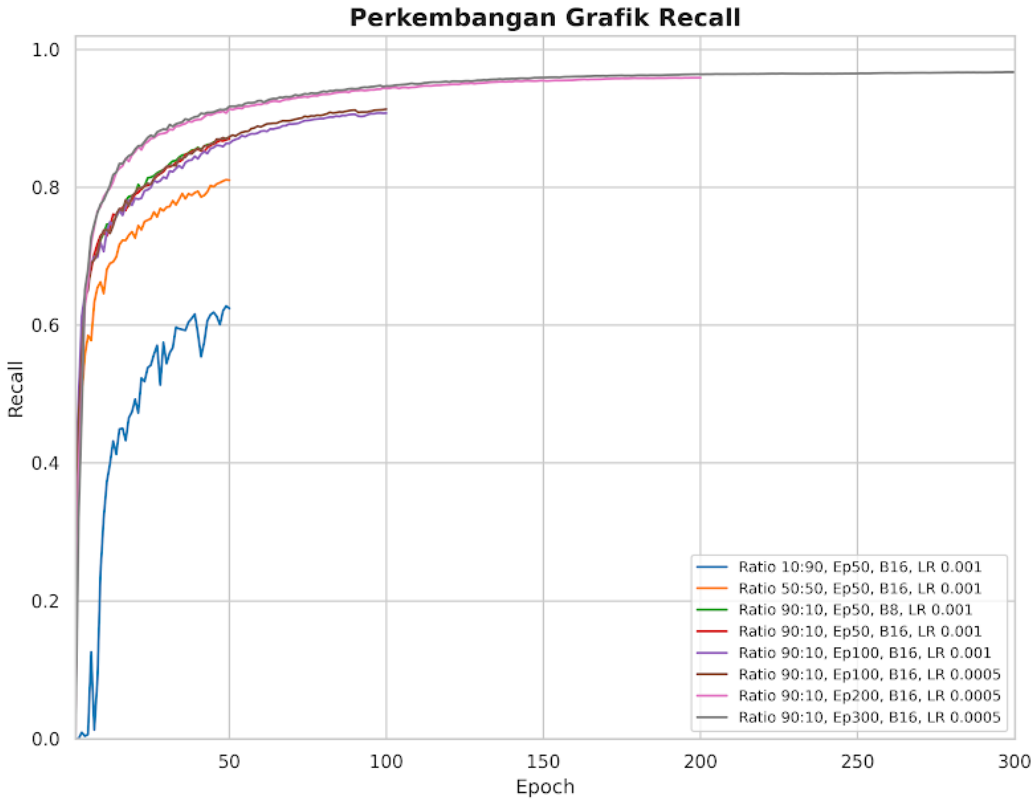
Perkembangan *precision* pada Gambar 5.3 menunjukkan bahwa seluruh konfigurasi rasio 90:10 memiliki performa yang lebih baik dibandingkan rasio lainnya. Kurva merah, ungu, coklat, pink, dan abu-abu secara konsisten berada di atas kurva biru dan oranye. Konfigurasi terbaik ditunjukkan oleh kurva abu-abu (epoch 300) dan pink (epoch 200), yang mampu mencapai *precision* mendekati 0.98. Namun demikian, peningkatan setelah epoch 200 relatif kecil dibandingkan tambahan waktu pelatihan yang diperlukan. Sementara itu, kurva hijau yang mewakili rasio 90:10, epoch 50, batch size 8, learning rate 0.001 menunjukkan performa sedikit lebih rendah dibandingkan kurva merah dengan *batch size* 16,

mengindikasikan bahwa ukuran *batch* yang lebih besar menghasilkan pembaruan bobot yang lebih stabil.



Gambar 5.3 Perkembangan Precision Per Epoch

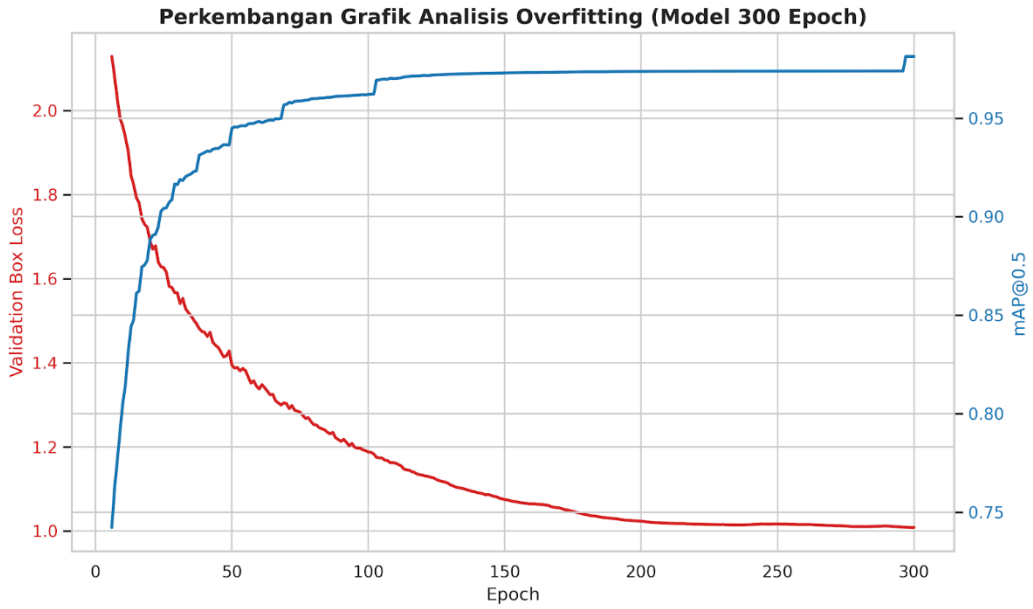
Pada Gambar 5.4, perkembangan *recall* menunjukkan pola yang serupa dengan *precision*. Kurva biru (rasio 10:90) berhenti pada kisaran 0.62, sedangkan kurva oranye (rasio 50:50) mencapai sekitar 0.81. Konfigurasi rasio 90:10 menunjukkan performa yang jauh lebih baik dengan kurva merah mencapai sekitar 0.90, kurva coklat sekitar 0.91, dan kurva pink serta abu-abu mendekati 0.96–0.97. Hal ini menunjukkan bahwa peningkatan jumlah data latih dan durasi pelatihan memperbaiki kemampuan model dalam menemukan wajah yang seharusnya terdeteksi, bahkan pada kondisi keramaian padat dan *occlusion* tinggi.



Gambar 5.4 Perkembangan Recall Per Epoch

Untuk mengevaluasi kemungkinan terjadinya *overfitting*, dilakukan analisis tambahan pada konfigurasi pelatihan hingga 300 *epoch* sebagaimana ditunjukkan pada Gambar 5.5. Kurva mAP@0.5 terus meningkat hingga mendekati 0.98, sementara *validation box loss* terus menurun hingga mencapai titik minimum pada kisaran epoch 220–222. Setelah titik tersebut, peningkatan akurasi menjadi sangat kecil dan kurva mulai memasuki fase *plateau*. Berdasarkan analisis perkembangan *validation loss*, gejala *overfitting* mulai muncul setelah epoch 222, ketika model semakin menyesuaikan diri terhadap data pelatihan tanpa memperoleh peningkatan kemampuan generalisasi yang signifikan. Dengan demikian, meskipun konfigurasi 300 *epoch* menghasilkan nilai metrik tertinggi, konfigurasi 90:10, epoch 200, batch

size 16, learning rate 0.0005 dapat dianggap sebagai titik optimum antara akurasi dan kemampuan generalisasi model.



Gambar 5.5 Analisis Overfitting Berdasarkan Validation Loss dan mAP@0.5

Secara keseluruhan, hasil evaluasi menunjukkan bahwa performa model YOLOv9 dengan *Attention Mechanism* dipengaruhi secara signifikan oleh distribusi data pelatihan dan konfigurasi parameter pelatihan. Rasio data 90:10 secara konsisten menghasilkan performa terbaik dibandingkan rasio lainnya. Peningkatan jumlah *epoch* hingga 200 terbukti meningkatkan kemampuan generalisasi model secara signifikan, sedangkan penurunan *learning rate* menjadi 0.0005 menghasilkan proses konvergensi yang lebih halus dan stabil. Berdasarkan seluruh metrik evaluasi dan analisis *overfitting*, konfigurasi rasio 90:10, epoch 200, batch size 16, dan learning rate 0.0005 merupakan konfigurasi paling optimal untuk menghasilkan model deteksi wajah yang akurat dan robust pada lingkungan keramaian padat.

5.3 Pembahasan Hasil YOLOv9 dengan *Attention Mechanism*

Hasil pengujian menunjukkan bahwa performa model YOLOv9 dengan *Attention Mechanism* dipengaruhi secara signifikan oleh distribusi data pelatihan dan konfigurasi parameter pelatihan yang digunakan. Variasi rasio data, jumlah *epoch*, *batch size*, dan *learning rate* menghasilkan perbedaan tingkat konvergensi serta kemampuan generalisasi model yang cukup jelas. Secara umum, konfigurasi dengan rasio data 90:10 menunjukkan performa yang lebih stabil dibandingkan rasio lainnya, sedangkan peningkatan jumlah *epoch* dan optimasi *learning rate* memberikan kontribusi terhadap peningkatan kualitas deteksi wajah secara keseluruhan.

No	Konfigurasi Model	Precision	Recall	F1-Score	mAP@0.5	Kondisi Model
1	Ratio 10:90, Ep50, B16, LR0.001	0.7335	0.6245	0.6771	0.6855	Generalisasi lemah
2	Ratio 50:50, Ep50, B16, LR0.001	0.8868	0.8105	0.8471	0.8728	Cukup stabil
3	Ratio 90:10, Ep50, B8, LR0.001	0.9326	0.8727	0.9016	0.9065	Stabil
4	Ratio 90:10, Ep50, B16, LR0.001	0.9287	0.8706	0.8987	0.9264	Optimal
5	Ratio 90:10, Ep100, B16, LR0.001	0.9545	0.9078	0.9306	0.9359	Sangat stabil

No	Konfigurasi Model	Precision	Recall	F1-Score	mAP@0.5	Kondisi Model
6	Ratio 90:10, Ep100, B16, LR0.0005	0.9576	0.9136	0.9351	0.9374	Sangat robust
7	Ratio 90:10, Ep200, B16, LR0.0005	0.9720	0.9704	0.9712	0.9733	Optimal maksimum
8	Ratio 90:10, Ep300, B16, LR0.0005	0.9780	0.9745	0.9762	0.9814	Overfitting mulai muncul

Tabel 5.1 Perbandingan Hasil Evaluasi YOLOv9 dengan *Attention Mechanism*

Berdasarkan Tabel 5.1 baris 1-4, pengaruh rasio data pelatihan terlihat sangat dominan terhadap performa model. Pada konfigurasi yang setara (*epoch* 50, *batch size* 16, dan *learning rate* 0.001), baris 1 dengan rasio 10:90 hanya menghasilkan mAP@0.5 sebesar 0.6855 dan F1-score sebesar 0.6771. Ketika rasio ditingkatkan menjadi 50:50 (baris 2), nilai mAP@0.5 meningkat menjadi 0.8728 dan F1-score menjadi 0.8471. Peningkatan terbesar terjadi pada rasio 90:10 (baris 4) yang mencapai mAP@0.5 sebesar 0.9264 dan F1-score sebesar 0.8987. Hasil ini menunjukkan bahwa jumlah data latih yang lebih besar memungkinkan model membangun representasi fitur wajah yang lebih kaya sehingga mampu mengenali wajah pada berbagai kondisi pencahayaan, variasi ukuran objek, dan *occlusion* yang tinggi.

Peningkatan jumlah *epoch* memberikan dampak yang signifikan terhadap kualitas konvergensi model. Berdasarkan Tabel 5.1 baris 4 dan baris 5, peningkatan *epoch* dari 50 menjadi 100 pada konfigurasi rasio 90:10, *batch size* 16, dan *learning*

rate 0.001 meningkatkan mAP@0.5 dari 0.9264 menjadi 0.9359 serta F1-score dari 0.8987 menjadi 0.9306. Peningkatan lebih lanjut terlihat pada baris 7, di mana konfigurasi *epoch* 200 dengan *learning rate* 0.0005 menghasilkan mAP@0.5 sebesar 0.9733 dan F1-score sebesar 0.9712. Hasil ini menunjukkan bahwa *Attention Mechanism* memerlukan proses pelatihan yang cukup panjang untuk memaksimalkan penguatan fitur wajah pada kondisi keramaian yang kompleks.

Pengaruh *batch size* dapat diamati pada Tabel 5.1 baris 3 dan baris 4. Kedua konfigurasi menggunakan rasio 90:10, *epoch* 50, dan *learning rate* 0.001, namun berbeda pada ukuran *batch*. Penggunaan *batch size* 8 pada baris 3 menghasilkan mAP@0.5 sebesar 0.9065 dan F1-score sebesar 0.9016, sedangkan *batch size* 16 pada baris 4 menghasilkan mAP@0.5 sebesar 0.9264 dan F1-score sebesar 0.8987. Meskipun nilai F1-score keduanya relatif berdekatan, *batch size* 16 menghasilkan akurasi deteksi yang lebih tinggi dan proses konvergensi yang lebih stabil karena pembaruan gradien dilakukan berdasarkan jumlah sampel yang lebih besar.

Pengaruh *learning rate* terlihat pada Tabel 5.1 baris 5 dan baris 6. Pada konfigurasi yang sama, yaitu rasio 90:10, *epoch* 100, dan *batch size* 16, penurunan *learning rate* dari 0.001 (baris 5) menjadi 0.0005 (baris 6) meningkatkan mAP@0.5 dari 0.9359 menjadi 0.9374 dan F1-score dari 0.9306 menjadi 0.9351. Selain itu, nilai *validation box loss* menurun dari 1.3616 menjadi 1.3171. Hasil ini menunjukkan bahwa *learning rate* yang lebih kecil menghasilkan proses pembaruan bobot yang lebih halus sehingga model dapat melakukan *fine-tuning* secara lebih presisi pada tahap akhir pelatihan.

Analisis pelatihan jangka panjang dapat diamati pada Tabel 5.1 baris 7 dan baris 8. Konfigurasi pada baris 7 (*epoch* 200) menghasilkan mAP@0.5 sebesar 0.9733 dan F1-score sebesar 0.9712, sedangkan konfigurasi pada baris 8 (*epoch* 300) menghasilkan mAP@0.5 sebesar 0.9814 dan F1-score sebesar 0.9762. Walaupun terjadi peningkatan nilai metrik, analisis *validation loss* menunjukkan bahwa gejala *overfitting* mulai muncul setelah kisaran *epoch* 220–222. Oleh karena itu, konfigurasi pada baris 7 dapat dianggap sebagai titik optimum karena memberikan keseimbangan terbaik antara akurasi deteksi dan kemampuan generalisasi model.

Secara konseptual, hasil penelitian menunjukkan bahwa performa YOLOv9 dengan *Attention Mechanism* tidak hanya dipengaruhi oleh kekuatan arsitektur model, tetapi juga oleh kualitas distribusi data dan strategi konfigurasi pelatihan. *Attention Mechanism* terbukti efektif dalam memperkuat fitur wajah yang relevan pada lingkungan keramaian dengan tingkat *occlusion* tinggi, namun efektivitas tersebut hanya dapat dicapai secara optimal ketika didukung oleh jumlah data latih yang memadai dan konfigurasi parameter yang tepat. Berdasarkan seluruh hasil evaluasi dan analisis *overfitting*, konfigurasi rasio 90:10, epoch 200, batch size 16, dan learning rate 0.0005 merupakan konfigurasi paling optimal karena memberikan keseimbangan terbaik antara akurasi deteksi, stabilitas konvergensi, dan kemampuan generalisasi model.

BAB VI

IMPLEMENTASI DAN EVALUASI PELACAKAN MENGGUNAKAN MOBILEFACENET DAN *DEEPSORT*

Bab ini membahas implementasi dan evaluasi pelacakan wajah menggunakan integrasi MobileFaceNet dan *DeepSORT* sebagai tahap lanjutan setelah deteksi YOLOv9 dengan *attention mechanism*. Fokus utama adalah menjaga konsistensi ID wajah antar frame pada lingkungan keramaian padat yang dipengaruhi occlusion dan pergerakan dinamis. Evaluasi dilakukan menggunakan metrik tracking pada berbagai rasio data serta pengujian performa real-time berbasis FPS untuk menilai stabilitas dan efisiensi sistem.

6.1 Evaluasi Pelacakan

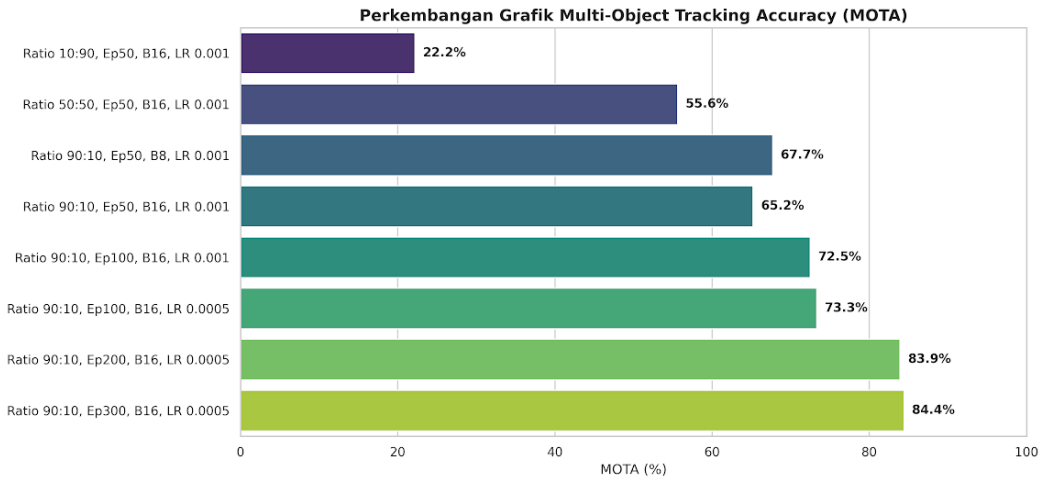
Evaluasi pelacakan pada penelitian ini dilakukan untuk menganalisis pengaruh variasi rasio data dan parameter pelatihan terhadap stabilitas sistem pelacakan berbasis integrasi YOLOv9 dengan *Attention Mechanism*, MobileFaceNet, dan DeepSORT. Pengujian dilakukan menggunakan delapan konfigurasi model yang mencakup variasi rasio data *training-validation* (90:10, 50:50, dan 10:90), jumlah *epoch*, *batch size*, dan *learning rate*. Evaluasi dilakukan pada sekuens video uji HT21 menggunakan metrik pelacakan standar yaitu *Multi-Object Tracking Accuracy* (MOTA), *ID F1-Score* (IDF1), dan *Identity Switches* (IDSW). Hasil evaluasi rata-rata dari seluruh konfigurasi ditampilkan pada Tabel 6.1.

No	Konfigurasi Model	MOTA (%)	IDF1 (%)	Total IDSW
1	Ratio 10:90, Ep50, B16, LR0.001	22.2	41.1	7,603
2	Ratio 50:50, Ep50, B16, LR0.001	55.6	61.2	4,816
3	Ratio 90:10, Ep50, B8, LR0.001	67.7	70.1	3,679
4	Ratio 90:10, Ep50, B16, LR0.001	65.2	68.2	4,046
5	Ratio 90:10, Ep100, B16, LR0.001	72.5	74.0	3,307
6	Ratio 90:10, Ep100, B16, LR0.0005	73.3	74.1	3,234
7	Ratio 90:10, Ep200, B16, LR0.0005	83.9	81.9	1,766
8	Ratio 90:10, Ep300, B16, LR0.0005	84.4	82.3	1,781

Tabel 6.1 Hasil Evaluasi Pelacakan MobileFaceNet-*DeepSORT*

Berdasarkan Tabel 6.1, terlihat bahwa performa pelacakan memiliki hubungan yang sangat kuat dengan kualitas model deteksi yang dihasilkan pada Bab V. Konfigurasi pada baris 7 (rasio 90:10, epoch 200, batch size 16, learning rate 0.0005) menghasilkan keseimbangan terbaik antara akurasi pelacakan dan konsistensi identitas dengan nilai MOTA sebesar 83.9%, IDF1 sebesar 81.9%, serta jumlah IDSW terendah sebesar 1.766. Sementara itu, konfigurasi pada baris 8 memang menghasilkan nilai MOTA dan IDF1 sedikit lebih tinggi, namun jumlah IDSW kembali meningkat menjadi 1.781. Kondisi ini menunjukkan bahwa setelah

epoch 200 sistem mulai memasuki fase kejenuhan yang sejalan dengan indikasi *overfitting* pada model deteksi.

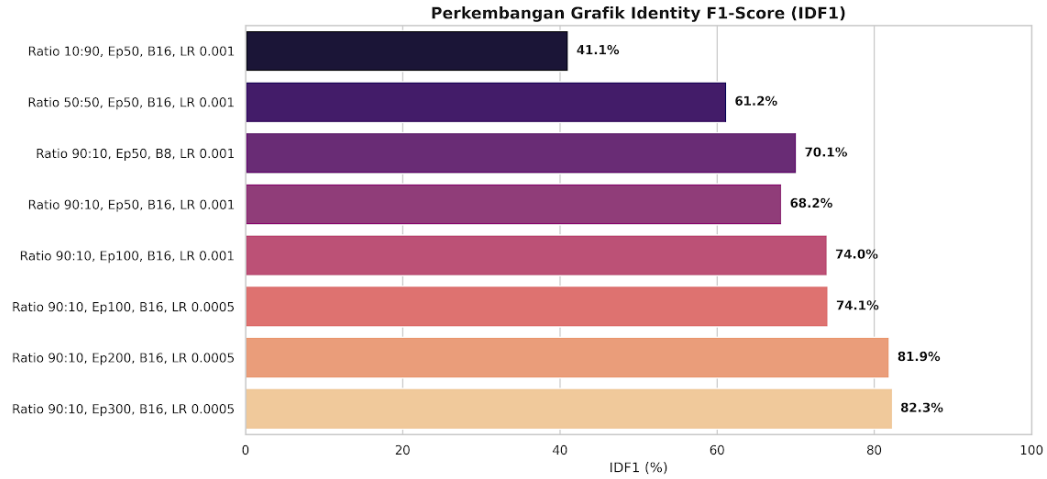


Gambar 6.1 Perbandingan MOTA Berdasarkan Variasi Konfigurasi Model

Pada Gambar 6.1 terlihat bahwa nilai MOTA meningkat secara konsisten seiring peningkatan kualitas konfigurasi pelatihan. Batang ungu tua yang merepresentasikan konfigurasi rasio 10:90, epoch 50, batch size 16, learning rate 0.001 menghasilkan nilai MOTA terendah sebesar 22.2%. Batang biru tua untuk konfigurasi rasio 50:50 meningkat menjadi 55.6%, sedangkan batang biru kehijauan dan hijau toska yang merepresentasikan rasio 90:10 pada epoch 50 menghasilkan nilai MOTA sebesar 67.7% dan 65.2%.

Peningkatan yang lebih signifikan terlihat pada batang hijau (epoch 100, learning rate 0.001) yang mencapai 72.5% dan batang hijau muda (epoch 100, learning rate 0.0005) yang mencapai 73.3%. Lompatan performa terbesar terjadi pada batang hijau terang yang merepresentasikan konfigurasi epoch 200, dengan nilai MOTA mencapai 83.9%. Selanjutnya, batang hijau kekuningan pada konfigurasi epoch 300 hanya meningkat menjadi 84.4%. Kenaikan sebesar 0.5%

tersebut relatif kecil dibanding tambahan waktu pelatihan yang diperlukan, sehingga menunjukkan terjadinya fase *plateau* performa

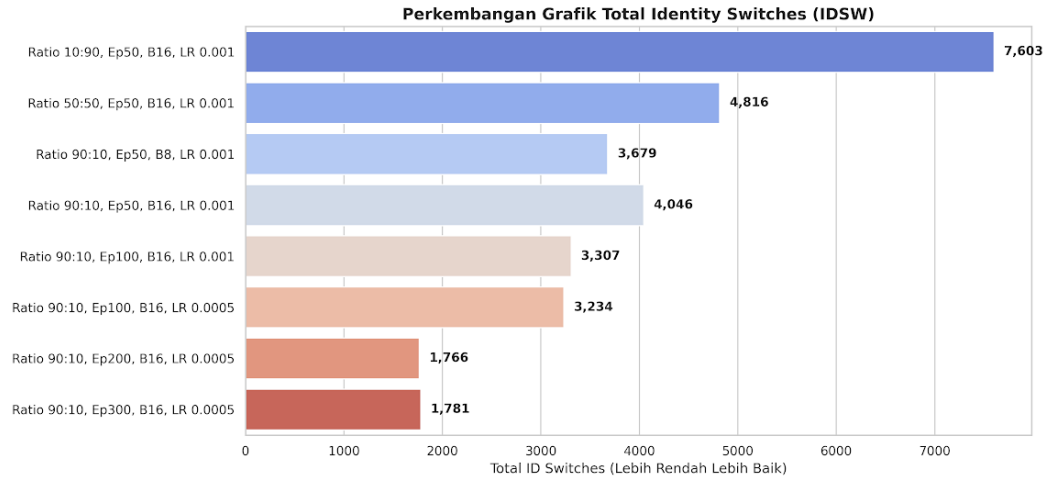


Gambar 6.2 Perbandingan IDF1 Berdasarkan Variasi Konfigurasi Model

Perkembangan nilai IDF1 pada Gambar 6.2 menunjukkan pola yang hampir identik dengan MOTA. Batang ungu tua yang merepresentasikan rasio 10:90 menghasilkan nilai IDF1 terendah sebesar 41.1%, menunjukkan bahwa sistem kesulitan mempertahankan identitas objek secara konsisten ketika kualitas deteksi rendah. Batang ungu untuk rasio 50:50 meningkat menjadi 61.2%, sedangkan konfigurasi rasio 90:10 pada epoch 50 menghasilkan nilai IDF1 sebesar 68.2% hingga 70.1%.

Peningkatan kualitas deteksi pada konfigurasi epoch 100 menyebabkan nilai IDF1 meningkat menjadi 74.0% dan 74.1%. Lompatan terbesar terlihat pada batang oranye muda yang merepresentasikan konfigurasi epoch 200, dengan nilai IDF1 mencapai 81.9%. Batang krem pada konfigurasi epoch 300 menghasilkan nilai tertinggi sebesar 82.3%, namun peningkatannya hanya 0.4% dibanding epoch 200. Hasil ini menunjukkan bahwa MobileFaceNet dan DeepSORT memperoleh

manfaat terbesar ketika kualitas deteksi telah mencapai kondisi optimal pada epoch 200.



Gambar 6.3 Perbandingan ID Switch Berdasarkan Rasio Data

Gambar 6.3 menunjukkan tren yang berlawanan dengan MOTA dan IDF1, karena semakin rendah nilai IDSW menunjukkan performa pelacakan yang semakin baik. Batang biru tua pada konfigurasi rasio 10:90 menghasilkan jumlah IDSW tertinggi sebesar 7.603, yang menunjukkan bahwa sistem sering kehilangan identitas objek dan memberikan label baru pada objek yang sama. Batang biru muda pada rasio 50:50 menurun menjadi 4.816, sedangkan konfigurasi rasio 90:10 pada epoch 50 menghasilkan nilai antara 3.679 hingga 4.046.

Peningkatan epoch menjadi 100 menyebabkan jumlah IDSW turun menjadi 3.307 dan 3.234. Penurunan paling drastis terlihat pada batang oranye muda yang merepresentasikan konfigurasi epoch 200, di mana jumlah IDSW berhasil ditekan hingga 1.766. Menariknya, pada batang merah bata yang merepresentasikan konfigurasi epoch 300, jumlah IDSW kembali meningkat menjadi 1.781. Walaupun peningkatannya relatif kecil, fenomena ini menunjukkan bahwa pelatihan yang

terlalu lama mulai mengurangi kestabilan identitas akibat munculnya gejala *overfitting* pada model deteksi YOLOv9. Ketika *bounding box* mulai mengalami pergeseran kecil dan ketidakkonsistenan pada fase akhir pelatihan, proses asosiasi identitas pada DeepSORT menjadi lebih rentan terhadap kesalahan sehingga memicu kenaikan IDSW.

Secara keseluruhan, hasil evaluasi pelacakan menunjukkan adanya korelasi yang sangat kuat antara kualitas deteksi dan kestabilan pelacakan. Peningkatan kualitas deteksi pada Bab V secara langsung meningkatkan performa *MobileFaceNet-DeepSORT* yang ditunjukkan oleh kenaikan nilai MOTA dan IDF1 serta penurunan jumlah IDSW. Berdasarkan seluruh hasil evaluasi, konfigurasi rasio 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 (Tabel 6.1 baris 7) merupakan konfigurasi paling optimal karena menghasilkan keseimbangan terbaik antara akurasi pelacakan, konsistensi identitas, dan efisiensi sistem. Meskipun konfigurasi *epoch* 300 menghasilkan nilai MOTA dan IDF1 sedikit lebih tinggi, peningkatan tersebut tidak sebanding dengan tambahan waktu pelatihan serta mulai menunjukkan gejala degradasi pelacakan melalui kenaikan kembali nilai IDSW.

6.2 Evaluasi Performa Berdasarkan FPS

Evaluasi performa sistem pelacakan dilakukan menggunakan metrik *Frames Per Second* (FPS) untuk mengukur kemampuan pemrosesan komputasional pada tahap inferensi. FPS merepresentasikan jumlah *frame* yang dapat diproses dalam satu detik dan menjadi indikator utama efisiensi sistem dalam skenario pemantauan keramaian berbasis video. Pada sistem yang mengintegrasikan

YOLOv9 dengan , MobileFaceNet, dan *DeepSORT*, beban komputasi tidak hanya berasal dari proses deteksi wajah, tetapi juga dari ekstraksi *embedding* identitas serta proses asosiasi antar-*frame*. Oleh karena itu, nilai FPS mencerminkan kompleksitas keseluruhan *pipeline* sistem, bukan hanya kecepatan deteksi objek.

Berbeda dengan evaluasi deteksi dan pelacakan yang dilakukan pada beberapa konfigurasi eksperimen, pengujian FPS pada penelitian ini difokuskan pada konfigurasi model terbaik, yaitu rasio 90:10 dengan epoch 200, batch size 16, dan learning rate 0.0005. Konfigurasi ini dipilih karena berdasarkan hasil evaluasi pada Bab V dan Subbab 6.1 menghasilkan keseimbangan terbaik antara akurasi deteksi, performa pelacakan, dan kemampuan generalisasi model sebelum memasuki fase *overfitting*. Pendekatan ini dipilih karena tujuan pengujian FPS adalah mengevaluasi efisiensi sistem pada kondisi performa paling optimal, bukan membandingkan kualitas model. Selain itu, seluruh konfigurasi eksperimen menggunakan arsitektur model, ukuran masukan, serta *pipeline* inferensi yang sama, sehingga perbedaan FPS antar konfigurasi diperkirakan tidak akan berubah secara signifikan dan cenderung dipengaruhi oleh kompleksitas visual data uji dibanding parameter pelatihan. Dengan demikian, pengujian FPS pada model terbaik dinilai telah cukup representatif untuk menggambarkan kemampuan komputasional sistem secara keseluruhan.

Hasil pengujian FPS pada lima *sequence* dataset HT21 disajikan pada Tabel 6.2. Berdasarkan hasil pengujian, terlihat bahwa kecepatan pemrosesan sistem bervariasi antar *sequence*, yang dipengaruhi oleh jumlah objek aktif, kepadatan wajah, serta kompleksitas pergerakan pada masing-masing video uji.

No	Sequence	Total Frame	Avg FPS
1	HT21-11	585	1.72
2	HT21-12	2080	1.45
3	HT21-13	1000	1.64
4	HT21-14	1050	1.57
5	HT21-15	1008	2.38

Tabel 6.2 Evaluasi FPS pada Model Terbaik

Berdasarkan Tabel 6.2, *sequence* HT21-15 menghasilkan performa pemrosesan tertinggi dengan rata-rata FPS sebesar 2.38. Kondisi ini mengindikasikan bahwa kompleksitas visual pada *sequence* tersebut relatif lebih ringan, sehingga proses deteksi wajah, ekstraksi *embedding*, dan asosiasi identitas dapat dilakukan lebih cepat. Sebaliknya, HT21-12 menunjukkan nilai FPS terendah sebesar 1.45, yang mengindikasikan bahwa jumlah *frame* yang besar, kepadatan objek tinggi, serta intensitas interaksi antarobjek menyebabkan beban komputasi meningkat secara signifikan. Pada kondisi keramaian padat, *DeepSORT* harus melakukan proses asosiasi identitas yang lebih kompleks akibat banyaknya *bounding box* yang saling berdekatan dan mengalami *occlusion*.

Untuk memberikan gambaran visual terhadap performa inferensi sistem, ilustrasi pengujian FPS ditampilkan pada Gambar 6.4 menggunakan *sequence* dengan tingkat kepadatan tinggi.

Gambar 6.4 Visualisasi FPS pada Dataset HT21-14 Menggunakan Konfigurasi Model Terbaik

Pemilihan konfigurasi *epoch* 200 sebagai dasar evaluasi FPS juga didukung oleh hasil analisis pelacakan yang menunjukkan bahwa konfigurasi tersebut menghasilkan nilai MOTA sebesar 83.9%, IDF1 sebesar 81.9%, dan jumlah ID *Switches* terendah sebesar 1.766. Meskipun konfigurasi *epoch* 300 menghasilkan peningkatan akurasi yang sangat kecil, jumlah ID *Switches* kembali meningkat dan waktu pelatihan bertambah secara signifikan. Oleh karena itu, pengujian FPS difokuskan pada model *epoch* 200 karena lebih representatif terhadap kondisi implementasi sistem yang optimal.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa bottleneck komputasi pada sistem lebih banyak dipengaruhi oleh jumlah objek yang harus diproses, tingkat *occlusion*, dan kompleksitas asosiasi multiobjek oleh *DeepSORT* dibandingkan parameter pelatihan model. Walaupun sistem menghasilkan rata-rata FPS pada kisaran 1-2 FPS dan belum memenuhi standar *real-time* konvensional, performa tersebut masih relevan untuk kebutuhan *crowd surveillance* berbasis

analitik, evaluasi kepadatan, serta pelacakan identitas berbasis interval *frame* pada lingkungan keramaian padat. Dengan mempertimbangkan hasil deteksi, pelacakan, dan efisiensi komputasi secara bersamaan, konfigurasi rasio 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 merupakan konfigurasi yang paling optimal untuk diimplementasikan pada penelitian ini.

6.3 Hasil dan Pembahasan

Hasil penelitian menunjukkan adanya hubungan yang kuat antara kualitas deteksi wajah, stabilitas pelacakan identitas, dan performa komputasi pada sistem terintegrasi yang terdiri atas tahap pra-pemrosesan, deteksi wajah berbasis YOLOv9 dengan *Attention Mechanism*, ekstraksi fitur menggunakan MobileFaceNet, serta pelacakan identitas berbasis DeepSORT. Evaluasi dilakukan melalui beberapa konfigurasi eksperimen yang meliputi variasi rasio data *training-validation* (90:10, 50:50, dan 10:90) serta variasi parameter pelatihan berupa jumlah *epoch*, *batch size*, dan *learning rate*. Hasil pengujian menunjukkan bahwa distribusi data pelatihan dan konfigurasi parameter memiliki pengaruh signifikan terhadap kemampuan generalisasi model deteksi, konsistensi identitas pelacakan, dan tingkat kesalahan asosiasi objek pada lingkungan keramaian padat.

Pada tahap deteksi, konfigurasi rasio 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 menghasilkan keseimbangan performa terbaik dengan *precision* sebesar 97.20%, *recall* sebesar 97.04%, *F1-score* sebesar 97.12%, dan mAP@0.5 sebesar 97.33%. Peningkatan performa ini menunjukkan bahwa jumlah data latih yang memadai, durasi pelatihan yang lebih panjang, dan laju

pembelajaran yang lebih kecil mampu meningkatkan kemampuan model dalam membangun representasi fitur wajah yang lebih stabil dan diskriminatif. Integrasi *Attention Mechanism* juga terbukti meningkatkan selektivitas fitur spasial pada area wajah sehingga model lebih tahan terhadap gangguan *background noise*, perubahan pencahayaan, variasi ukuran objek, serta kondisi *occlusion* yang kompleks. Sebaliknya, konfigurasi rasio 10:90 menghasilkan performa deteksi paling rendah dengan mAP@0.5 sebesar 68.55% dan F1-score sebesar 67.71%, yang menunjukkan bahwa keterbatasan data pelatihan menyebabkan model kesulitan membangun representasi fitur secara optimal.

Kualitas deteksi yang dihasilkan pada tahap sebelumnya terbukti memberikan dampak langsung terhadap stabilitas pelacakan identitas. Hal tersebut terlihat dari hubungan yang konsisten antara peningkatan nilai mAP@0.5 dan F1-score dengan kenaikan nilai MOTA dan IDF1 pada tahap pelacakan. Konfigurasi rasio 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 menghasilkan nilai MOTA sebesar 83.9%, IDF1 sebesar 81.9%, serta jumlah *Identity Switches* (IDSW) terendah sebesar 1.766. Kondisi ini menunjukkan bahwa *bounding box* yang lebih presisi menghasilkan proses *cropping* wajah yang lebih konsisten sehingga MobileFaceNet mampu membangun *embedding* identitas yang lebih representatif. Dengan kualitas fitur yang lebih stabil, *DeepSORT* menjadi lebih efektif dalam mempertahankan identitas objek antar-*frame* meskipun terjadi pergerakan dinamis, tumpang tindih objek, maupun *occlusion*. Sebaliknya, konfigurasi rasio 10:90 hanya menghasilkan MOTA sebesar 22.2%, IDF1 sebesar 41.1%, dan IDSW

sebesar 7.603, yang menunjukkan bahwa ketidakstabilan deteksi memperbesar kesalahan asosiasi identitas selama proses pelacakan.

Pola hubungan antara deteksi dan pelacakan pada penelitian ini memperlihatkan bahwa performa sistem *tracking-by-detection* sangat bergantung pada kualitas deteksi awal. *MobileFaceNet* mampu menghasilkan *embedding* wajah yang diskriminatif, namun representasi tersebut tetap sensitif terhadap kesalahan posisi *bounding box*. Kesalahan kecil pada tahap deteksi dapat menyebabkan wajah terpotong secara tidak konsisten sehingga meningkatkan kemungkinan perubahan identitas pada proses asosiasi *DeepSORT*. Fenomena ini terlihat jelas pada konfigurasi dengan kualitas deteksi rendah yang menghasilkan nilai IDSW tinggi. Sebaliknya, ketika kualitas deteksi meningkat secara signifikan pada konfigurasi *epoch* 200, jumlah IDSW berhasil ditekan hingga lebih dari 75% dibandingkan konfigurasi rasio 10:90. Temuan ini memperkuat bahwa peningkatan kualitas deteksi secara langsung berkontribusi terhadap stabilitas pelacakan identitas.

Analisis tambahan terhadap konfigurasi *epoch* 300 menunjukkan bahwa nilai MOTA meningkat menjadi 84.4% dan IDF1 menjadi 82.3%, namun jumlah IDSW kembali meningkat menjadi 1.781. Meskipun secara nominal nilai akurasi pelacakan sedikit lebih tinggi dibandingkan *epoch* 200, peningkatan tersebut relatif kecil dan tidak diikuti oleh peningkatan konsistensi identitas. Fenomena ini sejalan dengan hasil analisis pada Bab V yang menunjukkan gejala *overfitting* mulai muncul setelah kisaran epoch 220-222. Ketika model deteksi mulai mengalami *overfitting*, posisi *bounding box* menjadi sedikit kurang konsisten sehingga memengaruhi proses asosiasi identitas pada *DeepSORT* dan menyebabkan

peningkatan kembali jumlah IDSW. Oleh karena itu, konfigurasi *epoch* 200 dapat dianggap sebagai titik optimum sebelum terjadinya degradasi performa pelacakan.

Dari sisi efisiensi komputasi, pengujian FPS dilakukan pada konfigurasi model terbaik untuk mengevaluasi kemampuan inferensi sistem pada kondisi performa optimal. Hasil pengujian menunjukkan rata-rata kecepatan pemrosesan berada pada kisaran 1–2 FPS pada seluruh *sequence* HT21. Variasi FPS lebih dipengaruhi oleh kompleksitas visual video uji, seperti jumlah wajah aktif, tingkat kepadatan objek, dan intensitas *overlap*, dibandingkan oleh konfigurasi parameter pelatihan model. *Sequence* dengan jumlah objek lebih sedikit menunjukkan FPS yang lebih tinggi, sedangkan lingkungan dengan tingkat kepadatan tinggi menghasilkan penurunan performa akibat meningkatnya beban komputasi pada proses deteksi, ekstraksi *embedding*, dan asosiasi identitas oleh *DeepSORT*. Temuan ini menunjukkan bahwa *bottleneck* komputasi utama terletak pada proses inferensi dan *multi-object association*, bukan pada konfigurasi pelatihan model.

Secara keseluruhan, hasil penelitian menunjukkan bahwa integrasi tahap pra-pemrosesan, YOLOv9 dengan *Attention Mechanism*, *MobileFaceNet*, dan *DeepSORT* mampu menghasilkan sistem deteksi dan pelacakan wajah yang stabil pada lingkungan keramaian padat. Berdasarkan hasil deteksi, pelacakan, dan analisis *overfitting*, konfigurasi rasio 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 merupakan konfigurasi yang paling optimal karena menghasilkan keseimbangan terbaik antara kualitas deteksi, kestabilan pelacakan identitas, dan efisiensi komputasi. Temuan ini menegaskan bahwa keberhasilan sistem *tracking-by-detection* tidak hanya ditentukan oleh kekuatan arsitektur model, tetapi juga

sangat dipengaruhi oleh kualitas distribusi data pelatihan dan strategi optimasi parameter yang digunakan selama proses pelatihan.

6.4 Integrasi Dalam Islam

Integrasi penelitian deteksi dan pelacakan wajah pada lingkungan keramaian dengan nilai-nilai Islam dapat dipahami dalam kerangka hubungan antara ilmu pengetahuan, kemaslahatan sosial, dan tanggung jawab manusia sebagai khalifah di muka bumi. Dalam perspektif Islam, pengembangan teknologi merupakan bagian dari aktivitas muamalah yang bertujuan menghadirkan manfaat (*maslahah*) serta meminimalkan kerusakan (*mafsadah*) bagi kehidupan manusia. Oleh karena itu, pengembangan sistem deteksi dan pelacakan wajah pada lingkungan keramaian tidak hanya dipahami sebagai inovasi teknis dalam bidang *computer vision*, tetapi juga sebagai bentuk ikhtiar untuk menjaga keamanan, keteraturan, dan keselamatan publik. Hal ini sejalan dengan urgensi penelitian yang telah dijelaskan pada pendahuluan, yaitu tingginya risiko kepadatan massa, kehilangan individu, gangguan keamanan, serta keterbatasan pengawasan manual dalam lingkungan keramaian padat.

Dalam perspektif Islam, ilmu pengetahuan dan teknologi juga dapat menjadi sarana untuk mengenali tanda-tanda kebesaran Allah SWT (ayat *kauniyyah*). Keunikan wajah manusia yang berbeda pada setiap individu menunjukkan kompleksitas ciptaan Allah yang luar biasa. Allah SWT berfirman dalam Surah Al-Rum ayat 22:

وَمِنْ اٰيٰتِهٖ خَلْقُ السَّمٰوٰتِ وَالْاَرْضِ وَاخْتِلَافُ اَلْسِنَتِكُمْ وَاَلْوَانِكُمْۗ اِنَّ فِيْ ذٰلِكَ لَاٰيٰتٍ لِّلْعٰلِمِيْنَ ﴿٢٢﴾

Artinya: *"Dan di antara tanda-tanda (kebesaran)-Nya ialah penciptaan langit dan bumi, perbedaan bahasamu dan warna kulitmu. Sungguh, pada yang demikian itu benar-benar terdapat tanda-tanda bagi orang-orang yang mengetahui."*

Ayat tersebut menunjukkan bahwa keberagaman karakteristik manusia, termasuk identitas visual seperti wajah, merupakan tanda kekuasaan Allah SWT yang mengandung hikmah dan keteraturan. Dalam konteks penelitian ini, kemampuan sistem untuk mengenali pola wajah secara otomatis melalui pendekatan *deep learning* dapat dipahami sebagai bagian dari pemanfaatan ilmu pengetahuan untuk memahami kompleksitas ciptaan Allah, sekaligus meningkatkan kesadaran manusia terhadap kebesaran-Nya. Implementasi nyata dari prinsip ini dapat dilihat pada pengembangan sistem pengawasan cerdas berbasis pengenalan wajah di berbagai pusat keramaian dunia, termasuk sistem *crowd management* yang diterapkan di Masjidil Haram, Mekah, sejak 2019, yang memanfaatkan teknologi *computer vision* untuk memantau kepadatan jamaah dan mencegah insiden fatal selama pelaksanaan ibadah Haji dan Umrah Alharbey et al. (2022).

Selain menunjukkan kebesaran Allah, pemanfaatan teknologi dalam Islam juga harus dilandasi rasa tanggung jawab, amanah, dan ketakwaan kepada Allah SWT. Aspek khauf (rasa takut kepada Allah) menjadi landasan moral agar teknologi tidak digunakan untuk kezaliman, penyalahgunaan identitas, diskriminasi, ataupun pelanggaran hak individu. Dalam konteks teknologi pengenalan wajah, prinsip *al-mizan* sebagaimana dijelaskan oleh Al-Ṭabari tercermin pada tuntutan akan sistem yang presisi, terukur, dan bertanggung jawab (Al-Tabari, 2007). Ketepatan model dalam mendeteksi dan melacak wajah menjadi

bentuk implementasi keadilan teknis, karena kesalahan pengukuran atau keputusan sistem dapat menimbulkan dampak nyata terhadap keselamatan, hak, dan kehormatan individu di ruang publik. Dengan demikian, capaian *precision* sebesar 95,76%, mAP@0.5 sebesar 93,74%, MOTA sebesar 73,3%, dan IDF1 sebesar 74,1% yang diperoleh dalam penelitian ini bukan sekadar ukuran teknis, melainkan juga cerminan penerapan nilai keseimbangan, tanggung jawab, dan keadilan dalam pemanfaatan teknologi.

Integrasi penelitian ini dapat diletakkan secara mendalam dalam kerangka *maqaṣid al-syari'ah*, khususnya prinsip *hifz al-nafs* (menjaga jiwa) dan *hifz al-'ird* (menjaga kehormatan). Allah SWT menegaskan kemuliaan manusia dalam firman-Nya pada Surah Al-Isra' ayat 70:

وَلَقَدْ كَرَّمْنَا بَنِيْٓ اٰدَمَ وَحَمَلْنٰهُمْ فِى الْبَرِّ وَالْبَحْرِ وَرَزَقْنٰهُمْ مِّنَ الطَّيِّبٰتِ وَفَضَّلْنٰهُمْ عَلٰى كَثِيْرٍ مِّمَّنْ خَلَقْنَا تَفْضِيْلًا ۝٧٠

Artinya: *"Sungguh, Kami telah memuliakan anak cucu Adam dan Kami angkut mereka di darat dan di laut. Kami anugerahkan pula kepada mereka rezeki dari yang baik-baik dan Kami lebihkan mereka di atas banyak makhluk yang Kami ciptakan dengan kelebihan yang sempurna."*

Menurut penjelasan dalam Tafsir al-Qur'an *al-'Azhim* karya Ibnu Katsir, kemuliaan tersebut mencakup pemberian akal, kemampuan berpikir, dan sistem sosial yang memungkinkan manusia menjaga ketertiban dan keselamatan (Kathir, 2000). Dalam konteks penelitian ini, pengembangan sistem deteksi dan pelacakan wajah berbasis YOLOv9, *attention mechanism*, dan *DeepSORT* dapat dipahami sebagai upaya menjaga kemuliaan manusia melalui perlindungan keselamatan publik. Contoh nyata urgensi ini tercermin dalam tragedi Halloween Itaewon, Seoul (2022), di mana 159 jiwa meninggal akibat kegagalan *crowd management* manual

yang tidak mampu mendeteksi kepadatan kritis secara *real-time*. Sistem deteksi berbasis kecerdasan buatan seperti yang dikembangkan dalam penelitian ini berpotensi memberikan peringatan dini sehingga insiden serupa dapat dicegah, yang selaras dengan prinsip *hifz al-nafs* dalam *maqaṣid al-syari'ah*.

Prinsip perlindungan jiwa ditegaskan pula dalam firman Allah SWT Surah Al-Ma'idah ayat 32:

مِنْ اَجْلِ ذٰلِكَ كَتَبْنَا عَلٰى بَنِيْٓ اِسْرَاۤءِيْلَ اَنَّهٗ مَنْ قَتَلَ نَفْسًاۢ بِغَيْرِ نَفْسٍ اَوْ فَسَادٍ فِى الْاَرْضِ فَكَاَنَّمَا قَتَلَ النَّاسَ جَمِيْعًاۗ وَمَنْ اَحْيَاهَا فَكَاَنَّمَآ اَحْيَا النَّاسَ جَمِيْعًاۗ وَلَقَدْ جَاۤءَتْهُمْ رُسُلُنَا بِالْبَيِّنٰتِ ثُمَّ اِنَّ كَثِيْرًا مِّنْهُمْ بَعْدَ ذٰلِكَ فِى الْاَرْضِ لَمُسْرِفُوْنَ ۝٣٢

Artinya: "*Oleh karena itu, Kami menetapkan (suatu hukum) bagi Bani Israil bahwa siapa yang membunuh seseorang bukan karena (orang yang dibunuh itu) telah membunuh orang lain atau karena telah berbuat kerusakan di bumi, maka seakan-akan dia telah membunuh semua manusia. Sebaliknya, siapa yang memelihara kehidupan seorang manusia, dia seakan-akan telah memelihara kehidupan semua manusia. Sungguh, rasul-rasul Kami benar-benar telah datang kepada mereka dengan (membawa) keterangan-keterangan yang jelas. Kemudian, sesungguhnya banyak di antara mereka setelah itu melampaui batas di bumi*"

Dalam *Al-Jami' li Aḥkam* al-Qur'an karya Al-Qurtubi, ayat ini dijelaskan sebagai landasan utama *hifz al-nafs* dalam *maqaṣid al-syari'ah*. Setiap upaya pencegahan bahaya dan penyelamatan manusia dari risiko termasuk implementasi langsung dari tujuan syariat (Al-Qurṭubi, 2003). Dalam lingkungan keramaian padat, potensi desak-desakan, kehilangan individu rentan, kepanikan massal, maupun gangguan keamanan menjadi ancaman nyata terhadap keselamatan jiwa. Oleh karena itu, sistem berbasis kecerdasan buatan yang mampu mendeteksi dan melacak individu secara *real-time* merupakan bentuk ikhtiar preventif yang sejalan dengan *maqasid*, khususnya *hifz al-nafs*.

Dalam teori *maqasid* klasik yang dirumuskan oleh Al-Ghazali, tujuan utama syariat adalah menjaga lima kebutuhan dasar manusia, yaitu agama (*hifz al-din*), jiwa (*hifz al-nafs*), akal (*hifz al-'aql*), keturunan (*hifz al-nasl*), dan harta (*hifz al-mal*) (Al-Ghazali, 2008). Kelima prinsip ini menempatkan keselamatan dan kesejahteraan manusia sebagai tujuan utama syariat, sehingga setiap upaya yang berkontribusi pada pencegahan bahaya memiliki nilai kemaslahatan yang tinggi (Kamali, 2008). Berikut diuraikan relevansi masing-masing prinsip dalam konteks penelitian ini.

Pertama, prinsip *hifz al-din* tercermin melalui dukungan teknologi terhadap keamanan dan keteraturan pelaksanaan aktivitas keagamaan pada ruang publik yang padat. Sistem pengawasan berbasis *computer vision* telah diimplementasikan secara nyata di Masjidil Haram untuk mengelola pergerakan jutaan jamaah Haji dan Umrah setiap tahunnya, sehingga pelaksanaan ibadah dapat berlangsung lebih aman dan tertib. Sistem deteksi dan pelacakan wajah yang dikembangkan dalam penelitian ini berpotensi memperkuat infrastruktur keamanan serupa pada konteks lokal, seperti shalat Idul Fitri, shalat Idul Adha, maupun kegiatan keagamaan massal lainnya.

Kedua, prinsip *hifz al-nafs* diwujudkan melalui kemampuan sistem dalam membantu deteksi dini, pemantauan berkelanjutan, dan identifikasi individu secara *real-time* guna meminimalkan risiko kehilangan orang, kepanikan massal, maupun insiden keselamatan pada lingkungan keramaian. Nilai MOTA sebesar 83,9% dan IDF1 sebesar 81,9% yang dicapai dalam penelitian ini menunjukkan kemampuan

sistem dalam menjaga konsistensi identitas objek secara efektif, yang secara langsung mendukung fungsi deteksi dini ancaman keselamatan jiwa.

Ketiga, prinsip *hifz al-'aql* tercermin pada pemanfaatan ilmu pengetahuan dan kecerdasan buatan secara rasional untuk mendukung pengambilan keputusan berbasis data dalam pengelolaan keamanan publik. Integrasi YOLOv9 dengan *attention mechanism* menghasilkan kemampuan analisis visual yang adaptif, memungkinkan petugas keamanan membuat keputusan yang lebih tepat dan terukur berdasarkan informasi *real-time* dari sistem, bukan semata mengandalkan penilaian subjektif di lapangan.

Keempat, prinsip *hifz al-nasl* berkaitan dengan kemampuan sistem dalam membantu pelacakan individu rentan, seperti anak-anak, lansia, atau penyandang disabilitas yang terpisah dari keluarga di tengah keramaian. Dalam konteks festival, konser, atau kegiatan massal, kehilangan anak di keramaian merupakan masalah nyata yang dilaporkan secara rutin. Kemampuan *DeepSORT* dalam mempertahankan konsistensi identitas objek selama pelacakan secara langsung mendukung fungsi ini, sehingga anggota keluarga yang terpisah dapat dilacak dan dipertemukan kembali dengan lebih cepat.

Kelima, prinsip *hifz al-mal* diwujudkan melalui peningkatan keamanan lingkungan publik yang dapat membantu mengurangi tindak kriminal, kehilangan barang, maupun risiko kerugian material akibat kondisi keramaian yang tidak terkendali. Kajian yang dilakukan di kota-kota yang telah mengimplementasikan sistem pengawasan berbasis AI, seperti di beberapa wilayah di London dan Shenzhen, menunjukkan penurunan signifikan angka kejahatan di area publik yang

dipantau secara cerdas (Webster, 2009). Hal ini menegaskan bahwa pengembangan sistem serupa memiliki dampak ekonomi dan keamanan yang nyata bagi masyarakat.

Selain kelima prinsip *maqasid* di atas, penelitian ini juga berkaitan erat dengan prinsip *hifz al-'ird* (menjaga kehormatan manusia), yang dalam konteks teknologi pengenalan wajah memiliki relevansi yang sangat kritis. Kehormatan manusia mencakup perlindungan identitas, martabat, dan hak individu agar tidak mengalami perlakuan tidak adil akibat kesalahan sistem. Dalam lingkungan keramaian, kesalahan deteksi (*false positive*) atau kesalahan pelacakan dapat menimbulkan salah identifikasi, stigma sosial, maupun tuduhan yang merugikan individu tertentu. Kasus-kasus salah tangkap akibat kesalahan sistem *facial recognition* telah terdokumentasi di berbagai negara, sebuah studi oleh Buolamwini dan Gebru (2018) menunjukkan bahwa sistem pengenalan wajah memiliki tingkat kesalahan yang jauh lebih tinggi pada individu berkulit gelap dan perempuan, mengindikasikan potensi diskriminasi algoritmik yang bertentangan dengan prinsip keadilan dalam Islam. Oleh karena itu, pengembangan model yang presisi dan berimbang sebagaimana dicerminkan dalam nilai *precision* 97,20 % dan *recall* 97,04% pada penelitian ini menjadi keharusan etis sekaligus teknis sebagai implementasi nyata dari *hifz al-'ird* dalam *maqaṣid al-syari'ah* (Kamali, 2008).

Meskipun teknologi pengenalan wajah memberikan manfaat yang besar, Islam juga menekankan prinsip *la dharara wa la dhirar* (tidak boleh menimbulkan mudarat dan tidak boleh membalas mudarat dengan mudarat) sebagai batasan normatif dalam pemanfaatan teknologi. Dalam konteks ini, penggunaan sistem

deteksi dan pelacakan wajah harus disertai dengan tata kelola yang bertanggung jawab, termasuk perlindungan data biometrik, pembatasan akses, transparansi penggunaan, serta persetujuan dari pihak yang dipantau dalam konteks yang relevan. Prinsip ini penting untuk mencegah teknologi pengenalan wajah disalahgunakan sebagai instrumen pengawasan represif yang melanggar hak privasi individu. Habib (2025) mengembangkan kerangka etika AI berbasis *Maqasid al-Shariʿa* yang menekankan perlindungan lima nilai dasar sebagai landasan normatif dalam merespons tantangan bias algoritmik, pelanggaran privasi, dan problem akuntabilitas sistem AI. Dalam kerangka ini, sistem yang dikembangkan pada penelitian ini harus dipahami sebagai alat yang difungsikan untuk kemaslahatan publik dalam batas-batas etika yang jelas, bukan sebagai instrumen pengawasan tanpa batas.

Nilai tanggung jawab sosial juga ditegaskan dalam hadits Rasulullah SAW yang diriwayatkan oleh Bukhari dan Muslim:

الْمُسْلِمُ أَخُو الْمُسْلِمِ لَا يَظْلِمُهُ وَلَا يُسْلِمُهُ، مَنْ كَانَ فِي حَاجَةِ أَخِيهِ كَانَ اللَّهُ فِي حَاجَتِهِ، وَمَنْ فَرَّجَ عَنْ مُسْلِمٍ كُرْبَةً فَرَّجَ اللَّهُ عَنْهُ بِهَا كُرْبَةً مِنْ كُرَبِ يَوْمِ الْقِيَامَةِ، وَمَنْ سَتَرَ مُسْلِمًا سَتَرَهُ اللَّهُ يَوْمَ الْقِيَامَةِ

Artinya : *"Seorang muslim adalah saudara bagi muslim lainnya, dia tidak menzaliminya dan tidak membiarkannya (disakiti/dizalimi). Barangsiapa yang membantu memenuhi kebutuhan saudaranya, maka Allah akan memenuhi kebutuhannya. Barang siapa yang menghilangkan satu kesusahan seorang muslim, maka Allah akan menghilangkan darinya satu kesusahan dari kesusahan-kesusahan di hari kiamat. Dan barang siapa yang menutupi (aib) seorang muslim, maka Allah akan menutupi aibnya pada hari kiamat"*

Penjelasan dalam *Fath al-Bari* karya Ibn Hajar al-Asqalani menegaskan bahwa larangan 'tidak membiarkannya' bermakna kewajiban aktif mencegah bahaya terhadap sesama (Ibn Ḥajar, 2001). Dengan demikian, sistem pengawasan

berbasis algoritma dalam penelitian ini dapat dipandang sebagai bentuk tanggung jawab kolektif (*fardhu kifayah*) untuk mengurangi risiko keselamatan dan memberikan manfaat nyata bagi masyarakat. Lebih dari itu, frasa "menutupi aib seorang muslim" dalam hadits ini juga mengandung dimensi etika privasi yang relevan: sistem yang dikembangkan tidak boleh digunakan untuk membuka, menyebarkan, atau menyalahgunakan data identitas individu, melainkan semata-mata diarahkan untuk perlindungan keselamatan publik.

Konsep *maqasid* modern sebagaimana dikembangkan oleh Jasser Auda menekankan pendekatan sistem (*systems approach*) dalam memahami syariat sebagai instrumen kemaslahatan sosial yang dinamis (Auda, 2008). Teknologi dalam kerangka ini bukan sekadar inovasi teknis, tetapi sarana menjaga kesejahteraan manusia secara menyeluruh. Sejalan dengan itu, Mohammad Hashim Kamali menjelaskan bahwa *maqaṣid al-syari'ah* melindungi kepentingan dasar manusia, termasuk jiwa dan kehormatan, dalam konteks kehidupan kontemporer yang terus berkembang (Kamali, 2008). Keduanya menekankan bahwa penerapan *maqasid* tidak bersifat statis, melainkan harus merespons tantangan zaman termasuk revolusi kecerdasan buatan yang tengah berlangsung saat ini.

Dengan demikian, penelitian ini tidak hanya memiliki kontribusi teknis dalam pengembangan sistem deteksi dan pelacakan wajah berbasis YOLOv9, *attention mechanism*, dan *DeepSORT*, tetapi juga memiliki legitimasi normatif yang kuat dalam perspektif Islam. Capaian *precision* 97,20%, *recall* 97,04%, *mAP@0.5* sebesar 97,33%, *MOTA* 83,9%, dan *IDF1* 81,9% yang diperoleh merupakan wujud nyata dari prinsip *al-mizan* yaitu presisi yang terukur sebagai bentuk keadilan teknis

dalam menjaga keselamatan dan kehormatan individu di ruang publik. Pemanfaatan teknologi dalam penelitian ini diarahkan sebagai sarana kemaslahatan untuk menjaga keselamatan jiwa (*hifz al-nafs*), kehormatan identitas (*hifz al-'ird*), keamanan ibadah (*hifz al-din*), perlindungan keluarga rentan (*hifz al-nasl*), dan keamanan harta benda (*hifz al-mal*), sekaligus dibatasi oleh prinsip *la dharar* agar tidak menjadi instrumen kezaliman. Dengan mengintegrasikan keunggulan teknis dan landasan normatif Islam, penelitian ini diharapkan berkontribusi nyata pada pengembangan teknologi kecerdasan buatan yang bertanggung jawab, berkeadilan, dan berpihak pada kemaslahatan umat.

BAB VII
KESIMPULAN DAN SARAN

7.1 Kesimpulan

Penelitian ini bertujuan untuk merancang dan menguji sistem deteksi serta pelacakan wajah pada lingkungan keramaian berbasis *deep learning* menggunakan YOLOv9 dengan *Attention Mechanism* untuk deteksi wajah, serta integrasi *MobileFaceNet* dan *DeepSORT* untuk pelacakan identitas. Berdasarkan rangkaian eksperimen yang meliputi tahap pra-pemrosesan, variasi rasio data pelatihan, optimasi parameter pelatihan model, serta evaluasi performa deteksi dan pelacakan, diperoleh beberapa kesimpulan sebagai berikut.

1. Tahap pra-pemrosesan berhasil meningkatkan kualitas dan konsistensi data masukan sehingga mendukung kestabilan proses pelatihan pada lingkungan keramaian dengan kondisi visual yang kompleks. Evaluasi kualitas citra menunjukkan nilai PSNR pada rentang 33.72–41.00 dB, SSIM sebesar 0.9719–0.9888, serta *Detection Score* (DET) pada rentang 45.92–207.59. Hasil tersebut menunjukkan bahwa struktur visual wajah tetap terjaga dan layak digunakan sebagai masukan sistem deteksi wajah.
2. Implementasi YOLOv9 dengan *Attention Mechanism* menunjukkan bahwa performa deteksi sangat dipengaruhi oleh distribusi data dan konfigurasi parameter pelatihan. Konfigurasi terbaik diperoleh pada rasio data 90:10, *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 dengan nilai *precision* sebesar 97.20%, *recall* sebesar 97.04%, F1-score sebesar 97.12%, dan mAP@0.5 sebesar 97.33%. Hasil ini menunjukkan bahwa kombinasi jumlah data latih yang memadai, durasi pelatihan yang cukup panjang, serta

laju pembelajaran yang lebih kecil mampu meningkatkan kemampuan generalisasi model dalam mendeteksi wajah pada lingkungan keramaian padat. Analisis kurva pelatihan juga menunjukkan bahwa *epoch* 200 merupakan titik optimum sebelum model mulai menunjukkan kecenderungan *overfitting* pada pelatihan yang lebih panjang.

3. Pada tahap pelacakan, integrasi *MobileFaceNet* dan *DeepSORT* berhasil mempertahankan konsistensi identitas wajah antar-frame pada lingkungan keramaian yang dinamis. Konfigurasi terbaik menghasilkan nilai *MOTA* sebesar 83.9%, IDF1 sebesar 81.9%, serta jumlah *Identity Switches* (*IDSW*) sebesar 1.766. Hasil tersebut menunjukkan bahwa kualitas deteksi awal memiliki pengaruh langsung terhadap kestabilan pelacakan, karena *bounding box* yang lebih presisi menghasilkan *embedding* wajah yang lebih representatif dan mempermudah proses asosiasi identitas oleh *DeepSORT*.
4. Evaluasi performa komputasi menunjukkan bahwa sistem memiliki kompleksitas inferensi yang relatif tinggi dengan rata-rata FPS berada pada kisaran 1.45–2.38 FPS pada *sequence* HT21. Variasi performa FPS lebih dipengaruhi oleh tingkat kepadatan objek, jumlah wajah aktif, dan kompleksitas proses asosiasi identitas dibandingkan parameter pelatihan model. Meskipun belum mencapai standar real-time konvensional, sistem masih relevan digunakan untuk kebutuhan *crowd surveillance*, *monitoring* kepadatan, serta analisis pelacakan identitas berbasis interval *frame*.

Secara keseluruhan, penelitian ini berhasil menjawab rumusan masalah dengan merancang dan menguji sistem deteksi serta pelacakan wajah pada

lingkungan keramaian menggunakan integrasi YOLOv9 dengan *Attention Mechanism*, *MobileFaceNet*, dan *DeepSORT*. Sistem yang dikembangkan mampu mendeteksi wajah secara akurat pada kondisi kepadatan tinggi, *occlusion*, dan variasi pencahayaan, sekaligus mempertahankan identitas wajah secara berkelanjutan antar-*frame* video. Hasil evaluasi menunjukkan bahwa konfigurasi rasio 90:10 dengan *epoch* 200, *batch size* 16, dan *learning rate* 0.0005 memberikan performa paling optimal, sehingga membuktikan bahwa *Attention Mechanism* mampu meningkatkan kualitas ekstraksi fitur wajah, sedangkan *MobileFaceNet* dan *DeepSORT* mampu menjaga kontinuitas identitas secara efektif pada lingkungan keramaian padat.

7.2 Saran

Berdasarkan hasil penelitian yang telah diperoleh, beberapa saran yang dapat dikembangkan pada penelitian selanjutnya adalah sebagai berikut.

1. Optimalisasi Kecepatan Komputasi perlu dilakukan pada arsitektur YOLOv9, *MobileFaceNet*, dan *DeepSORT* melalui teknik seperti *model pruning*, *quantization*, *knowledge distillation*, atau penggunaan arsitektur yang lebih ringan untuk meningkatkan performa inferensi. Mengingat sistem masih menghasilkan rata-rata FPS pada kisaran 1-2 FPS, peningkatan efisiensi komputasi diperlukan agar sistem dapat mendukung implementasi *real-time* pada lingkungan keramaian yang dinamis.
2. Eksplorasi Strategi Pelatihan yang Lebih Adaptif dapat dilakukan melalui penerapan *learning rate scheduler*, *early stopping*, *warm-up training*, atau

metode optimasi lainnya untuk memperoleh titik konvergensi yang lebih optimal. Hasil penelitian menunjukkan bahwa peningkatan jumlah *epoch* mampu meningkatkan performa hingga titik tertentu, namun pelatihan yang terlalu panjang berpotensi menimbulkan gejala *overfitting*.

3. Evaluasi pada Lingkungan Komputasi yang Beragam perlu dilakukan menggunakan perangkat dengan spesifikasi berbeda, seperti GPU kelas tinggi, *edge computing*, maupun perangkat *embedded* seperti NVIDIA *Jetson*. Pengujian tersebut penting untuk mengevaluasi skalabilitas sistem dan kesiapannya pada implementasi lapangan yang memiliki keterbatasan sumber daya komputasi.
4. Pengayaan dan Diversifikasi Dataset disarankan dengan menambahkan variasi kondisi pencahayaan ekstrem, sudut pandang kamera yang lebih beragam, resolusi rendah, pergerakan kamera dinamis, serta tingkat *occlusion* yang lebih tinggi. Pengayaan data diharapkan dapat meningkatkan kemampuan generalisasi model pada kondisi keramaian yang lebih kompleks dibandingkan dataset HT21.
5. Pengembangan Modul *Re-Identification* dapat dilakukan dengan mengeksplorasi model pengenalan wajah yang lebih kuat, seperti *ArcFace*, *AdaFace*, *RetinaFace*, maupun pendekatan berbasis *Transformer* dan *attention-enhanced embedding network*. Pendekatan tersebut berpotensi meningkatkan kualitas representasi identitas wajah sehingga nilai IDF1 dapat ditingkatkan dan jumlah *Identity Switches* dapat dikurangi pada kondisi kepadatan yang sangat tinggi.

6. Pengembangan Sistem *End-to-End* yang Lebih Terintegrasi perlu dilakukan dengan menggabungkan proses deteksi, ekstraksi fitur, dan pelacakan dalam satu *pipeline* inferensi yang lebih efisien. Integrasi tersebut diharapkan mampu mengurangi latensi pemrosesan, meningkatkan efisiensi penggunaan memori, serta mendukung implementasi sistem secara *semi real-time* maupun *real-time*.
7. Pengembangan Analisis Kerumunan Berbasis Pelacakan Wajah dapat menjadi arah penelitian lanjutan dengan memanfaatkan hasil pelacakan untuk menganalisis kepadatan, pola pergerakan, identifikasi area rawan penumpukan massa, maupun deteksi perilaku abnormal pada lingkungan keramaian. Pengembangan ini dapat memperluas manfaat sistem dari sekadar deteksi dan pelacakan wajah menjadi sistem pendukung pengambilan keputusan pada pengelolaan kerumunan.

Dengan pengembangan lebih lanjut, sistem deteksi dan pelacakan wajah pada lingkungan keramaian ini berpotensi diterapkan pada pengawasan area publik, manajemen kerumunan, sistem keamanan berbasis video, monitoring kepadatan, serta analisis perilaku kerumunan secara otomatis dan berkelanjutan pada lingkungan dengan mobilitas tinggi seperti transportasi publik, acara massal, pusat perbelanjaan, maupun area ibadah berskala besar.

DAFTAR PUSTAKA

Alansari, M., Alnuaimi, K., Ganapathi, I., Alansari, S., Javed, S., Shoufan, A., Zweiri, Y., & Werghi, N. (2025). EfficientFaceV2S: A lightweight model and a benchmarking approach for drone-captured face recognition. *Expert Systems With Applications*, 126786. doi.org/10.1016/j.eswa.2025.126786

Alharbey, R., Banjar, A., Said, Y., Atri, M., Alshdadi, A., & Abid, M. (2022). Human faces detection and tracking for crowd management in Hajj and Umrah. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *71*(3), 6275–6291. doi.org/10.32604/cmc.2022.024272

Almuqren, L., Hamza, M. A., Mohamed, A., & Abdelmageed, A. A. (2023). Automated Video-Based Face Detection Using Harris Hawks Optimization with Deep Learning. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *75*(3), 4917–4933. doi.org/10.32604/cmc.2023.037738

Al-Ghazali, I. (2008). *Al-Mustashfa Jilid 1: Rujukan Utama Ushul Fikih*. Pustaka Al-Kautsar.

Al-Qurṭūbī, M. B. A. (2003). *Al-Jāmi' Li Aḥkām Al-Qurān (H. S. Al-Bukhari, Ed.)*. Arab Saudi: Dar Alim Al-Kutub.

Alsabei, A. A., Alsubait, T. M., & Alhakami, H. H. (2025). Enhancing crowd safety at Hajj: Real-Time Detection of Abnormal Behavior using YOLOV9. *IEEE Access*, 1. doi.org/10.1109/access.2025.3545256

Al-Tabari, M. B. J. (2007). *Tafsir al-Tabari: Jamiʿ al-bayan ʿan taʾwil ay al-Qurʾan (Terjemahan)*. Jakarta: Pustaka Azzam.

Auda, J. (2008). *Maqasid Al-Shariah as Philosophy of Islamic Law: A Systems Approach*. International Institute of Islamic Thought (IIIT).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *arXiv.org*. arxiv.org/abs/1409.0473v7

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. doi.org/10.1109/icip.2016.7533003

Cahyani, P. A., Mardiana, M., Wintoro, P. B., & Muhammad, M. A. (2024). Sistem Perhitungan Kendaraan Menggunakan Algortima YOLOv5 dan *DeepSORT*. *Jurnal Teknik Informatika Dan Sistem Informasi*, *10*(1). doi.org/10.28932/jutisi.v10i1.7519

Chusna, N. L., & Khumaidi, A. (2024). Comparison of convolutional neural network models for feasibility of selling orchids. *ILKOM Jurnal Ilmiah*, *16*(3), 296–304. doi.org/10.33096/ilkom.v16i3.2006.296-304

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2015). ARCFaCE: Additive angular margin loss for deep face recognition. In *JOURNAL OF LATEX CLASS FILES* (Vol. 14, Issue 8, pp. 1–2).

Djarah, D., Benmakhlouf, A., Zidani, G., & Khettache, L. (2024). Online Multi-object Tracking with YOLOv9 and *DeepSORT* Optimized by Optical Flow. *Engineering Technology & Applied Science Research*, *14*(6), 17922–17930. doi.org/10.48084/etasr.8770

Fabrianto, L., Prihandayani, T. W., Rasenda, R., & Faizah, N. M. (2025). Attention-based convolutional neural networks for interpretable classification of maritime equipment. *ejournal.isha.or.id*. doi.org/10.35335/mandiri.v14i1.426

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R. R., Cheng, M., & Hu, S. (2022). *Attention mechanism*s in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. doi.org/10.1007/s41095-022-0271-y

Habib, Z. (2025). Ethics of Artificial Intelligence in Maqāṣid Al-Sharīa's perspective. *KARSA Journal of Social and Islamic Culture*, *33*(1), 105–134. doi.org/10.19105/karsa.v33i1.19617

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv (Cornell University)*. doi.org/10.48550/arxiv.1704.04861

Ibn Ḥajar, A.-'Asqalani. (2001). *Fatḥ al-Bārī bi Syarḥ Ṣaḥīḥ al-Bukhārī (Vol. 5)*. Dār al-Salām.

Kamali, M. H. (*2008). Maqasid al-Shariah: A Beginner's Guide (2nd ed.). Internatio*nal Institute of Islamic Thought (IIIT).

Kathir, I. (2000). *Tafsīr al-Qur'an al-'Aẓim (Vol. 5)*. Dar Ṭayyibah.

Kurniawan, F., Astawa, I. N. G. A., Atmaja, I. M. a. D. S., & Wibawa, A. P. (2023). Facemask Detection using the YOLO-v5 Algorithm: Assessing Dataset Variation and R esolutions. *Register Jurnal Ilmiah Teknologi Sistem Informasi*, *9*(2), 95–102. doi.org/10.26594/register.v9i2.3249

Narayanan, S. P., Manikandan, M. S., & Cenkeramaddi, L. R. (2025). YOLOV9-Based human face detection and counting under Human-Animal faces, complex imaging environments, and image qualities. *IEEE Access*, 1. doi.org/10.1109/access.2025.3591247

Nasir, R., Jalil, Z., Nasir, M., Alsubait, T., Ashraf, M., & Saleem, S. (2025). An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8. *Artificial Intelligence Review*, *58*(7). https://doi.org/10.1007/s10462-025-11206-w

Ngeni, F., Mwakalonge, J., & Siuhi, S. (2024). Solving traffic data occlusion problems in computer vision algorithms using *DeepSORT* and quantum computing. *Journal of Traffic and Transportation Engineering (English Edition)*, *11*(1), 1–15. doi.org/10.1016/j.jtte.2023.05.006

Salman, H. A., & Kalakech, A. (2024). Image Enhancement using Convolution Neural Networks. *Babylonian Journal of Machine Learning*, *2024*, 30–47. doi.org/10.58496/bjml/2024/003

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. (2018). *MobileNetV2: Inverted Residuals and Linear Bottle*neck*s* (pp. 4510–4520). doi.org/10.1109/cvpr.2018.00474

Shukla, R. K., & Tiwari, A. K. (2022). Masked Face Recognition Using MobileNet V2 with Transfer Learning. *Computer Systems Science and Engineering*, *45*(1), 293–309. doi.org/10.32604/csse.2023.027986

Singh, P., & Reibman, A. R. (2024). Task-aware image quality estimators for face detection. *EURASIP Journal on Image and Video Processing*, *2024*(1). doi.org/10.1186/s13640-024-00660-1

Somoal, M. G., & Dzikrillah, A. R. (2025). Komparasi MobileNETV2 dengan Kustomisasi Transfer Learning dan Hyperparameter untuk Identifikasi Tumor Otak. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *12*(1), 229–240. doi.org/10.25126/jtiik.2025129582

Sundararaman, R., De Almeida Braga, C., Marchand, E., & Pettre, J. (2021). Tracking Pedestrian Heads in Dense Crowd. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 3864–3874. doi.org/10.1109/cvpr46437.2021.00386

Susetianingtias, D. T., Patriya, E., & Arianty, R. (2023). Combination of YOLOV3 algorithm and Blob detection technique in calculating nile tilapia seeds. *ILKOM Jurnal Ilmiah*, *15*(2), 317–325. doi.org/10.33096/ilkom.v15i2.1634.317-325

Thaher, T., Saffarini, M., Mafarja, M., Alashbi, A., Mohamed, A. H., & El-Saleh, A. A. (2025). A hybrid approach for heavily occluded face detection using histogram of oriented gradients and deep learning models. *Computer Modeling in Engineering & Sciences*, *144*(2), 2359–2394. doi.org/10.32604/cmes.2025.065388

Tommy, Siregar, R., & Syahputra, E. R. (2025). Low-Resolution face image reconstruction using Multi-Stage FSRCNN to improve face detection and tracking accuracy in CCTV surveillance. *JOIV International Journal on Informatics Visualization*, *9*(3), 1022. doi.org/10.62527/joiv.9.3.3160

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Google Brain, Google Research, Gomez, A. N., University of Toronto, Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. arxiv.org/pdf/1706.03762.pdf

Wang, Y., Li, Y., & Zou, H. (2023). Masked face recognition system based on *attention mechanism*. *Information*, *14*(2), 87. doi.org/10.3390/info14020087

Wei, Y., Li, H., He, Y., Li, L., Lyu, Q., & Yang, Y. (2025). Robust face mask detection in complex scenarios using YOLOv8 and context-aware convolutions. *Scientific Reports*, *15*(1). doi.org/10.1038/s41598-025-04768-w

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. doi.org/10.1109/icip.2017.8296962

Xiao, J., Jiang, G., & Liu, H. (2022). A Lightweight Face Recognition Model based on MobileFaceNet for Limited Computation Environment. *EAI Endorsed Transactions on Internet of Things*, *7*(27), 1–9. doi.org/10.4108/eai.28-2-2022.173547

Xue, S. (2025). Facial recognition for surveillance videos based on RetinaFace and MobileFaceNet. *Proceedings Volume 13513, the International Conference Optoelectronic Information and Optical Engineering (OIOE2024)*, 155. doi.org/10.1117/12.3048760

Zhang, G., Zhang, S., & He, X. (2024). Research on improved MobileFaceNet facial recognition algorithm. *Proceedings Volume 13395, International Conference on Optics, Electronics, and Communication Engineering (OECE 2024)*, 7. doi.org/10.1117/12.3048257