EVALUASI KEPUASAN PELANGGAN BERDASARKAN EKSPRESI WAJAH MENGGUNAKAN *REAL TIME DETECTION TRANSFORMER* (RT-DETR)

SKRIPSI

diajukan untuk memenuhi bagian dari syarat memperoleh gelar sarjana komputer pada Program Studi Ilmu Komputer

Oleh:

Shafa Meira Wahyono

2007723

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

2025

EVALUASI KEPUASAN PELANGGAN BERDASARKAN EKSPRESI WAJAH MENGGUNAKAN *REAL TIME DETECTION TRANSFORMER* (RT-DETR)

Oleh

Shafa Meira Wahyono

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memeroleh gelar Sarjana pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

Januari 2025

SHAFA MEIRA WAHYONO

EVALUASI KEPUASAN PELANGGAN BERDASARKAN EKSPRESI WAJAH MENGGUNAKAN *REAL TIME DETECTION TRANSFORMER* (RT-DETR)

disetujui dan disahkan oleh pembimbing:

Pembimbing I

Prof. Dr. Munir, M. IT.

NIP. 196603252001121001

Pembimbing II

Yaya Wihardi, M. Kom.

NIP. 198903252015041001

Mengetahui,

Ketua Program Studi Ilmu Komputer

Dr. Muhammad Nursalman, M.T.

NIP. 197909292006041002

PERNYATAAN BEBAS PLAGIARISME

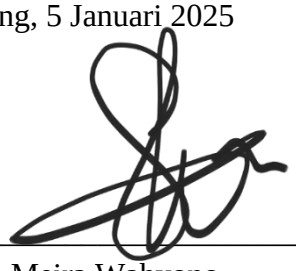
Saya yang bertanda tangan di bawah ini:

Nama	: Shafa Meira Wahyono
NIM	: 2007723
Program Studi	: Ilmu Komputer
Judul Karya	: Evaluasi Kepuasan Pelanggan Berdasarkan Ekspresi Wajah Menggunakan *Real Time Detection Transformer* (RT-DETR)

Dengan ini menyatakan bahwa karya tulis ini merupakan hasil kerja saya sendiri. Saya menjamin bahwa seluruh isi karya ini, baik sebagian maupun keseluruhan, bukan merupakan plagiarisme dari karya orang lain, kecuali pada bagian yang telah dinyatakan dan disebutkan sumbernya dengan jelas.

Jika di kemudian hari ditemukan pelanggaran terhadap etika akademik atau unsur plagiarisme, saya bersedia menerima sanksi sesuai peraturan yang berlaku di Universitas Pendidikan Indonesia.

Bandung, 5 Januari 2025

Tanda tangan: 

Shafa Meira Wahyono

KATA PENGANTAR

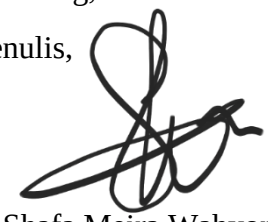
Puji dan syukur penulis panjatkan kepada Allah SWT karena telah memberikan berkah dan rahmat-Nya. Shalawat dan salam semoga terlimpahcurahkan kepada baginda tercinta kita yaitu Nabi Muhammad SAW. Tanpa pertolongan-Nya, tentu penulis tidak akan sanggup untuk menyelesaikan skripsi yang berjudul "Evaluasi Kepuasan Pelanggan Berdasarkan Ekspresi Wajah Menggunakan *Real Time Detection Transformer* (RT-DETR)" dengan baik.

Penulisan skripsi ini memiliki tujuan sebagai salah satu syarat untuk memeroleh gelar Sarjana Ilmu Komputer (S.Kom) pada jenjang studi Strata-1 pada Program Studi Ilmu Komputer di Universitas Pendidikan Indonesia.

Dalam penyusunan skripsi ini, penulis telah berusaha semaksimal kemampuan yang dimiliki. Namun tidak bisa dipungkiri sebagai manusia, penulis pun tidak luput dari kesalahan. Kesalahan tersebut bisa berupa isi, tata bahasa, tanda baca, dan lain-lain. Oleh karena itu, penulis mengharapkan kritik dan saran dari pembaca, agar skripsi ini dapat menjadi lebih baik lagi. Dengan demikian, semoga skripsi ini bisa bermanfaat bagi penulis dan pembaca. Terima kasih

Bandung, 5 Januari 2025

Penulis,



Shafa Meira Wahyono

UCAPAN TERIMA KASIH

Alhamdulillahirabbil'alamin, puji dan syukur atas kehadirat Allah SWT karena telah memberikan berkah dan rahmat-Nya dalam proses penulisan skripsi ini. Pada proses penulisan skripsi ini, penulis menerima banyak bantuan serta dukungan dari Allah SWT, keluarga, serta banyak pihak lain. Dengan demikian, sudah sepantasnya penulis mengucapkan terima kasih serta penghargaan yang pantas, kepada:

1. Kedua orang tua serta adik penulis yang selalu memberikan doa, dukungan, bantuan, serta semangat kepada penulis dalam menjalankan perkuliahan dan penulisan skripsi ini.
2. Bapak Prof. Dr. Munir, M. IT. selaku dosen pembimbing I yang senantiasa membimbing dan memberi masukan yang bermanfaat untuk penulis.
3. Bapak Yaya Wihardi, M. Kom. selaku dosen pembimbing II yang senantiasa membimbing dan memberi masukan serta saran yang bermanfaat selama penulisan skripsi ini.
4. Ibu Rosa Ariani Sukamto, M.T. selaku dosen pembimbing akademik yang telah memberikan semangat dan motivasi serta saran selama perkuliahan penulis.
5. Ibu Dr. Rani Megasari, M.T., selaku demisioner ketua Departemen Ilmu Komputer Universitas Pendidikan Indonesia.
6. Bapak Dr. Muhammad Nursalman, M.T. selaku ketua Program Studi Ilmu Komputer Universitas Pendidikan Indonesia.
7. Seluruh jajaran dosen serta staff Departemen Pendidikan Ilmu Komputer Universitas Pendidikan Indonesia yang tidak bisa penulis tuliskan satu-persatu yang telah senantiasa membantu, mengarahkan, serta membimbing penulis selama menempuh pendidikan Sarjana.
8. Sahabat penulis, Alfi Amaliandini, yang senantiasa memberikan bantuan, motivasi, dan dukungan kepada penulis dari sekolah menengah pertama hingga pendidikan Sarjana.
9. Teman-teman magang PDDIKTI yang senantiasa memberikan bantuan dan dukungan kepada penulis selama penulisan skripsi serta telah membantu dan memberikan pelajaran yang berharga bagi penulis selama kegiatan magang.

10. Teman-teman Ilmu Komputer angkatan 2020 yang telah berjuang bersama penulis dalam kegiatan perkuliahan maupun luar perkuliahan.
11. Rekan-rekan Badan Eksekutif Mahasiswa periode 2022-2023 yang telah memberikan kepercayaan kepada penulis untuk menjadi Ketua Biro Teknologi dan telah memberikan pengalaman serta pengetahuan baru dalam bidang organisasi.
12. Rekan-rekan kelompok persiapan Duolingo English Test untuk persiapan mengikuti kegiatan IISMA yang telah berbagi semangat serta ilmu yang bermanfaat.
13. Serta seluruh pihak lain yang tidak dapat penulis sebutkan yang telah membantu serta memberikan dukungan dalam perkuliahan serta penulisan skripsi ini hingga selesai.

Akhir kata, penulis berharap agar skripsi ini dapat memberikan manfaat serta menambah ilmu bagi para pembaca. Sekadar tulisan tidak dapat menggambarkan rasa terima kasih penulis kepada seluruh pihak yang telah mendukung. Semoga Tuhan Yang Maha Esa senantiasa memberikan kebahagiaan dan kesehatan untuk sekarang dan seterusnya. *Aamiin ya rabbal alamin.*

EVALUASI KEPUASAN PELANGGAN BERDASARKAN EKSPRESI WAJAH MENGGUNAKAN *REAL TIME DETECTION TRANSFORMER* (RT-DETR)

Oleh
Shafa Meira Wahyono – shafameira@upi.edu
2007723

ABSTRAK

Ekspresi wajah merupakan indikator paling baik dalam mengetahui perasaan manusia karena hanya teridentifikasi dalam waktu singkat yaitu 0,5 detik. Maka dari itu, ekspresi wajah dapat digunakan sebagai indikator evaluasi kepuasan pelanggan. Namun, karena perubahan ekspresi wajah yang cepat, penggunaan teknologi untuk pengenalan wajah serta klasifikasi ekspresi menjadi lebih ideal. Selain itu, pengimplementasian teknologi di dunia nyata sering kali dihadapi tantangan, seperti keterbatasan perangkat keras. Berdasarkan permasalahan tersebut, dilakukan penelitian untuk membangun model kecerdasan buatan yang mampu mengevaluasi kepuasan pelanggan berdasarkan pengenalan ekspresi wajah dengan memerhatikan kemungkinan keterbatasan sumber daya yang dapat muncul ketika pengimplementasian. Pertama, model dikembangkan untuk mendeteksi wajah menggunakan metode RT-DETR (*Real Time-Detection Transformer*) dengan *backbone* ResNet-18 dan LCNet-0.25 (*Lightweight CPU Convolutional Neural Network*). Kedua, model mengklasifikasikan 2 tipe ekspresi wajah, yaitu 7 ekspresi wajah dan 3 ekspresi wajah, menggunakan metode *Real-Time* CNN. Penelitian yang dilakukan, menunjukan bahwa metode RT-DETR dengan backbone ResNet-18 mendapatkan kinerja terbaik dengan 35.0% AP, 6.64 FPS, dan parameter 20M serta *Real-Time* CNN dengan 3 ekspresi wajah dengan performa *micro-average* F1-score 68.4%. Kombinasi dari model RT-DETR dan model *Real-Time* CNN memiliki kinerja 2.1% AP dan 4.7 FPS.

Kata Kunci: Detection Transformer, Pengenalan Ekspresi Wajah, Real Time.

CUSTOMER SATISFACTION EVALUATION THROUGH FACIAL EXPRESSION USING REAL TIME DETECTION TRANSFORMER (RT-DETR)

Arranged by
Shafa Meira Wahyono – shafameira@upi.edu
2007723

ABSTRACT

Facial expression is the best indicator to recognize emotion because it is only recognizable in a such short time, which is about 0,5 seconds. Therefore, facial expression is usable as customer satisfaction evaluation indicator. However, as facial expression changes quickly, it would be more ideal to use technology to recognize the faces and classify the expression. Furthermore, technology implementation in real-world was often meet with computation limitation. Based on those problems, this research is done to build Artificial Intelligence model that can evaluate customer satisfaction through facial expression, while keeping attention on the possibility of resource limitation when implementing. Firstly, a model is developed to detect faces using methods, such as RT-DETR (Real Time-Detection Transformer) with backbone ResNet-18 and LCNet-0.25 (Lightweight CPU Convolutional Neural Network. Secondly, a model will classify 2 types of facial expression, such as 7 type of facial expression and 3 type of facial expression, using Real-Time CNN method. RT-DETR with ResNet-18 as backbone achieves best performance with 35.0% AP, 6.64 FPS, and parameter of 20M, while Real-Time CNN with 3 types facial expression achieves micro-average F1-score of 68.4%. The combination of RT-DETR and Real Time-CNN achieves 4.7% AP dan 5.94 FPS.

Keywords: Customer Satisfaction, Detection Transformer, Facial Expression Recognition, Real Time

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR SINGKATAN

AI	: *Artificial Intelligence*
ANN	: *Artificial Neural Network*
AP	: *Average Precision*
BBOX	: *Bounding Box*
CCTV	: *Closed-Circuit Television*
CNN	: *Convolutional Neural Network*
COCO	: *Common Objects in Context Dataset*
CPU	: *Central Processing Unit*
DETR	: *Detection Transformer*
DINO	: DETR *with Improved DeNoising Anchor*
FACS	: *Facial Action Coding System*
FER	: *Facial Expression Recognition*
FER-2013	: *Facial Expression Recognition*-2013
FFN	: *Feed Forward Network*
FPS	: *Frame per Second*
GPU	: *Graphics Processing Unit*
IMED	: *Indonesian Mixed Emotion Dataset*
IoU	: *Intersection over Union*
JSON	: JavaScript *Object Notation*
JST	: Jaringan Syaraf Tiruan
KB	: *Kilo Bytes*
LCNet-0.25	: *Lightweight CPU Convolutional Neural Network* dengan skala 0.25
L-DETR	: *Lightweight Detection Transformer*
M	: *Millions*
MB	: *Mega Bytes*
MKLDNN	: *Math Kernel Library for Deep Neural Networks*
NLP	: *Natural Language Processing*
OCR	: *Optical Character Recognition*

ONNX	: *Open Neural Network Exchange*
PP-LCNET	: *PaddlePaddle*-LCNet
Real Time-CNN	: *Real Time-Convolutional Neural Network*
ResNet-18	: *Residual Network* dengan 18 *layers*
ROI	: *Region of Interest*
RT-DETR	: *Real Time Detection Transformer*
SMOTE	: *Synthetic Minority Over-sampling Technique*
YAML	: *Yet Another Markup Language*
YOLO	: *You Only Look Once*

BAB I
PENDAHULUAN

1.1. Latar Belakang

Wajah dipercaya mampu menyampaikan emosi yang dirasakan oleh seseorang (Revina & Emmanuel, 2021). Menurut Darwin, beberapa tindakan yang umumnya dilakukan karena kebiasaan, dapat dikendalikan hingga batas tertentu (Darwin, 1898, hlm. 48-49). Selanjutnya, teori tersebut dikembangkan lebih jauh oleh (Ekman & Friesen, 1969, hlm. 93) dan mengusulkan bahwa wajah, tangan, dan kaki merupakan bagian tubuh yang tindakannya bertentangan dengan pendapat Darwin.

Dalam penelitiannya mengenai relasi ekspresi wajah dan pergerakan tubuh dengan psikoterapi, dinyatakan bahwa wajah memiliki visibilitas terbaik dalam penyampaian ekspresi dibandingkan dengan tangan atau kaki. Hal tersebut karena umumnya wajah hanya bisa ditutupi oleh riasan wajah atau rambut yang tidak menghalangi dalam pendeteksian ekspresi. Berbeda dengan tangan dan kaki yang bisa tertutup oleh objek lain, seperti pakaian, hal ini menjadikan wajah lebih bisa diandalkan dalam mengidentifikasi ekspresi pada seseorang. Dalam keperluan pengenalan ekspresi wajah, kecepatan mendeteksi merupakan aspek krusial karena ekspresi wajah yang mudah untuk diidentifikasi hanya muncul kurang dari 1 detik, kerap kali sekitar 0,5 detik (Ekman & Friesen, 1969, hlm. 93-94).

Konsep mengenai dasar ekspresi wajah pertama kali dinyatakan pada penelitian Ekman (1970). Pada penelitian tersebut ditemukan bahwa terdapat enam jenis ekspresi dasar, yaitu senang, marah, takut, muak, terkejut, dan sedih. Selain itu, ekspresi-ekspresi tersebut selalu ditemukan pada investigasi yang mempelajari mengenai ekspresi wajah pada semua jenis kultur (Ekman, 1970, hlm. 156; Ekman & Friesen, 1975, hlm. 98).

Dalam konteks kepuasan pelanggan, secara umum dapat dilakukan dengan cara mengevaluasi melalui pertanyaan yang ditanyakan melalui surat elektronik (*e-mail)*, telefon, wawancara tatap muka, atau perangkat layar sentuh. Evaluasi kepuasan pelanggan juga dapat dilakukan dengan memberikan pertanyaan berupa skala penilaian agar pelanggan dapat memberikan evaluasi terhadap jasa atau

barang yang diterima (Slim dkk., 2018, hlm. 1). Namun, cara-cara tersebut bukan cara terbaik untuk mendapatkan penilaian yang jujur dari pelanggan karena pelanggan dapat memberikan penilaian yang tidak jujur. Maka dari itu, untuk mendapatkan penilaian yang jujur (Ringler, 2021), evaluasi kepuasan pelanggan dilakukan melalui pengenalan ekspresi wajah. Kemudian, dengan mempertimbangkan pengenalan ekspresi wajah dalam waktu yang singkat, penelitian ini menggunakan kecerdasan buatan untuk mengembangkan sistem evaluasi yang efektif.

Pengenalan ekspresi wajah atau *facial expression recognition* (FER) merupakan metode kecerdasan buatan yang dapat digunakan pada berbagai macam tempat yang membutuhkan evaluasi kepuasan pelanggan. Contohnya adalah penelitian FER yang digunakan untuk menilai efisiensi iklan serta agar iklan tersebut bisa lebih interaktif dengan pelanggan (Vinh & Tran Dac Thinh, 2019). Kemudian, terdapat juga penelitian FER untuk menilai kepuasan pengunjung bandara melalui kamera video keamanan (Sugianto dkk., 2018). Selanjutnya, terdapat juga penelitian mengenai FER untuk pertunjukan opera (Ceccacci dkk., 2023). Selain digunakan sebagai penilaian dari kepuasan penonton terhadap pertunjukan opera, hasil dari analisa FER juga digunakan sebagai penentu pemilihan dari spesifikasi karakteristik untuk pertunjukan selanjutnya. Setelah itu, sektor pariwisata juga menggunakan FER sebagai alat dalam mengevaluasi kepuasan turis ketika mengunjungi situs warisan UNESCO (González-Rodríguez dkk., 2020). Penelitian FER juga digunakan sebagai evaluasi kepuasan pelanggan terhadap konsep, makanan, serta suasana dari restoran tak berawak (Chang dkk., 2023). Selain hasil berdasarkan akurasi, hasil dari penelitian mengenai pengukuran kepuasan nasabah bank melalui analisis ekspresi wajah, digunakan untuk mengevaluasi performa bank (Karadağ, 2018, hlm. 3). Dengan menerapkan evaluasi kepuasan pelanggan yang didapatkan melalui FER, bank pada penelitian tersebut dapat mengevaluasi kinerja pegawai serta mengidentifikasi ketidakpuasan nasabah terhadap suatu layanan.

Terdapat beberapa penelitian mengenai pengenalan ekspresi wajah untuk evaluasi kepuasan pelanggan menggunakan berbagai metode kecerdasan buatan.

Dengan menggunakan *Improved Support Vector Machine,* penelitian ini berhasil mencapai akurasi hingga 82.14% (Chetana dkk., 2022, hlm. 1). Selanjutnya, penggunaan metode Inception-V3 berbasis *Convolutional Neural Networks* untuk mengenali ekspresi wajah dengan menggunakan set data CK+, FER-2013, dan JAFFE, berhasil mencapai akurasi hingga 99.5% (Meena dkk., 2023, hlm. 1).

Dengan penelitian-penelitian terdahulu yang telah menggunakan berbagai macam metode untuk keperluan pengenalan ekspresi wajah, pada penelitian ini akan digunakan metode menggunakan pendekatan DETR. Dengan mengadaptasi metode Transformer (Vaswani dkk., 2017) pada bidang *computer vision*, DETR (*Detection Transformer)* (Carion dkk., 2020) berhasil menjadi metode yang memengaruhi penelitian ranah *computer vision* secara signifikan. Cara kerja metode DETR selaras dengan perkembangan tren, yaitu metode berbasis Transformer, yang sedang populer dalam penggunaan *deep learning*. DETR bekerja dengan menggunakan Transformer yang dikenal luas menggunakan sistem Attention. Selain itu, DETR juga memiliki keunggulan berupa mekanisme *end-to-end* yang tidak memerlukan pemrosesan lebih dari satu tahap seperti halnya mekanisme Faster R-CNN. Walaupun belum terdapat penelitian pengenalan ekspresi wajah dengan menggunakan DETR, penelitian lain dengan pendekatan Attention, yaitu metode DAN (*Distract your Attention Network)* (Wen dkk., 2023), berhasil mencapai ukuran parameter sebesar 19.72M serta akurasi sebesar 89.70% pada set data RAF-DB. Metode DETR juga memiliki kekurangan dari segi kecepatan yang menjadikannya kurang ideal untuk sistem yang membutuhkan pendeteksian dalam waktu singkat (Lv dkk., 2023, hlm. 2).

Salah satu penelitian metode berbasis DETR terkini, yaitu RT-DETR (*Real Time-Detection Transformer),* berhasil mengatasi kekurangan tersebut dan menjadikan metode berbasis Transformer unggul dalam segi kecepatan serta akurasi dibandingkan dengan YOLO (Lv dkk., 2023). Keunggulan yang dimiliki metode RT-DETR, juga berhasil mengalahkan metode terkini yang lain, seperti DETR *baseline,* Efficient DETR, DINO, dan sebagainya.

Dengan kesesuaian metode RT-DETR untuk kebutuhan pengenalan ekspresi wajah, yaitu kecepatan dan efisiensi, hal ini dianggap sesuai dengan manfaat

pengimplementasian perangkat lunak di perangkat *edge*. Berdasarkan survey mengenai pengenalan ekspresi wajah, merekomendasikan pengembangan FER untuk perangkat *edge*. Manfaat dari pengembangan FER dengan komputasi rendah merupakan meningkatkan pemrosesan data, memastikan hasil keputusan terjadi dalam waktu *real-time,* serta meningkatkan keamanan data (Sajjad dkk., 2023, hlm. 833). Namun dalam pelaksanaannya, penggunaan sistem kecerdasan buatan pada perangkat *edge* dibatasi oleh banyak tantangan. Hal ini disebabkan oleh berbagai keterbatasan, antara lain seperti komputasi, penyimpanan memori, dan pengaksesan data. Kemudian, jaringan syaraf tiruan dan kecerdasan buatan memiliki karakteristik, yaitu membutuhkan penyimpanan serta pengaksesan parameter dalam jumlah yang besar (Plastiras dkk., 2018, hlm. 3). Permasalahan ini disimpulkan oleh latar belakang penelitian Li dkk. (2022, hlm. 1), yaitu merancang model *light-weight* yang membutuhkan konsumsi sumber daya yang rendah dengan performa yang tinggi merupakan masalah utama dalam pengembangan model untuk penggunaan luring.

Kemudian, terdapat juga masalah lain terkait dengan kebutuhan sumber daya untuk pengembangan perangkat lunak. Sebuah penelitian Thompson dkk. (2023, hlm. 1) mengenai kebutuhan komputasi terhadap pengembangan *deep learning* mengemukakan bahwa kebutuhan komputasi telah meningkat secara dramatis dan peningkatan tersebut merupakan aspek krusial pada peningkatan performa *deep learning.* Penelitian tersebut menyimpulkan, apabila tren tersebut berlanjut, pertumbuhan beban komputasi akan dengan cepat menjadi tantangan secara teknis, ekonomi, serta lingkungan alam.

Berdasarkan penemuan tersebut, penelitian ini mengajukan metode LCNet (*Lightweight* CPU *Convolutional Neural Network*) sebagai variabel dalam eksperimen RT-DETR (Cui dkk., 2021). Penelitian L-DETR oleh Li dkk. (2022), yaitu pengimplementasian metode LCNet dengan DETR, berhasil mengurangi parameter sebanyak 26% dan 46% lebih sedikit dibandingkan dengan DETR yang memiliki backbone ResNet-50 dan ResNet-18. Selain itu, keseluruhan penelitian L-DETR dilakukan pada perangkat yang ringan yaitu 1 unit CPU dan 2 unit GPU, jika dibandingkan dengan penelitian DETR yang membutuhkan 16 unit GPU V100.

Penelitian ini berupaya untuk menjembatani kesenjangan penelitian subjek pengenalan ekspresi wajah atau FER pada metode dengan pendekatan DETR (RT-DETR). Selain itu, metode LCNet memiliki potensi untuk mengurangi besar ukuran model RT-DETR agar mudah digunakan pada dunia nyata dengan penggunaan perangkat yang ringan pada pengembangannya. Dengan keterbatasan dari set data yang tersedia untuk pengembangan penelitian ini, eksperimen terdiri dari dua tahap, yaitu tahap deteksi wajah menggunakan metode RT-DETR dan tahap klasifikasi ekspresi wajah menggunakan metode yang terpisah.

1.2. Rumusan Masalah

Mengacu pada latar belakang, rumusan masalah berikut akan diselesaikan pada penelitian ini:

1. Bagaimana pengimplementasian dua model (deteksi wajah dan pengklasifikasian ekspresi) pada proses evaluasi kepuasan pelanggan berdasarkan ekspresi wajah?
2. Bagaimana pengimplementasian keseluruhan model pengenalan ekspresi wajah pada proses evaluasi kepuasan pelanggan berdasarkan ekspresi wajah?
3. Bagaimana performa dua model (deteksi wajah dan pengklasifikasian ekspresi) pada proses evaluasi kepuasan pelanggan berdasarkan ekspresi wajah?
4. Bagaimana performa keseluruhan model pengenalan ekspresi wajah pada proses evaluasi kepuasan pelanggan berdasarkan ekspresi wajah?

1.3. Tujuan Penelitian

Mengacu pada rumusan masalah, berikut merupakan tujuan dari pelaksanaan penelitian ini:

1. Mengetahui pengimplementasian masing-masing model (deteksi wajah dan pengklasifikasian ekspresi).
2. Mengetahui pengimplementasian keseluruhan model.
3. Mengevaluasi dan menganalisa performa masing-masing model (deteksi wajah dan pengklasifikasian ekspresi) untuk pengenalan ekspresi wajah.
4. Mengevaluasi performa keseluruhan model untuk pengenalan ekspresi wajah.

1.4. Manfaat Penelitian

Dari penelitian ini, diharapkan dapat menghasilkan manfaat seperti berikut:

1. Bagi peneliti

Peneliti diharapkan mendapatkan pengetahuan baru mengenai pengenalan ekspresi wajah serta memahami proses dan alur kerja dari penggabungan model deteksi wajah dan klasifikasi ekspresi.

2. Bagi pihak lain

Penelitian ini diharapkan dapat menjadi rujukan bagi penelitian pengenalan ekspresi wajah selanjutnya.

1.5. Ruang Lingkup Penelitian

Berikut merupakan batasan masalah dari penelitian ini:

1. Digunakan 2 jenis jumlah ekspresi wajah, yaitu 7 jenis ekspresi dan 3 jenis ekspresi. Rincian jenis ekspresi adalah sebagai berikut: 1) 7 jenis ekspresi: bahagia, sedih, terkejut, takut, marah, muak dan netral. 2) 3 jenis ekspresi: positif, negatif, dan netral.
2. Set data yang digunakan untuk pengembangan dan pengujian merupakan citra tidak bergerak (*still image).*
3. Lingkup implementasi dilaksanakan pada ruangan dengan pencahayaan yang cukup agar kamera dapat menangkap ekspresi wajah pelanggan.

1.6. Sistematika Penelitian

Berikut merupakan sistematika penulisan yang disusun untuk mempermudah pemahaman skripsi ini.

A. BAB I PENDAHULUAN

Bab ini menjelaskan mengenai pengenalan ekspresi wajah untuk kepuasan pelanggan yang terdiri dari latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, serta sistematika penelitian.

B. BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan mengenai metode serta landasan teori yang akan digunakan sebagai referensi ataupun materi pada penelitian ini, seperti pengenalan wajah, pengenalan ekspresi wajah atau *facial expression recognition* (FER), kepuasan pelanggan, pengolahan citra digital, metode pengenalan objek, yaitu RT-DETR (*Real Time-Detection Transformer)* serta LCNet (*Lightweight* CPU *Convolutional Neural Network*).

C. BAB III METODE PENELITIAN

Bab ini menjelaskan mengenai semua tahapan pembangunan sistem pengenalan ekspresi wajah untuk kepuasan pelanggan menggunakan metode RT-DETR (*Real Time-Detection Transformer)* dengan komponen ResNet-18 dan LCNet (*Lightweight* CPU *Convolutional Neural Network*), yang terdiri dari desain penelitian, rancangan model dan eksperimen, serta lingkungan komputasi eksperimen.

D. BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai hasil yang telah didapat dari penelitian yang telah dilakukan. Terdiri dari pengolahan data, implementasi metode, eksperimen, dan evaluasi.

E. BAB V SIMPULAN DAN SARAN

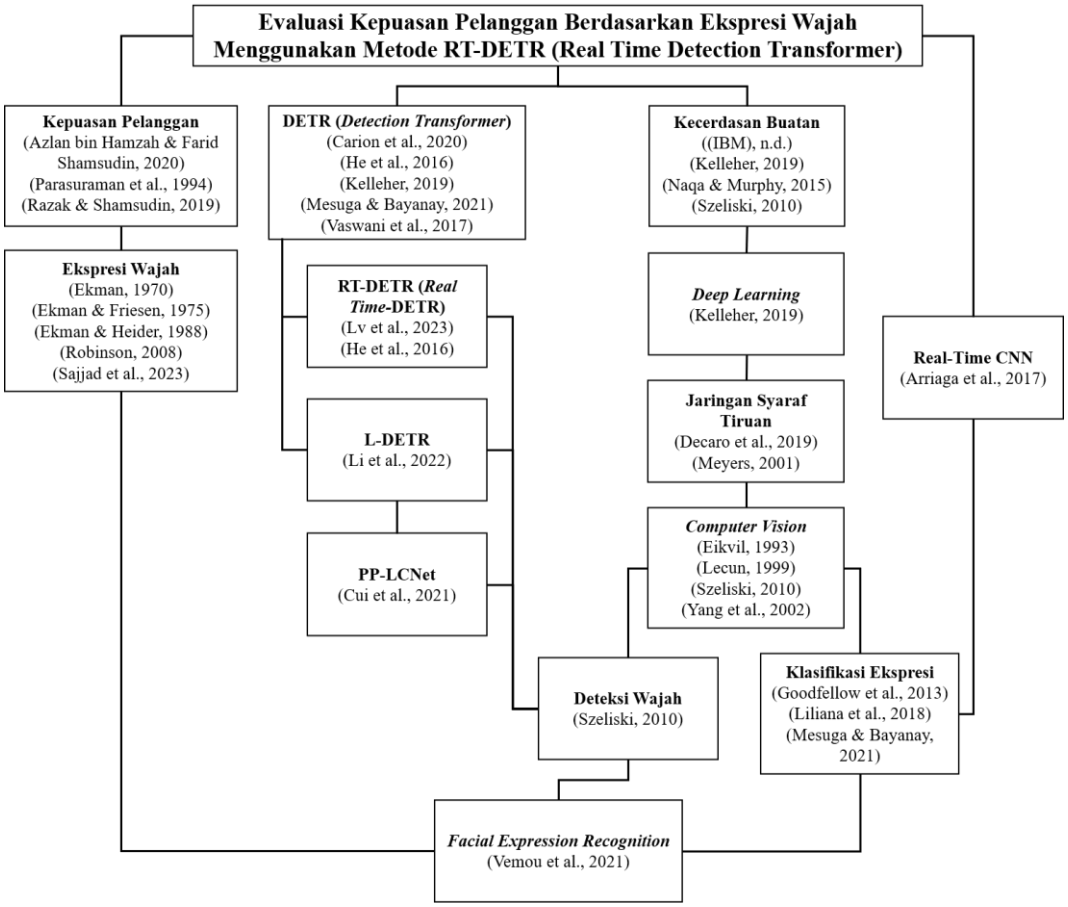
Bab ini menjelaskan mengenai simpulan akhir dari hasil penelitian yang telah dilakukan serta saran dari peneliti untuk penelitian yang selanjutnya.

BAB II
TINJAUAN PUSTAKA

Bab ini merupakan penguraian serta penjabaran dari berbagai teori yang melandasi atau menjadi argumen penelitian yang dilakukan.

2.1. Peta Literatur

Peta literatur pada Gambar 2.1 berikut berisi mengenai referensi literatur yang dimuat pada subbab ini.



Gambar 2.1 Peta Literatur

2.2. Penelitian Terkait

Terdapat beberapa penelitian terkait mengenai pengembangan *facial expression recognition* (FER) menggunakan pendekatan kecerdasan buatan. Penelitian FER menggunakan *Improved Support Vector Machine*, berhasil

mencapai akurasi hingga 82.14% (Chetana dkk., 2022). Dengan menggunakan set data FER-2013, penelitian ini menguji metode *Improved support vector machine* (SVM) dan *Flattened convolutional neural networks* (CNN). Berdasarkan hasil yang didapatkan, *Improved* SVM berhasil mengalahkan *Flattened* CNN yang hanya mencapai akurasi sebesar 61.16%.

Kemudian, penelitian dengan pendekatan Inception-V3 berbasis CNN dan menggunakan proses *transfer learning* pada *layer* klasifikasi terakhir pada Inception-V3. Dengan menggunakan set data CK+, FER-2013, dan JAFFE, penelitian ini berhasil mencapai akurasi pada masing-masing set data sebesar 99.5%, 73%, dan 86% (Meena dkk., 2023).

Metode pendeteksian objek berbasis Transformer, yaitu DETR (*Detection Transformer),* menggunakan pendekatan *end-to-end* (Carion dkk., 2020). Penerapan *end-to-end* memungkinkan DETR untuk melakukan seluruh proses pada satu sistem yang sama. DETR menerapkan tiga komponen, yaitu *backbone,* Transformer, dan *Feed-Forward Networks (FFNs).*

Penelitian mengenai pendeteksian ekspresi wajah berbasis pendekatan Transformer (Wen dkk., 2023) menggunakan metode DAN (*Distract your Attention Network*) berhasil mencapai akurasi sebesar 89.70% pada set data RAF-DB beserta parameter sebesar 19.72M. Hasil dari penelitian tersebut membuktikan bahwa metode berbasis Transformer terbukti dapat mencapai keakuratan yang tinggi untuk deteksi ekspresi wajah. Walaupun dengan keunggulannya, DETR memiliki kelemahan dalam kecepatan deteksi yang disebabkan oleh NMS (*Non-Maximum Supervision*) (Lv dkk., 2023 hlm. 2).

Untuk mengatasi kelemahan DETR dalam kecepatan deteksi, Baidu melaksanakan penelitian yang menggabungkan DETR dengan sifat *real-time* (Lv dkk., 2023), yaitu RT-DETR (*Real-Time Detection Transformer*). RT-DETR mengatasi kelemahan DETR yang disebabkan oleh NMS dengan meningkatkan kinerja dua komponen, yaitu Encoder *hybrid* yang dapat memproses *multi-scale features* secara efisien dan *IoU-aware query selection* yang memperbaiki inisialisasi *object queries.* Metode RT-DETR berhasil mencapai kecepatan hingga

114 FPS serta akurasi hingga 53% AP yang mengalahkan YOLO dan DETR *baseline*.

Pada penelitian ini, dilakukan eksperimen untuk mengoptimalkan ukuran model agar mudah digunakan pada kegiatan sehari-hari. Pada umumnya, pengimplementasian sistem *machine learning* pada kehidupan sehari-hari akan melibatkan *framework* TensorFlow Lite karena ukurannya yang kecil dibandingkan dengan ukuran model lain (TensorFlow, 2019). Penelitian yang menggunakan TensorFlow Lite untuk deteksi objek secara *real-time* (Dai, 2020), menghasilkan model dengan ukuran parameter sebesar 6.1M. Ukuran tersebut jauh lebih kecil jika dibandingkan dengan besar model RT-DETR yaitu 42M dan DETR *baseline* yaitu 41M.

Maka dari itu, penelitian ini mengajukan metode *backbone* LCNet sebagai solusi dari permasalahan tersebut. Penelitian L-DETR oleh Li dkk. (2022), yaitu pengimplementasian metode LCNet dengan DETR, berhasil mengurangi parameter sebanyak 26% dan 46% lebih sedikit dibandingkan dengan DETR yang memiliki backbone ResNet-50 dan ResNet-18.

2.3. Kepuasan Pelanggan

Seiring dengan meningkatnya kompetisi pada sektor bisnis, salah satu aspek yang membantu organisasi dalam melakukan evaluasi bisnis adalah berdasarkan kepuasan pelanggan. Kepuasan pelanggan mengindikasikan baiknya pengalaman penggunaan produk dibandingkan dengan ekspektasi pembeli (Razak & Shamsudin, 2019, hlm. 31).

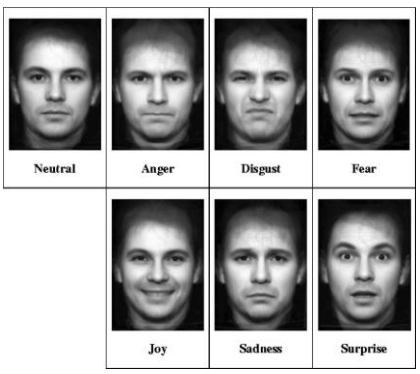
Dengan memperhatikan kepuasan pelanggan, sebuah organisasi akan mendapatkan beberapa efek sebagai berikut: 1) Organisasi dapat memuaskan kepercayaan yang diberikan oleh pelanggan; 2) Organisasi dapat mencapai kepuasan pelanggan yang lebih tinggi dibandingkan dengan kompetitor. Hal ini dikarenakan strategi bisnis mengharuskan organisasi untuk memiliki keunggulan kompetitif dibandingkan dengan kompetitornya; 3) Berdasarkan penelitian oleh Parasuraman dkk. (1994), kepuasan pelanggan penting agar organisasi dapat mempertahankan pelanggan dalam waktu yang lama; 4) *Market segmentation*

(pembagian pasar) diciptakan oleh pelanggan. Dengan memeroleh kepuasan pelanggan, organisasi bisnis dapat memiliki pangsa pasar; dan 5) Dengan memeroleh pangsa pasar, organisasi dapat meningkatkan pendapatan yang mana akan meningkatkan pendapatan pemilik saham (Hamzah & Shamsudin, 2020, hlm. 3).

Pelanggan setia akan memengaruhi perilaku orang-orang di sekitarnya agar mereka membeli produk atau jasa yang sama. Dengan meningkatnya konsumsi produk dan jasa yang ditawarkan oleh organisasi, kepuasan pelanggan akan membantu dalam meningkatkan penjualan serta keuntungan (Hamzah & Shamsudin, 2020, hlm. 3).

Umumnya, evaluasi kepuasan pelanggan dilakukan melalui surat elektronik, telepon, wawancara tatap muka, atau perangkat layar sentuh, sering kali dengan skala penilaian (Slim dkk., 2018 hlm. 1). Namun, metode-metode tersebut tidak selalu menghasilkan penilaian yang jujur (Ringler, 2021). Oleh karena itu, penelitian ini menggunakan pengenalan ekspresi wajah dan kecerdasan buatan untuk mengembangkan sistem evaluasi yang lebih efektif. Kemudian, berdasarkan penelitian mengenai kepuasan nasabah bank melalui pengenalan ekspresi wajah, analisa hasil yang diperoleh membantu bank tersebut untuk meningkatkan serta mengidentifikasi kekurangan dari pelayanan (Karadağ, 2018, hlm. 3). Maka dari itu, penelitian ini memiliki potensi untuk membantu bisnis dalam meningkatkan layanannya.

2.4. Ekspresi Wajah



Gambar 2.2 Enam ekspresi dasar dan ekspresi netral

Konsep mengenai dasar ekspresi wajah pertama kali dinyatakan pada penelitian oleh Ekman (1970). Pada penelitian tersebut ditemukan bahwa terdapat enam jenis ekspresi dasar, yaitu senang, marah, takut, muak, terkejut, dan sedih. Selain itu, ekspresi-ekspresi tersebut selalu ditemukan pada investigasi yang mempelajari mengenai ekspresi wajah pada semua jenis kultur (Ekman, 1970, hlm. 156; Ekman & Friesen, 1975, hlm. 98).

Berdasarkan penelitian-penelitian oleh Ekman mengenai ekspresi wajah, dapat dipahami bahwa konsep mengenai dasar ekspresi wajah mengalami perubahan seiring dengan bertambahnya temuan dari penelitian lain. Dimulai dari 6 ekspresi (Ekman, 1970, hlm. 156) yang terdiri dari senang, marah, takut, muak, terkejut, dan sedih hingga bertambahnya 1 ekspresi baru, yaitu malu *(contempt)* (Ekman & Heider, 1988).

Walaupun Ekman & Heider (1988) mengemukakan bahwa ekspresi dasar terdiri dari senang, marah, takut, muak, terkejut, sedih, dan malu, mayoritas set data FER hanya menggunakan jenis pengklasifikasian yang terdiri dari 6 ekspresi wajah dasar yang sesuai dengan penelitian Ekman (1970). Berdasarkan survey Sajjad dkk. (2023, hlm. 830), ditemukan bahwa set data FER secara umum menggunakan 7 jenis kelas ekspresi, yang terdiri dari 1) 6 ekspresi dasar yang terdiri dari senang, marah, takut, muak, terkejut, dan sedih; dan 2) 1 ekspresi tambahan, yaitu netral.

Selanjutnya, terdapat penelitian, oleh Robinson (2008, hlm. 155), yang memisahkan 6 ekspresi dasar tersebut menjadi 2 kelas, yaitu positif dan negatif. Dengan terbaginya emosi menjadi dua kelas dan 1 kelas tambahan, yaitu netral, maka akan memudahkan sistem untuk menilai apakah pelanggan puas atau tidak puas. Rincian penilaian kepuasan pelanggan berdasarkan kelas emosi adalah sebagai berikut:

1) Emosi positif

   Emosi positif menandakan ekspresi senang dan terkejut. Dengan terdeteksinya emosi ini, maka menandakan pelanggan puas.

2) Emosi negatif

   Emosi negatif menandakan ekspresi sedih, marah, muak, dan takut. Dengan terdeteksinya emosi ini, maka menandakan pelanggan tidak puas.

3) Emosi netral

   Emosi netral menandakan ekspresi netral. Dengan terdeteksinya emosi ini, maka menandakan pelanggan tidak sedang memiliki emosi.

2.5. Kecerdasan Buatan

Kecerdasan buatan atau bisa dalam bahasa Inggris disebut sebagai *Artificial Intelligence* merupakan teknologi yang memungkinkan komputer atau mesin untuk menyimulasikan kecerdasan serta kemampuan penyelesaian masalah yang dimiliki oleh manusia ((IBM), n.d.). Selanjutnya, kecerdasan buatan juga merupakan salah satu cabang ilmu pengetahuan bidang komputer yang memiliki tujuan untuk meneliti serta mengembangkan perangkat lunak agar menyerupai kecerdasan yang dimiliki oleh makhluk hidup (manusia).

Kecerdasan buatan memiliki beberapa cabang yaitu sebagai berikut:

1) *Machine learning*

   Machine learning, berdasarkan Naqa & Murphy (2015), merupakan cabang ilmu pengetahuan dari algoritma komputasi yang dirancang untuk menyimulasikan kecerdasan manusia dengan belajar dari lingkungan di sekitarnya. Selanjutnya, *machine learning* juga merupakan proses komputasi yang menggunakan data input untuk menyelesaikan tugas yang diberikan, tanpa harus diprogram secara eksplisit (*hard coded*) untuk menghasilkan *output* tertentu.

2) *Deep learning*

   Menurut Kelleher (2019), *Deep Learning* merupakan sub-bidang dari kecerdasan buatan yang fokus dalam pembuatan model jaringan syaraf tiruan besar yang mampu membuat keputusan akurat berdasarkan data. Secara umum, *deep learning* cocok dengan skenario yang memiliki data kompleks serta dengan jumlah yang besar. Salah satu contoh pengimplementasian *deep learning* adalah untuk penggunaan deteksi wajah pada kamera digital. Selain itu, pada sektor kesehatan, *deep learning* juga digunakan untuk memproses foto medis, seperti *X-rays*, *CT scans*, dan *MRI scans*.

3) *Computer Vision*

Computer Vision merupakan cara untuk mendeskripsikan lingkungan yang dilihat oleh mahluk hidup dalam satu citra atau lebih serta untuk merekonstruksi semua properti yang ada, seperti bentuk, pencahayaan, dan distribusi warna (Szeliski, 2010, hlm. 3).

2.6. *Deep Learning*

Pengertian *deep learning* menurut Kelleher (2019) adalah salah satu cabang dari ilmu kecerdasan buatan yang berkutat dengan model komputasi jaringan syaraf tiruan untuk membuat keputusan berdasarkan data. *Deep learning* juga merupakan salah satu cabang dari *machine learning* berdasarkan jaringan syaraf tiruan. Kosakata *deep* yang bisa diartikan dalam, digunakan untuk merepresentasikan banyaknya *layer* ketika mengembangkan jaringan syaraf tiruan. Metode yang digunakan pada jaringan syaraf tiruan, yaitu *supervised* dan *unsupervised* yang dengan rincian sebagai berikut:

1) *Supervised*

Tipe *machine learning* yang memiliki tujuan untuk mempelajari fungsi yang memetakan atribut-atribut input dengan sebuah prediksi akurat dari nilai yang hilang. Dalam penjelasan lain, *supervised* merupakan tipe algoritma *machine learning* yang dilatih menggunakan set data yang memiliki label.

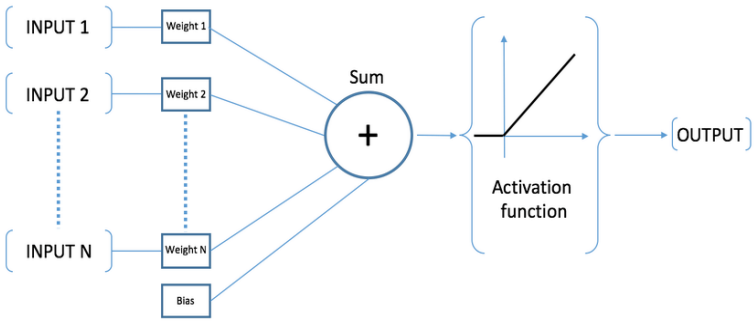
2) *Unsupervised*

Tipe *machine learning* yang memiliki tujuan untuk mengidentifikasi regularitas atau hal yang sering muncul, seperti kumpulan dari suatu hal yang sama. Dalam penjelasan lain, *unsupervised* merupakan tipe algoritma *machine learning* yang dilatih menggunakan set data yang tidak memiliki label.

2.7. Jaringan Syaraf Tiruan (*Artificial Neural Network*)

Jaringan syaraf tiruan (JST) atau dalam bahasa Inggris disebut sebagai *artificial neural network* (ANN) merupakan cabang dari subjek kecerdasan buatan yang mengikuti konsep cara kerja jaringan syaraf makhluk hidup. JST digunakan

untuk mengatasi masalah kompleks yang melibatkan pola pada tipe kategorisasi ataupun analisa tren (Meyers, 2001).



Gambar 2.3 Prinsip kerja jaringan syaraf tiruan

JST merupakan algoritma yang membuat keputusan berdasarkan set data yang sebelumnya telah diberikan dan berusaha untuk menemukan hubungan tersembunyi yang ada, walaupun tidak tertulis secara eksplisit. Secara umum, JST dibangun berdasarkan sebuah syaraf (*neuron*). Syaraf merupakan elemen pemrosesan yang disusun berdasarkan bermacam-macam koneksi. Pada setiap syaraf, input adalah set data asli ataupun berat (*weight*) dan bias yang diterima dari syaraf sebelumnya yang juga telah melewati fungsi aktivasi (Decaro dkk., 2019). Prinsip kerja ini diilustrasikan pada Gambar 2.3. JST memiliki berbagai macam variasi, salah satunya adalah *Feed Forward Network* (FFN).

2.7.1. *Convolutional Neural Network*

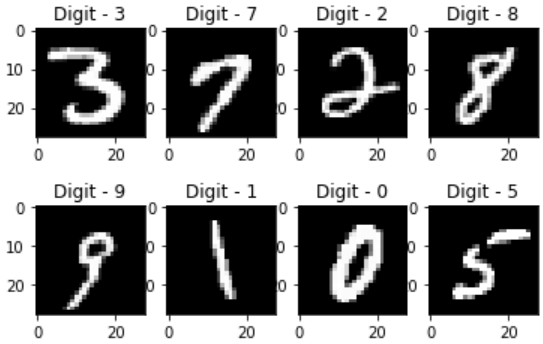
Convolutional Neural Network, yang juga memiliki singkatan CNN, dirancang untuk pengenalan citra dan pada awalnya diaplikasikan untuk mengatasi tantangan pengenalan digit angka yang ditulis dengan tangan. Menurut Venkatesan & Li (2018), CNN merupakan *neural network* yang memberikan sinyal input ke lapisan *convolutional pooling* yang kemudian lapisan terakhirnya memberikan *output* ke susunan *fully connected layers*. Selanjutnya *output* dari lapisan tersebut masuk lapisan *softmax*.

2.8. *Computer Vision*

Menurut Szeliski (2010, hlm. 3), *computer vision* merupakan cara untuk mendeskripsikan lingkungan yang dilihat oleh mahluk hidup dalam satu citra atau

lebih serta untuk merekonstruksi semua properti yang ada, seperti bentuk, pencahayaan, dan distribusi warna. Saat ini *computer vision* digunakan pada berbagai jenis pengimplementasian di dunia nyata, di antaranya seperti:

1) *Optical character recognition* (OCR)



Gambar 2.4 Tulisan tangan dari digit angka (Lecun, 1999)

Optical character recognition (OCR) menurut Eikvil (1993) merupakan sistem untuk mengenali karakter melalui tulisan tangan atau hasil cetak dari komputer secara optik. Contoh penggunaan OCR adalah pada pengenalan digit angka yang ditulis tangan (Gambar 2.4)

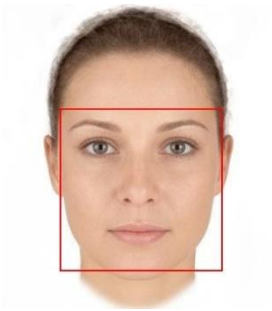
2) Deteksi wajah

Gambar 2.5 Deteksi wajah

Pemrosesan wajah dikembangkan dengan harapan bahwa identitas, keadaan, dan niat seseorang bisa diekstraksi dari sebuah citra dan komputer dapat memberikan respon yang sesuai berdasarkan hasil yang telah diekstraksi (M. H. Yang dkk., 2002, hlm. 1). Deteksi wajah memiliki tujuan utama, yaitu menentukan apakah terdapat wajah pada sebuah citra dan, jika

terdapat wajah, maka berikan lokasi serta ukuran wajah yang terdeteksi. Terdapat beberapa tantangan dalam pendeteksian wajah, salah satunya adalah ekspresi wajah karena penampilan wajah secara langsung dipengaruhi oleh ekspresi wajah orang tersebut. Pada Gambar 2.5 berikut merupakan contoh pendeteksian wajah yang ditandai dengan *bounding box* atau indikator yang menandakan adanya wajah yang terdeteksi.

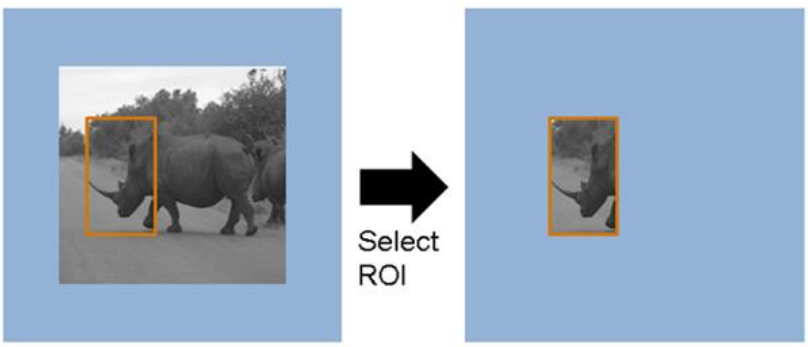
2.8.1. *Bounding Box*



Gambar 2.6 Contoh penggunaan *bounding box*

Bounding box merupakan persegi panjang yang mengelilingi sebuah objek (Ramla dkk., 2022). Secara intuitif, *bounding box* dapat direpresentasikan sebagai (x, y, w, h) dengan (x, y) sebagai titik koordinat dan (w, h) sebagai lebar serta panjang. Koordinat dari *bounding box* dikalkulasikan dari pojok kiri atas dari sebuah citra sebagai koordinat $(0, 0)$. Berikut (Gambar 2.6) merupakan contoh penggunaan *bounding box* yang ditandai dengan garis merah dari sebuah foto wajah.

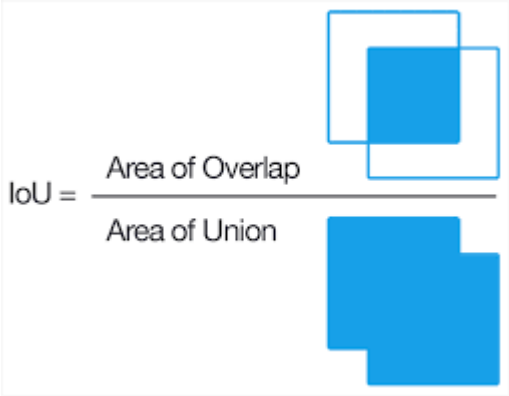
2.8.2. *Region of Interest* (ROI)



Gambar 2.7 Ilustrasi penggunaan ROI

Region of Interest (ROI) dapat dikatakan merupakan sebuah bagian atau batasan wilayah dari citra yang akan diproses secara khusus (Pratomo dkk., 2020). ROI dapat digunakan sebagai cara untuk mengurangi masalah tingginya waktu pemrosesan. Penggunaan metode ini dapat mengoptimalkan kinerja sistem dan tanpa ROI, pemrosesan pada citra akan dilakukan pada semua piksel yang akan menghabiskan waktu dan sumber daya.

2.8.3. *Intersection over Union* (IoU)



Gambar 2.8 Ilustrasi Intersection over Union (IoU)

IoU merupakan metrik yang digunakan pada deteksi objek dan digunakan untuk mengukur besaran tumpang tindih (*overlap*) antara *bounding box* prediksi dan *ground truth* (Terven dkk., 2023, hlm. 29). IoU dikalkukasi dengan cara membagi area yang beririsan dengan area gabungan, diilustrasikan pada gambar 2.8. Pada penelitian ini, metrik IoU digunakan bersamaan dengan metrik *average precision* (AP) yang memiliki tiga jenis *threshold*, 0.50 hingga 0.95 (standar COCO), 0.5 (standar PASCAL VOC), dan 0.75 (standar ketat).

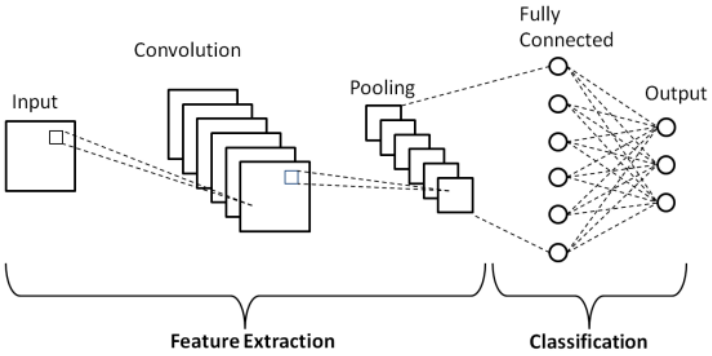
2.9. Deteksi Wajah

Deteksi wajah merupakan bagian dari cabang kecerdasan buatan yang bertujuan untuk menemukan serta mengidentifikasi wajah manusia dari citra atau video. Tahap pendeteksian wajah dilakukan sebelum tahap pengenalan ekspresi wajah. Walaupun tahap pengenalan ekspresi wajah dapat dilakukan secara langsung dan tanpa tahap deteksi wajah, proses tersebut akan memakan waktu yang lama. Hal tersebut karena algoritma pengenalan ekspresi wajah akan dieksekusi pada

setiap piksel dan skala. Sedangkan, jika menerapkan tahap deteksi wajah terlebih dahulu, tahap pengenalan ekspresi wajah hanya akan dilakukan pada piksel yang terindikasi adanya wajah (Szeliski, 2010, hlm. 578).

Hasil yang bisa didapat dari proses pendeteksian wajah adalah foto atau video dari wajah yang telah dideteksi, koordinat *bounding box*, serta keterangan yang menyatakan adanya wajah atau tidak pada foto atau video yang dideteksi. Ilustrasi hasil dari pendeteksian wajah, dapat dilihat pada Gambar 2.5.

2.10. Klasifikasi Ekspresi



Gambar 2.9 Arsitektur CNN

Tahap klasifikasi ekspresi dilakukan dengan menerima output berupa wajah yang terdeteksi dari tahap deteksi wajah. Secara umum, klasifikasi ekspresi terjadi ketika input wajah telah diproses melalui layer convolution. Penelitian Mesuga & Bayanay (2021) dapat menggambarkan apa yang umumnya terjadi pada model klasifikasi ekspresi. Dimulai dari arsitektur CNN yang terdiri dari dua bagian utama (Gambar 2.9), yaitu ekstraksi fitur dan klasifikasi. Bagian ekstraksi fitur bertugas untuk mengekstraksi fitur dari input citra. Selanjutnya, bagian klasifikasi bertugas untuk menentukan kelas dari citra. CNN merupakan komponen yang terdapat pada berbagai model yang memproses citra, seperti DETR.

Algoritma CNN digunakan pada penelitian ini sebagai model klasifikasi ekspresi. Pengembangan model menggunakan set data FER-2013 (Goodfellow dkk., 2013) yang pernah digunakan pada penelitian terdahulu (Chetana dkk., 2022; Meena dkk., 2023) serta set data IMED (Liliana dkk., 2018). Kedua set data tersebut

akan mengombinasikan keunggulannya masing-masing, yaitu pose wajah yang bervariasi serta wajah penduduk asli Indonesia, yang menjadikannya cocok untuk penilaian kepuasan pelanggan yang pada pelaksanaannya akan mengklasifikasi wajah melalui berbagai sudut dengan ciri khas wajah Indonesia.

2.11. *Facial Expression Recognition* (FER)

Facial expression recognition (FER) atau bisa disebut sebagai pengenalan ekspresi wajah merupakan teknologi yang digunakan untuk menganalisa sentimen yang bisa didapatkan melalui sumber, seperti foto dan video. Analisa dalam pengenalan ekspresi wajah terdiri dari tiga tahap, yaitu deteksi wajah, pendeteksian ekspresi wajah, dan pengklasifikasian ekspresi wajah (Vemou dkk., 2021, hlm. 1). Walaupun begitu, pada penelitian ini hanya terdapat dua tipe model, yaitu model deteksi wajah dan model klasifikasi ekspresi. Hal ini dikarenakan model ekspresi wajah sudah mencakup pendeteksian ekspresi wajah dan pengklasifikasian ekspresi wajah.

Sumber data foto atau video untuk pengembangan sistem pendeteksian ekspresi wajah berasal dari kamera keamanan atau kamera khusus pendeteksian ekspresi wajah. FER memiliki banyak pengimplementasian, seperti analisa perilaku pelanggan dan periklanan. Kemudian, pengembangan FER melibatkan sejumlah tantangan pada set data yang digunakan untuk *training*, seperti pencahayaan, pose wajah, halangan (*occlusion)*, penuaan pada wajah, serta resolusi rendah (Sajjad dkk., 2023, hlm. 818). Pada penelitian ini, FER digunakan untuk menganalisa perilaku yang terjadi ketika pelanggan sedang fokus pada produk ataupun pada suatu momen tertentu ketika jasa dilakukan.

2.12. Implementasi FER untuk Evaluasi Kepuasan Pelanggan

Menurut Lee dkk. (dalam Enholm dkk., 2022, hlm. 1718), kecerdasan buatan atau *Artificial Intelligence* (AI) merupakan teknologi yang inovatif dan dapat memungkinkan organisasi untuk berinovasi dan mentransformasikan proses bisnis. Tujuan dari semua proses bisnis adalah untuk mengubah masukan (*input*) menjadi keluaran (*output*) yang berharga dan teknologi baru diharapkan dapat memperbaiki

proses tersebut melalui transformasi yang radikal. Maka dari itu, AI juga dapat menjadi poros penggerak dalam perubahan pada struktur organisasi yang ada, memengaruhi penggunaan sumber daya manusia, serta memfasilitasi perubahan pada proses bisnis dan struktur organisasi.

Penggunaan AI juga dapat memengaruhi pada level operasional. Sebagai contoh, menemukan perubahan pada preferensi pelanggan. Hal ini disebabkan karena AI dapat digunakan untuk menganalisa pendapat, sikap, serta emosi yang berkaitan dengan produk atau pelayanan. Penelitian ini akan mengimplementasikan FER untuk keperluan evaluasi kepuasan pelanggan.

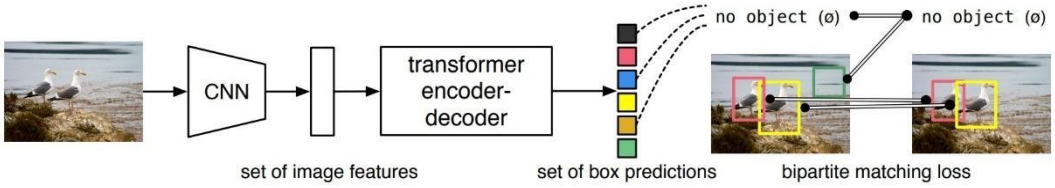
Dengan menggunakan teknologi FER untuk mengevaluasi kepuasan pelanggan, pelaku usaha dapat mengidentifikasi pelayanan yang dinilai oleh pelanggan sebagai tidak memuaskan dengan lebih cepat. Selain itu, dengan mengimplementasikan teknologi yang tidak memerlukan campur tangan manusia pada tahap operasional, pelaku usaha dapat mendapatkan hasil evaluasi kepuasan pelanggan yang akurat serta objektif. Berhubungan dengan penggunaan teknologi tanpa campur tangan manusia, hal ini dapat mengakibatkan efisiensi penggunaan manusia, yang secara langsung akan menghemat waktu serta biaya.

Contoh dari manfaat pengimplementasian FER untuk evaluasi kepuasan pelanggan, terjadi pada sebuah bank (Karadağ, 2018). Dengan menggunakan bantuan Microsoft Cognitive Services Face API, kondisi emosional nasabah dideteksi pada periode tertentu. Selanjutnya, hasil dari pendeteksian wajah yang berupa nama nasabah dan jenis emosi atau ekspresi yang dialami dianalisa berdasarkan sistem operasional bank yang bersangkutan. Hasil akhir dari analisa kepuasan pelanggan berdasarkan kepuasan pelanggan tersebut merupakan evaluasi dari performa staff dan identifikasi dari layanan yang bermasalah.

2.13. DETR (*Detection Transformer*)

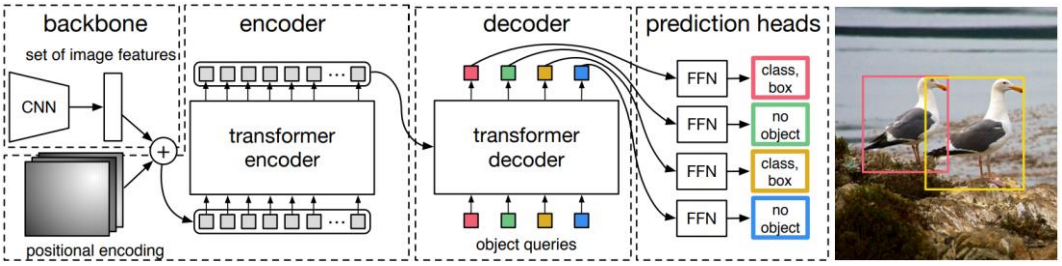
DETR yang memiliki kepanjangan *Detection Transformer* (Carion dkk., 2020) merupakan sebuah model *end-to-end deep learning* untuk pendeteksian objek. Penerapan *end-to-end* memungkinkan DETR untuk melakukan seluruh proses pada satu sistem yang sama. Transformer merupakan arsitektur ANN yang terkenal

dengan mekanisme *self-attention* yang memungkinkan untuk menangkap keterhubungan serta ketergantungan antar-elemen di dalam sebuah urutan atau kumpulan data.



Gambar 2.10 Cara kerja DETR

Secara umum, proses DETR diilustrasikan pada Gambar 2.10. DETR memprediksi semua objek pada input dalam waktu yang bersamaan dan terlatih secara *end-to-end* dengan *set loss function* yang bertugas untuk melakukan persamaan bipartite antara hasil prediksi dengan objek yang sesungguhnya.



Gambar 2.11 Arsitektur DETR

Kemudian, arsitektur DETR (Gambar 2.11) dimulai dari *backbone* CNN yang mempelajari representasi 2D dari citra input. Representasi tersebut selanjutnya di-*flatten* dan diberikan *positional encoding* sebelum dijadikan *input* untuk Transformer – Encoder. Selanjutnya, Transformer – Decoder mengambil *positional embedding* yang telah dipelajari dengan jumlah tertentu (*object queries*). Keluaran *embedding* Decoder kemudian diberikan ke *feed forward network* (FFN) untuk diprediksi sebagai objek (label kelas dan *bounding box*) atau bukan objek.

Adapun lebih lanjut, detail dari tiap komponen DETR oleh Carion dkk. (2020) sebagai berikut:

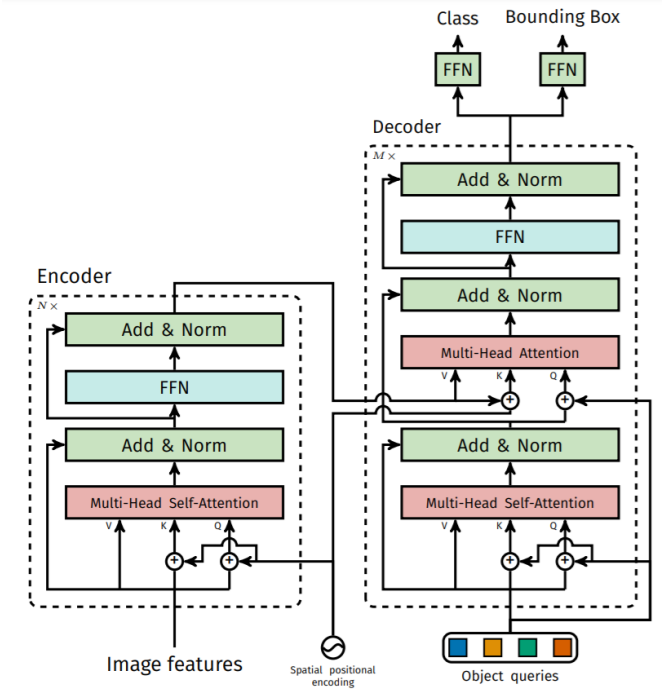
2.13.1. ResNet (*Residual Network)*

Backbone berfungsi untuk memproses input citra dan mengekstraksi *feature map* yang mengandung konten dari input citra tersebut. *Feature map* yang terekstraksi berjenis dari beberapa level abstraksi. Pada umumnya, *backbone* merupakan *convolutional neural network* yang menggunakan arsitektur spesifik berupa ResNet (*Residual Neural Network).*

ResNet atau yang memiliki kepanjangan *Residual Network* (He dkk., 2016) merupakan perkembangan dari arsitektur CNN, dirancang untuk mengatasi masalah *vanishing/exploding gradient.* Masalah tersebut kerap muncul dalam pengembangan metode *deep learning* dengan *layer* yang banyak. Arsitektur ResNet memperkenalkan konsep *Residual Blocks* serta menggunakan teknik *skip connections*. *Skip connections* menghubungkan *layer* aktivasi dan *layer* yang lebih jauh, dengan melewati beberapa *layer* di antaranya. Semua ini membentuk *residual block* dan ResNet dibuat dengan menyusun *residual block* tersebut.

Terdapat beberapa jenis dari ResNet, seperti ResNet-18 dan ResNet-50. Jenis-jenis dari ResNet didapat dari jumlah *layer* yang membentuk jenis arsitektur ResNet tersebut. Sebagai contoh, ResNet-50 terbentuk dari 50 *layer*, hal yang sama juga berlaku untuk ResNet-18. Penelitian ini menggunakan ResNet-18 sebagai *backbone* karena memiliki jumlah parameter yang paling rendah dibandingkan dengan jenis ResNet yang lain.

2.13.2. Metode Transformer



Gambar 2.12 Arsitektur Transformer untuk deteksi objek

Transformer yang dikembangkan oleh Vaswani dkk. (2017) merupakan tipe arsitektur *neural network sequence-to-sequence* yang mentransformasikan atau mengubah urutan input menjadi urutan output. Secara singkat, metode ini dirancang untuk mengatasi masalah NLP karena berfungsi untuk menerima input urutan kata menjadi output prediksi kata. Namun, pada konteks deteksi objek, Transformer bertugas untuk menerima input citra dan menghasilkan output prediksi berupa keberadaan, kelas, serta *bounding box* objek pada citra. Metode pendeteksian objek yang menggunakan Transformer sebagai komponen adalah DETR (*Detection Transformer)* serta model lain yang serupa.

Transformer juga terdiri dari dua bagian, yaitu Encoder dan Decoder (Gambar 2.12) dengan rincian sebagai berikut:

2.13.2.1. Encoder

Secara umum, Encoder (bagian kiri dari Gambar 2.12) bertugas untuk menerima input dan memproses input tersebut agar bisa diproses oleh Decoder, hal ini berlaku untuk subjek apapun, baik pemrosesan citra, maupun bahasa. Pada

subjek deteksi objek, Encoder menerima input citra untuk selanjutnya diproses menjadi *query*, *key*, dan *value.* Setiap layer Encoder memiliki standard arsitektur yang terdiri dari modul *multi-head self-attention* dan *feed forward network* (FFN).

2.13.2.2. Decoder

Secara umum, Decoder (bagian kanan dari Gambar 2.12) bertugas untuk menerima output dari Encoder sebagai input, untuk selanjutnya diproses menjadi sebagai hasil prediksi. Setelah menerima input yang berasal dari output Encoder, yaitu *query*, *key*, dan value, Decoder selanjutnya akan memproses menjadi kehadiran, kelas, serta *bounding box* dari objek yang ada pada citra. Pada Decoder untuk DETR, terdapat perubahan dari Transformer *baseline*, yaitu Decoder DETR menerjemahkan (*decodes*) N objek secara paralel pada setiap *layer* Decoder. Sementara pada Transformer *baseline* menggunakan model autoregresi yang menghasilkan prediksi output sebanyak satu elemen secara satu per satu.

2.13.3. *Feed-Forward Networks (FFNs)*

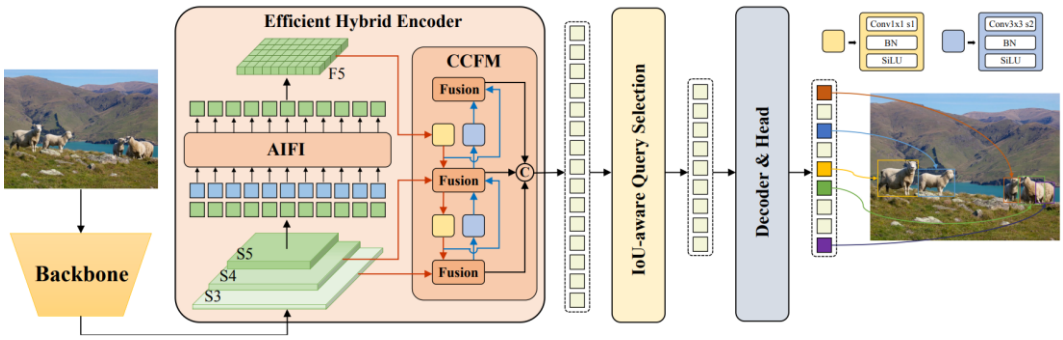
Tahap terakhir pada DETR adalah *feed-forward network* (FFN) yang merupakan variasi, tetapi masih versi umum dari jaringan syaraf tiruan. FFN melakukan prediksi akhir dengan rincian: 3-*layer* perceptron dengan fungsi aktivasi ReLU, *hidden dimension*, dan sebuah *layer* projeksi linier. Hasil dari prediksi yang dilakukan oleh FFN adalah koordinat pusat yang dinormalisasikan, lebar dan tinggi *bounding box* dengan memperhatikan ukuran input citra. Terakhir, *layer* linier akan memprediksi klasifikasi dengan fungsi softmax.

Secara umum, FFN juga dikenal sebagai *Multi-layer Perceptron.* FFN membentuk dasar dari banyak jenis jaringan syaraf tiruan yang digunakan pada masa kini, seperti *Convolutional Neural Networks* (penggunaan untuk *computer vision)* serta *Recurrent Neural Networks* (penggunaan untuk pemrosesan bahasa alami).

2.14. RT-DETR (*Real Time-Detection Transformer)*

RT-DETR atau yang memiliki kepanjangan *Real Time-Detection Transformer* merupakan varian DETR yang memiliki keunggulan untuk mendeteksi secara *real-time* (Lv dkk., 2023). RT-DETR dibuat untuk mengatasi kelemahan DETR *baseline* dalam kecepatan mendeteksi yang diakibatkan oleh *hyperparameter* NMS (*non-maximum supervision*).

Terdapat penelitian lain, serupa dengan RT-DETR yang mengedepankan pemrosesan secara *real-time,* yaitu DEYOv3 (DETR dengan YOLO) (Ouyang, 2023). Namun, metode ini tidak berhasil mengalahkan hasil dari RT-DETR. Dengan RT-DETR berhasil mencapai akurasi sebesar 53.0% pada set data COCO val2017, sedangkan akurasi DEYOv3 hanya mencapai 41.1%. Maka dari itu, dengan keunggulan RT-DETR yang dapat memproses deteksi hingga kecepatan *real-time* serta unggul dari model DETR *real-time* serupa dalam aspek akurasi, metode ini dianggap sesuai untuk digunakan dalam *facial expression recognition.*



Gambar 2.13 Arsitektur RT-DETR

Arsitektur RT-DETR (Gambar 2.13) terdiri dari *backbone, hybrid* encoder, dan decoder dengan *auxiliary prediction heads.* Tiga tahap terakhir yang digunakan sebagai fitur *output* dari *backbone* dimanfaatkan sebagai *input* untuk encoder. *Hybrid* encoder bertugas untuk mentransformasikan fitur *multi-scale* menjadi urutan fitur citra melalui interaksi *intra-scale.* Terakhir, IoU-*aware query selection* digunakan untuk memilih fitur citra dengan jumlah yang tetap dari urutan *output* encoder yang nantinya akan digunakan sebagai *object queries* awal untuk decoder.

Metode ini digunakan sebagai algoritma dalam pengembangan model deteksi wajah. Set data WIDER-face digunakan untuk pengembangan karena foto yang

digunakan diambil pada berbagai kondisi serta terdapat *multi-face* (S. Yang dkk., 2016). Kemudian, dengan adanya anotasi *bounding box* pada set data, menjadikannya cocok dalam pengembangan model.

RT-DETR mengatasi kelemahan DETR yang disebabkan oleh NMS dengan merancang dua komponen, yaitu Encoder *hybrid* dan IoU-*aware query selection.* Penjelasan lebih lanjut mengenai dua komponen tersebut adalah sebagai berikut:

2.14.1. Encoder *Hybrid*

Baidu memperkenalkan konsep pemisahan bagi interaksi fitur *multi-scale* menjadi dua tahap, yaitu interaksi *intra-scale* dan perpaduan *cross-scale*. Sebelumnya, konsep encoder hybrid didapatkan untuk membuktikan bahwa bagian dari metode *Deformable*-DETR, yaitu *multi-scale features,* interaksi *intra-scale* dan *cross-scale* secara bersamaan menghasilkan komputasi yang tidak efisien.

Pendekatan ini bertujuan untuk mengurangi biaya komputasi dan secara bersamaan untuk meningkatkan akurasi. Hybrid encoder terdiri dari dua komponen, yaitu *Attention-based Intra-scale Feature Interaction* (AIFI) yang berfungsi mengurangi redundansi komputasi dengan fokus pada interaksi *intra-scale* pada fitur level atas serta *CNN-based Cross-scale Feature-fusion Module* (CCFM) yang berfungsi untuk mengoptimalkan perpaduan *cross-scale* dengan menggunakan blok *fusion* yang terdiri dari lapisan CNN (Lv dkk., 2023, hlm. 5).

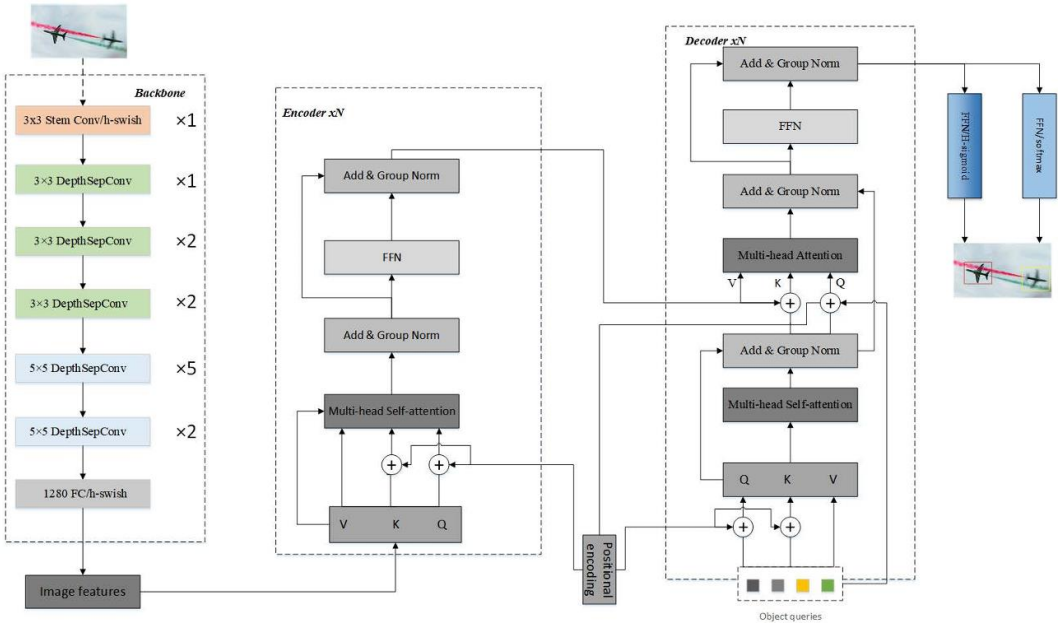
2.14.2. IoU-*aware Query Selection*

Untuk mengatasi *object query* yang sulit diinterpretasikan dan dioptimalkan, IoU *(intersection over union)-aware query selection* mengatur model agar nilai klasifikasi yang tinggi digunakan untuk fitur dengan nilai IOU yang tinggi dan nilai klasifikasi yang rendah digunakan untuk fitur dengan nilai IOU yang rendah ketika *training* model. Hal ini untuk memastikan bahwa fitur encoder yang terpilih bisa memiliki nilai klasifikasi serta IOU tinggi yang mana akan meningkatkan akurasi dari deteksi.

Berdasarkan analisis yang dilakukan pada penelitian RT-DETR, model yang dilatih dengan IoU-*aware Query Selection* dapat menghasilkan *encoder features*

yang berkualitas lebih baik. Selain itu, berdasarkan analisa kuantitatif, komponen ini dapat memberikan lebih banyak *encoder features* dengan klasifikasi akurat (skor klasifikasi yang tinggi) dan lokasi akurat (skor IoU yang tinggi) untuk *object queries*, yang mana akan meningkatkan akurasi dari detektor (Lv dkk., 2023, hlm. 6).

2.15. L-DETR (*Light-Weight Detector for End-to-End Object Detection With Transformers)*



Gambar 2.14 Arsitektur L-DETR

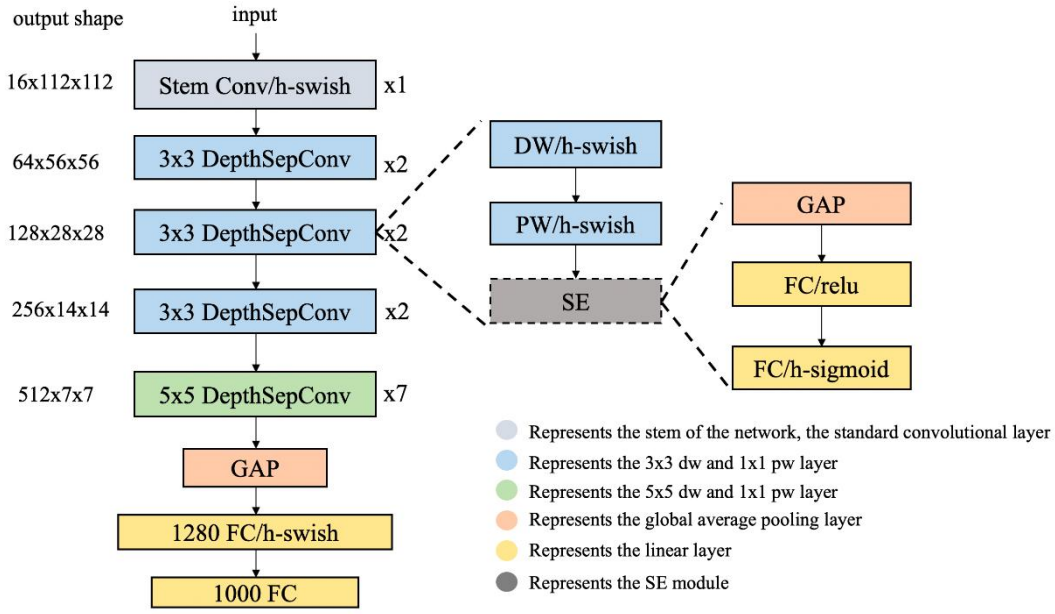
L-DETR yang dikembangkan oleh Li dkk. (2022) yang memiliki kepanjangan *Light-Weight Detector for End-to-End Object Detection With Transformers* merupakan metode DETR (Carion dkk., 2020) yang menggunakan komponen *backbone* berbasis PP-LCNet (*Lightweight CPU Convolutional Neural Network)* yang dikembangkan oleh Cui dkk. (2021). Metode ini dibangun dengan latar belakang masalah keterbatasan penggunaan model kecerdasan buatan dengan performa baik. Menurut penelitian ini, memastikan model dapat digunakan pada perangkat *edge* merupakan cara yang menjanjikan agar model kecerdasan buatan mendapatkan performa *real-time* yang baik. Selain itu, penelitian ini berpendapat bahwa merancang model *light-weight* yang membutuhkan konsumsi sumber daya

yang rendah dengan performa yang tinggi merupakan masalah utama dalam pengembangan model untuk penggunaan luring.

Metode ini terdiri dari dua bagian seperti yang ditunjukkan pada gambar 2.14. Bagian pertama merupakan *backbone* yang berbasis pada PP-LCNet yang telah diimprovisasi yang berfungsi untuk mengekstraksi *feature* data. Bagian kedua merupakan transformer yang telah diimprovisasi yang berfungsi untuk mengkalkulasi informasi global dan membuat prediksi final. Perbedaan *backbone* PP-LCNet dan L-DETR adalah *backbone* yang telah diimprovisasi hanya menggunakan lima modul DepthSepConv (3x3) serta menghilangkan *layer* GAP (*average group pooling*) dan FC (*full connection*).

Penelitian ini hanya akan menggunakan komponen *backbone* L-DETR pada eksperimennya. Hal ini disebabkan karena menggunakan komponen transformer yang telah diimprovisasi dari L-DETR pada metode RT-DETR, dapat mengubah arsitektur utama dari RT-DETR.

2.15.1. PP-LCNet (*Lightweight CPU Convolutional Neural Network*)



Gambar 2.15 Arsitektur PP-LCNet

PP-LCNet yang memiliki kepanjangan *Lightweight CPU Convolutional Neural Network* diadaptasi dari *framework* PaddlePaddle (PP) (Cui dkk., 2021).

Metode ini dikembangkan untuk mengatasi limitasi yang timbul karena MKLDNN (*Math Kernel Library for Deep Neural Networks*). Maka dari itu, PP-LCNet dibuat untuk memenuhi kriteria MKLDNN serta agar dapat mengakselerasi *framework* Deep Learning pada arsitektur Inter(R).

Dapat dilihat juga perbedaan arsitektur PP-LCNet (gambar 2.15) dengan *backbone* L-DETR (gambar 2.14). Perbedaan tersebut, yaitu lima modul DepthSepConv (3x3) serta *layer* GAP (*average group pooling*) dan FC (*full connection*).

Penelitian ini menggunakan *backbone* LCNet dengan skala 0.25 karena memiliki jumlah parameter yang paling sedikit, yaitu 19973379. Pemilihan skala tersebut sesuai dengan permasalahan latar belakang penelitian ini yaitu keterbatasan sumber daya dalam pengembangan perangkat lunak.

2.16. Real-Time CNN (*Convolutional Neural Networks*)

Real-time CNN yang dikembangkan oleh Arriaga dkk. (2017) merupakan metode yang digunakan untuk pengembangan model klasifikasi ekspresi. Metode ini dikembangkan untuk mendeteksi wajah dan mengklasifikasi ekspresi serta gender. Secara keseluruhan, alur metode ini menghabiskan waktu sebanyak (0.22 ± 0.0003 ms). Namun pada penelitian ini, hanya akan digunakan bagian klasifikasi ekspresi saja. Metode ini dikembangkan dengan framework OpenCV menggunakan algoritma CNN (*convolutional neural network*).

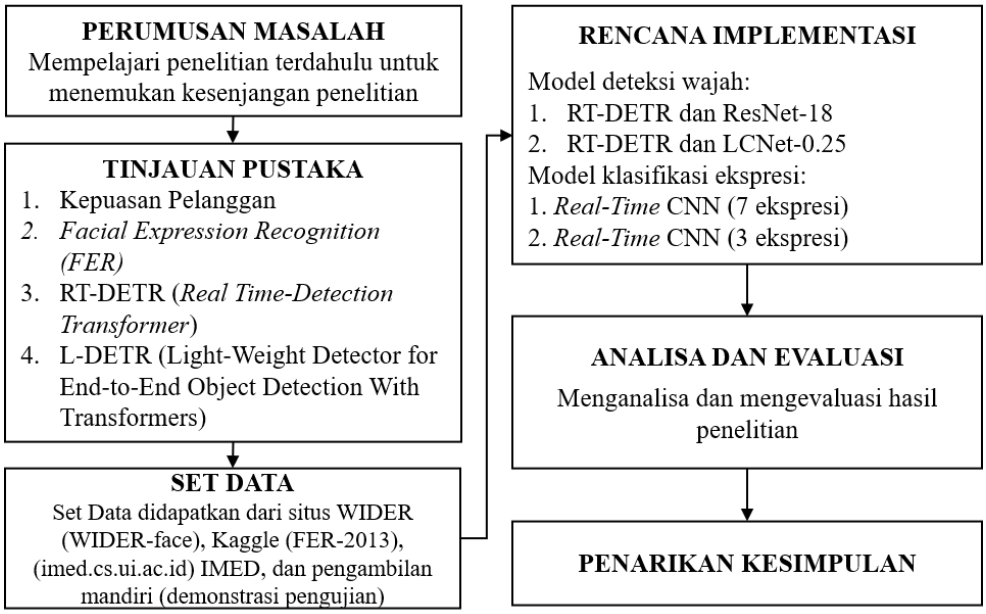
Dengan performa accuracy 66% pada klasifikasi set data FER-2013 serta kemampuannya untuk memprediksi secara real-time, Real Time-CNN dinilai mampu digunakan sebagai bagian model klasifikasi wajah untuk model pengenalan ekspresi wajah.

BAB III
METODE PENELITIAN

Bab ini menjelaskan semua tahapan pembangunan sistem pengenalan ekspresi wajah untuk kepuasan pelanggan. Diawali dengan pengembangan model deteksi wajah menggunakan metode RT-DETR (*Real Time-Detection Transformer*), ResNet-18, dan LCNet *(Lightweight CPU Convolutional Neural Network)* dengan set data WIDER-face. Kemudian, pengembangan model klasifikasi wajah menggunakan metode Real-Time CNN dengan set data FER-2013 (*Facial Expression Recognition*-2013) serta IMED (*The Indonesian Mixed Emotion Dataset*). Hingga terakhir, yaitu pengujian integrasi model deteksi wajah dan klasifikasi ekspresi menggunakan set data demonstrasi pengujian. Metodologi penelitian terdiri dari desain penelitian, rancangan model dan eksperimen, serta lingkungan komputasi eksperimen.

3.1. Desain Penelitian

Penelitian yang dilakukan melibatkan tujuh tahap pengerjaan. Tahapan tersebut terdiri dari perumusan masalah, tinjauan pustaka, pengumpulan data, rancangan model, implementasi, evaluasi, serta penarikan kesimpulan. Seluruh tahapan penelitian tersebut diilustrasikan pada gambar 3.1.



Gambar 3.1 Desain penelitian

Penjelasan lebih lanjut mengenai masing-masing tahapan adalah sebagai berikut:

3.1.1. Rumusan Masalah

Pada perumusan masalah, dilakukan proses identifikasi masalah dengan mempelajari penelitian terdahulu untuk menemukan kesenjangan penelitian yang dapat dijadikan basis dari penelitian yang dilakukan. Dari perumusan masalah ditemukan basis penelitian seperti, latar belakang, masalah utama, tujuan penelitian, dan metode yang akan digunakan.

3.1.2. Tinjauan Pustaka

Setelah perumusan masalah selesai dilakukan merupakan tinjauan pustaka yang berfungsi untuk menguraikan hubungan berbagai teori yang melandasi atau menjadi argumen penelitian. Tinjauan pustaka juga berisi penjelasan tentang variabel-variabel yang diteliti.

3.1.3. Pengumpulan Data

Penelitian ini membutuhkan data untuk melatih model agar dapat mendeteksi ekspresi wajah. Data yang digunakan, yaitu set data WIDER-face, FER-2013, IMED, dan set data demonstrasi pengujian.

3.1.4. Analisa dan Evaluasi

Tahap ini menganalisa serta mengevaluasi hasil eksperimen yang telah dilakukan. Diharapkan dari eksperimen yang telah dilakukan dapat menjawab rumusan masalah yang telah disusun.

3.1.5. Rancangan Implementasi

Pada tahap ini dibangun model sesuai rancangan serta eksperimen penelitian. Pembangunan model deteksi wajah melibatkan RT-DETR dengan ResNet-18 serta LCNet-0.25 dan model klasifikasi ekspresi melibatkan Real-Time CNN menggunakan 7 ekspresi dan 3 ekspresi.

Kemudian, eksperimen dilakukan berdasarkan rancangan model yang telah disusun. Dimulai dari praproses, *training*, pemilihan bobot model terbaik, dan *testing*. Selanjutnya, dilakukan eksperimen pada beberapa jenis model yang terdiri dari varian kombinasi metode yang telah dipilih pada rumusan masalah.

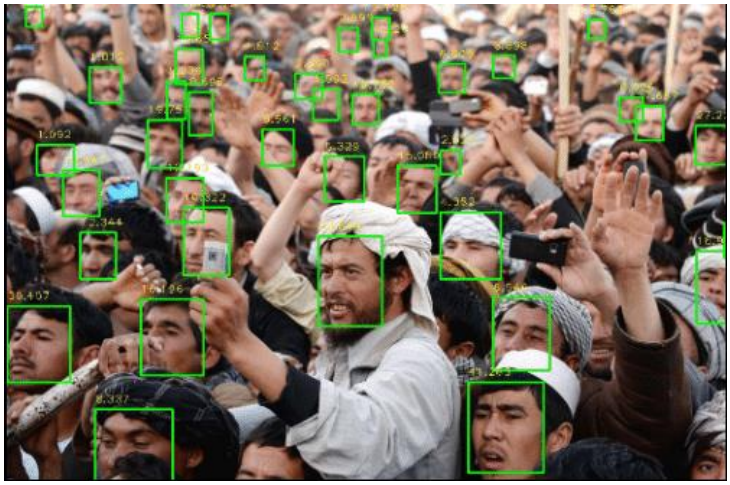
3.1.6. Penarikan Kesimpulan

Tahap terakhir dari penelitian adalah penarikan kesimpulan yang dilakukan dengan cara membandingkan hasil analisa dan evaluasi eksperimen dengan rumusan serta tujuan penelitian. Dari kesimpulan yang telah ditarik, terbentuk juga saran yang akan membantu dalam pengarahan penelitian selanjutnya.

3.2. Set Data

Pada penelitian ini, digunakan empat set data yang berbeda pada model deteksi serta klasifikasi wajah. Set data yang digunakan, yaitu WIDER-face (S. Yang dkk., 2016), FER-2013 (Goodfellow dkk., 2013), IMED (Liliana dkk., 2018) serta set data demonstrasi pengujian.

3.2.1. WIDER-face



Gambar 3.2 Sampel set data WIDER-face

Set data WIDER-face yang dikembangkan oleh S. Yang dkk. (2016) didapatkan dengan mengunduhnya secara langsung pada situs web terkait dan digunakan untuk pengembangan model deteksi wajah, dengan metode RT-DETR (*Real Time-Detection Transformer*), ResNet-18, serta LCNet *(Lightweight CPU Convolutional Neural Network)*. Evaluasi set data WIDER dilakukan dengan mengirimkan hasil prediksi model ke pengurus situs WIDER. Hasil evaluasi akan didapatkan dalam sebuah file MatLab, yang berisi nilai *precision* dan *recall*. Set data ini terdiri dari variasi kondisi seperti kemiringan, pose, ekspresi, halangan, serta pencahayaan. WIDER-face terbagi menjadi tiga bagian dengan persentase

training (40%), *validation* (10%), dan *testing* (50%). Gambar 3.2 merupakan sampel data dari set data WIDER-face.

Kemudian, set data ini digunakan karena memiliki anotasi *bounding box* yang mana sesuai dengan spesifikasi dari metode RT-DETR yang mendeteksi wajah dengan *bounding box*. Dengan kondisi foto pada set data, yaitu *wild* atau tidak mengikuti ketentuan/pose tertentu, menjadikan set data ini sesuai dengan kebutuhan sistem yang nantinya digunakan melalui CCTV yang mendeteksi wajah melalui sudut pandang atas.

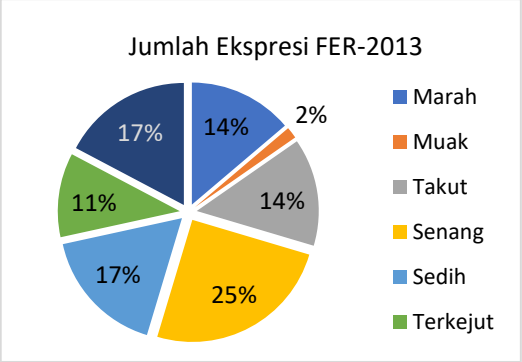
3.2.2. FER-2013 (*Facial Expression Recognition*-2013)

Gambar 3.3 Sampel set data FER-2013

Set data FER-2013 (*Facial Expression Recognition* 2013) yang dikembangkan oleh Goodfellow dkk. (2013) didapatkan dengan registrasi pada tantangan yang diadakan oleh Kaggle dan digunakan untuk pengembangan model klasifikasi ekspresi. Set data dibuat dengan menggunakan fitur pencarian citra pada Google dengan memasangkan nama ekspresi dengan gender, umur, atau etnis. Selanjutnya foto yang telah didapatkan dari Google diproses lebih lanjut agar tidak ada set data duplikat dan salah label. Hasil akhir dari set data adalah data berukuran 48x48 piksel dan terkonversi menjadi *grayscale*. FER-2013 terbagi menjadi tiga bagian dengan persentase training (80%), validation (10%), dan testing (10%). Gambar 3.3 merupakan contoh dari set data FER-2013.

Penelitian ini menggunakan set data FER-2013 yang juga pernah digunakan pada penelitian FER terdahulu (Meena dkk., 2023). Kemudian, kondisi foto pada set data merupakan *wild* atau tidak mengikuti ketentuan/pose tertentu. Kedua alasan tersebut menjadikan set data ini sesuai dengan tujuan pengembangan sistem, yaitu

pengimplementasian pengenalan ekspresi wajah melalui CCTV sudut pandang atas. Selanjutnya foto pada set data juga telah di-*crop* mengikuti batasan wajah yang menjadikan tahap praproses lebih mudah untuk dilakukan.



Gambar 3.4 Bagan jumlah ekspresi FER-2013

Keseluruhan set data FER-2013 berjumlah 35.887 data dengan rincian sebagai yaitu, 4953 foto untuk ekspresi marah, 547 foto untuk ekspresi muak, 5121 foto untuk ekspresi takut, 8989 foto untuk ekspresi senang, 6077 foto untuk ekspresi sedih, 4002 foto untuk ekspresi terkejut, dan 6198 foto untuk ekspresi netral Berdasarkan jumlah data tersebut, kelas muak mengalami *imbalance* dengan persentase yang hanya sebesar 1.52%. Gambar 3.4 merupakan bagan yang merepresentasikan jumlah ekspresi pada set data FER-2013.

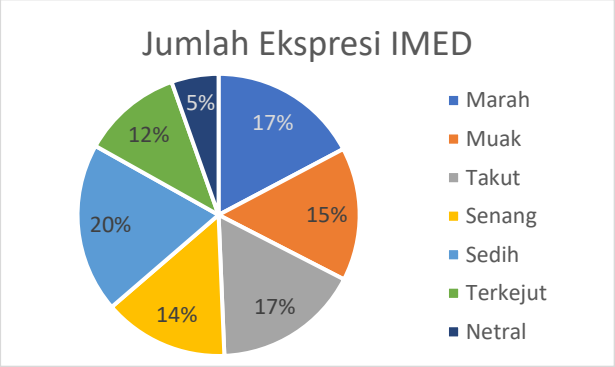
3.2.3. IMED (*Indonesian Mixed Emotion Dataset*)

Gambar 3.5 Sampel set data IMED

Set data IMED yang dibuat oleh Liliana dkk. (2018) didapatkan dengan registrasi di situs IMED dan mengajukan permintaan akses kepada pemilik dataset dan digunakan untuk pengembangan model klasifikasi ekspresi. Pembuatan data

dilakukan dengan mengikuti prosedur yang telah ditentukan, dengan partisipan berjumlah 15 orang yang terdiri dari 9 perempuan dan 6 laki-laki serta berbagai macam suku, yaitu 40% Jawa, 20% Sunda, 13.3% Padang, 13.3% Melayu, 6.7% Batak, dan 6.7% Manado. Terdapat 19 jenis ekspresi yang terdiri dari 7 ekspresi dasar dan 12 ekspresi campuran. Penelitian ini hanya menggunakan 7 ekspresi dasar, yaitu netral, senang, sedih, muak, marah, takut, dan terkejut. Gambar 3.5 merupakan contoh dari set data IMED.

Berdasarkan studi mengenai bias yang terjadi pada klasifikasi gender dan ras oleh Buolamwini & Gebru (2018), model klasifikasi komersil umumnya memiliki performa terbaik pada subjek pria berkulit putih serta terburuk pada perempuan berkulit gelap. Oleh karena itu, untuk mengantisipasi hal tersebut, digunakan IMED agar model bisa memiliki performa lebih baik pada wajah Indonesia karena keunikan etnis subjeknya yang memiliki karakteristik wajah orang Indonesia.



Gambar 3.6 Bagan jumlah ekspresi IMED

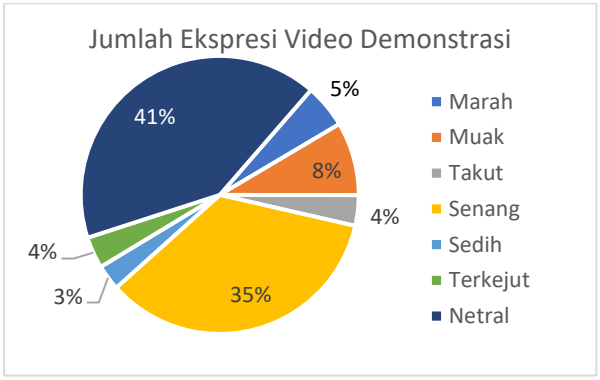
Keseluruhan set data IMED (7 ekspresi dasar) berjumlah 5956 data dengan rincian sebagai yaitu, 1028 foto untuk ekspresi marah, 911 foto untuk ekspresi muak, 1000 foto untuk ekspresi takut, 854 foto untuk ekspresi senang, 1158 foto untuk ekspresi sedih, 683 foto untuk ekspresi terkejut, dan 322 foto untuk ekspresi netral. Berdasarkan jumlah data tersebut, kelas netral mengalami *imbalance* dengan persentase yang hanya sebesar 5.4%. Gambar 3.6 merupakan bagan yang merepresentasikan jumlah ekspresi pada set data.

3.2.4. Set Data Demonstrasi Pengujian

Gambar 3.7 Sampel set data demonstrasi pengujian

Set data ini digunakan untuk pengujian model integrasi deteksi wajah dan klasifikasi wajah. Video diambil secara mandiri dengan subjek mahasiswa program studi Ilmu Komputer Universitas Pendidikan Indonesia. Semua subjek (aktor) memiliki karakteristik wajah orang Indonesia, yaitu kulit berwarna sawo matang hingga kuning kecoklatan, iris mata berwarna coklat kehitaman, rambut cenderung berwarna hitam hingga kecoklatan dengan tekstur lurus hingga ikal bergelombang, serta hidung lebar, tetapi tidak terlalu menonjol.

Agar wajah subjek dapat dievaluasi oleh sistem, kamera diletakkan di sebelah kiri kasir dengan posisi membelakangi cahaya. Subjek juga diharuskan untuk melepas kacamata, agar tidak terdapat pantulan cahaya dan wajah dapat terlihat jelas. Pada akhirnya, video berhasil diambil dengan durasi sepanjang 6 menit 44 detik, format 60 FPS, dan resolusi 1920 x 1080. Namun, karena terdapat subjek dengan wajah yang tidak terlihat jelas, maka video diedit dan menjadi berdurasi 4 menit 42 detik. Gambar 3.7 adalah sampel frame set data video demonstrasi pengujian.

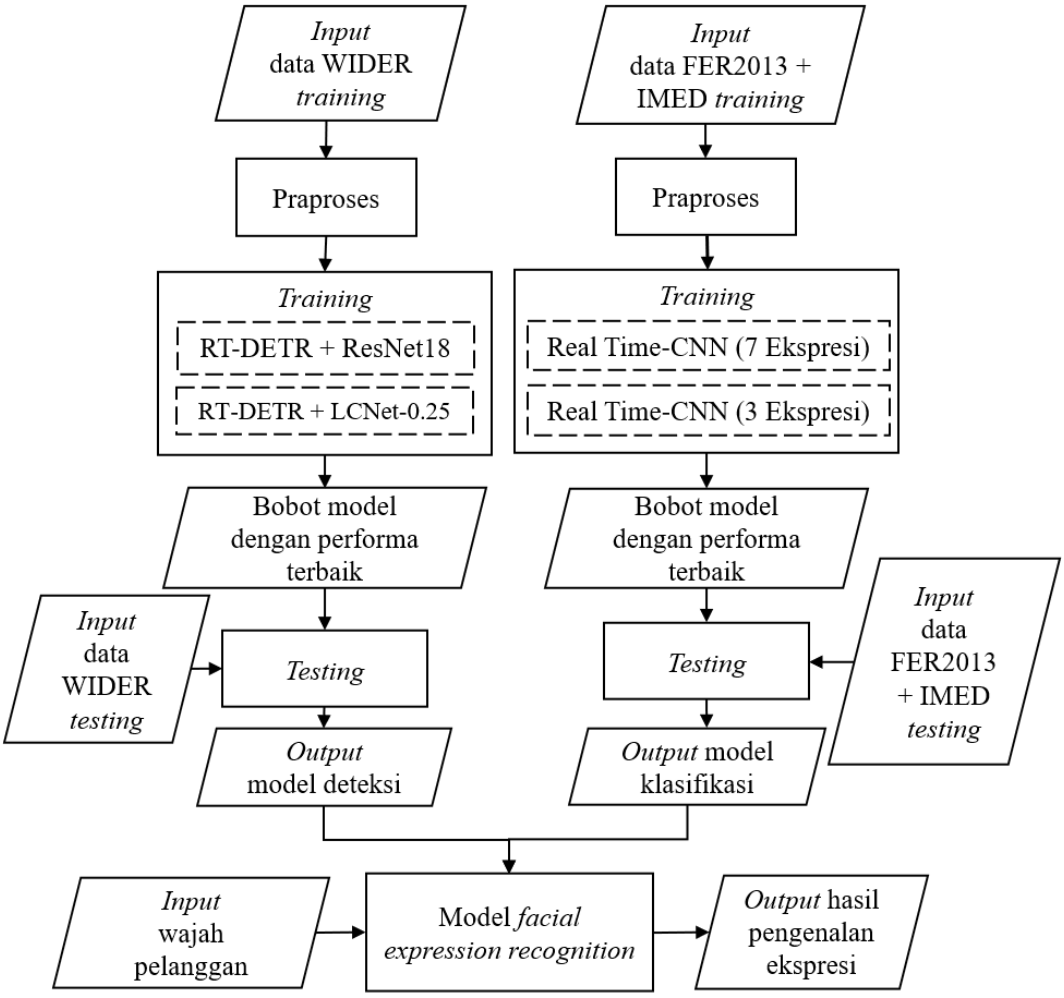


Gambar 3.8 Persentase jenis ekspresi

Pada video ini subjek diarahkan untuk berperilaku sebagai pelanggan pada minimarket, setiap subjek bergiliran menampilkan ekspresi netral, senang, sedih, muak, marah, takut, dan terkejut secara berurutan ketika sedang melakukan transaksi dengan kasir. Sayangnya karena subjek tidak memiliki keterampilan untuk berekspresi seperti aktor, ekspresi yang dipraktikkan subjek tidak serumit yang telah diarahkan. Pada gambar 3.8**,** dapat diketahui pembagian jenis ekspresi yang mampu dilakukan subjek.

Walaupun set data diambil dalam bentuk video, tetapi ketika digunakan untuk pengujian model pengenalan ekspresi wajah, video akan dikonversi menjadi *frames* atau citra (*still image).*

3.3. Rancangan Implementasi



Gambar 3.9 Rancangan implementasi

Pada tahap ini dibangun model sesuai pada rancangan gambar 3.9 yang melibatkan RT-DETR dengan ResNet-18 dan LCNet-0.25, serta Real-Time CNN dengan 7 ekspresi dan 3 ekspresi. Selanjutnya model diimplementasikan menggunakan set data WIDER-face, FER-2013 dan IMED, serta set data demonstrasi pengujian. Eksperimen terdiri dari dua tahap, yaitu tahap deteksi wajah dan klasifikasi ekspresi.

Walaupun metode utama dari penelitian ini, yaitu RT-DETR, memiliki kemampuan untuk deteksi dan klasifikasi secara langsung (*end-to-end*), tetapi set data yang digunakan memiliki keterbatasan. Keterbatasan set data tersebut yaitu tidak adanya set data deteksi wajah yang telah melibatkan jenis kelas ekspresi. Hal tersebut mengakibatkan harus menggunakan dua jenis set data untuk deteksi wajah dan klasifikasi ekspresi secara terpisah.

3.3.1. Model Deteksi Wajah

Model deteksi wajah dikembangkan menggunakan set data WIDER-face. Sebelum melalui tahap *training*, set data WIDER-face melalui tahap praproses yang berfungsi agar data anotasi sesuai dengan format COCO. Hal tersebut bertujuan agar set data dapat diterima oleh RT-DETR. Selanjutnya, pada tahap *training*, dilakukan eksperimen pada tiga kombinasi model, dengan rincian sebagai berikut:

1. RT-DETR dan *backbone* ResNet-18
2. RT-DETR dan *backbone* LCNet-0.25

Setelah *training* selesai dilakukan pada seluruh kombinasi model, selanjutnya dipilih bobot model terbaik berdasarkan performa, yaitu *epoch* terbaik berdasarkan besar IoU dan *average precision* (AP). Kemudian, dengan bobot model terbaik, dilakukan *testing* menggunakan set data WIDER-face. Evaluasi model dilakukan dengan metrik *average precision* (AP), FPS, dan ukuran model (parameter). AP yang digunakan memiliki *threshold* IoU 0.50 hingga 0.95 karena merupakan metrik evaluasi utama yang digunakan pada evaluasi COCO (Lin dkk., 2014).

3.3.2. Model Klasifikasi Ekspresi

Secara umum, pengimplementasian model dimulai dari *input* set data FER-2013 dan IMED *training* dan *testing* yang telah melewati praproses. Tahap praproses dilakukan agar set data IMED sesuai dengan format FER-2013. Format tersebut, yaitu *grayscale*, rasio 1:1, ukuran 48 x 48 piksel, serta dalam format *array* piksel yang disimpan dalam CSV.

Setelah set data IMED selesai dipraproses, maka data ini digabung dengan set data FER-2013. Memasuki tahap *training,* set data digunakan sebagai input untuk melatih model Real-Time CNN. Selanjutnya, model melalui tahap *testing* untuk menguji performa model klasifikasi ekspresi. Evaluasi model dilakukan dengan metrik *precision, recall,* dan F1-*score.*

Berdasarkan gambar 3.6, set data FER-2013 dan IMED, mengalami *imbalance* pada kelas netral. Selain pengembangan menggunakan 7 jenis ekspresi wajah (Ekman, 1970), untuk mempermudah pengevaluasian ekspresi wajah, model klasifikasi wajah juga akan dikembangkan menggunakan 3 jenis emosi saja, yaitu positif, negatif, dan netral. Pengkategorian 3 jenis ekspresi wajah dilakukan berdasarkan penelitian Robinson (2008) dengan rincian yang dapat dilihat pada tabel berikut. Setelah model klasifikasi ekspresi dikembangkan berdasarkan 2 tipe jenis ekspresi, yaitu 7 jenis ekspresi dan 3 jenis ekspresi, selanjutnya model dipilih berdasarkan performa terbaik.

Tabel 3.1 Rincian kategori 7 ekspresi dan 3 ekspresi

7 Ekspresi (Ekman, 1970)	**3 Ekspresi (Robinson, 2008)**
Senang	Positif
Kaget	
Takut	Negatif
Marah	
Sedih	
Muak	
Netral	Netral

3.3.3. Integrasi Model Deteksi Wajah dan Klasifikasi Ekspresi

Setelah model deteksi wajah dan klasifikasi ekspresi selesai dikembangkan, kedua model tersebut diintegrasikan agar terbentuk menjadi sebuah sistem

pengenalan ekspresi wajah. Kemudian, eksperimen dilanjutkan untuk melakukan evaluasi pengenalan ekspresi wajah pada set data demonstrasi pengujian yang telah melewati tahap praproses.

Supaya set data dapat digunakan sebagai data pengujian, video harus diubah menjadi *frame* dengan jumlah *frame* berdasarkan besar FPS yang dapat diproses oleh model integrasi. Selain itu, perlu ada anotasi *ground truth* berupa *bounding box* serta label kelas ekspresi. *Bounding box* dianotasi melalui model yang telah teruji untuk deteksi wajah, yaitu YuNet (Wu dkk., 2023) yang mencapai 81.1% mAP pada set data WIDER. Kemudian, anotasi label kelas ekspresi dilakukan secara manual oleh manusia berdasarkan sistem FACS (*Facial Action Coding System*) (Ekman & Friesen, 1978). Dilakukan juga, penentuan batas ROI agar model tidak memproses wajah pelanggan lain selain yang sedang bertransaksi.

Terakhir, model memproses *input* wajah pelanggan dan menghasilkan *output* berupa hasil pengenalan ekspresi wajah pelanggan. Setelah model selesai memprediksi posisi dan ekspresi wajah. Selanjutnya hasil akan dievaluasi untuk mendapatkan nilai FPS dan *average precision.*

3.4. Analisa dan Evaluasi

Tahap ini berfungsi untuk mengevaluasi kinerja model berdasarkan hasil eksperimen yang telah dilakukan. Berikut merupakan rincian dari seluruh metrik evaluasi yang digunakan:

1) *Accuracy*

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

2) *Precision*

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

3) *Recall*

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

4) F1-*score*

$$\text{F1} - score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

5) *Average precision* (AP)

$$AP = \sum_{k=1}^{k=n-1} [Recall_{k+1} - Recall_k] p_{interp}(r_{i+1})$$

6) *Framerate per second* (FPS)
7) Ukuran model atau jumlah parameter

3.5. Lingkungan Komputasi

Penelitian ini dilakukan pada beberapa perangkat keras dan perangkat lunak tertentu. Rincian spesifikasi dari seluruh perangkat yang dilakukan adalah sebagai berikut:

Tabel 3.2 Spesifikasi perangkat

Laptop pribadi	**Komputer Laboratorium Kecerdasan Buatan**
Perangkat keras	
● RAM 20GB ● SSD 477GB ● 8 CPU AMD Ryzen 5 3500U ● GPU AMD Radeon(TM) Vega 8 Graphics	● RAM 64GB ● SSD 1TB ● 16 CPU Intel(R) Core(TM) i7-6900K ● GPU Nvidia GeForce GTX 1080 Ti
Perangkat lunak	
● Windows 10 ● Google Colab ● Visual Studio Code ● Google Chrome ● Microsoft Edge	● Ubuntu 20.04 LTS ● Jupyter Hub

BAB IV
HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai hasil yang didapat dari penelitian yang telah dilakukan. Terdiri dari pengolahan data, implementasi metode, eksperimen, dan evaluasi. Keseluruhan eksperimen dilakukan pada perangkat komputer Laboratorium Kecerdasan Buatan yang tercantum pada tabel 3.2.

4.1 Praproses Set Data Pengembangan

Penelitian ini menggunakan tiga set data dalam eksperimennya, yaitu WIDER-face, FER-2013, dan IMED. Dengan rincian, set data WIDER-face digunakan untuk model deteksi wajah, set data FER-2013 dan IMED digunakan untuk model klasifikasi ekspresi. Sebelum memulai eksperimen, set data yang digunakan harus melewati tahap praproses terlebih dahulu agar sesuai dengan format input dari metode kecerdasan buatan yang digunakan.

4.1.1 WIDER-face

Secara umum set data WIDER-face terbagi menjadi empat bagian umum yaitu data yang terdiri dari WIDER_train (12876 foto), WIDER_val (3223 foto), dan WIDER_test (16097 foto), serta anotasi yang hanya terdiri dari wider_face_split. Pada penelitian ini, set data disesuaikan dengan kebutuhan RT-DETR yang hanya menerima input seperti format anotasi JSON set data COCO. Maka dari itu, anotasi dari WIDER-face disesuaikan dengan format anotasi tersebut. Format anotasi pada set data WIDER-face terdiri dari direktori file foto serta urutan angka *ground truth.* Sampel format anotasi orisinal serta rincian atribut dari *ground truth* dapat dilihat pada gambar 4.1 serta tabel 4.1.

```
...
13--Interview/13_Interview_Interview_Sequences_13_962.jpg
3
310 50 110 166 0 0 0 0 0 0
440 72 122 170 0 0 0 0 0 0
750 46 112 174 0 0 0 0 0 0
...
```

Gambar 4.1 Sampel anotasi orisinal WIDER-face

Tabel 4.1 Atribut anotasi *ground truth* WIDER-face

Nama atribut	Penjelasan
Bounding box	Jumlah *bounding box* yang ada pada foto
x_1	Koordinat awal *bounding box* pada sumbu x
y_1	Koordinat awal *bounding box* pada sumbu y
w	Lebar kotak bbox (*bounding box*)
h	Tinggi kotak bbox (*bounding box*)
Blur	Tingkatan *blur* (keburaman)
Expression	Jenis ekspresi
Illumination	Jenis pencahayaan
Invalid	Pilihan valid/invalid
Occlusion	Tingkat *occlusion* (halangan) yang menghalangi wajah
Pose	Jenis pose

Format set data WIDER disesuaikan menggunakan program Python. Program ini memiliki alur kerja dengan mengubah file anotasi dalam bentuk TXT menjadi anotasi dalam bentuk JSON. Gambar 4.2 merupakan *pseudocode* dari program yang digunakan.

```
Deklarasi:
img_flag, numdet_flag, start_det_count : boolean
det_count, numdet : integer
det_dict : dictionary
img_file : string
file_orisinal = direktori dari file anotasi orisinal WIDER
foreach file_orisinal do
          if det_count = '0 0 0 0 0 0 0 0 0 0' do //jika tidak ada wajah yang terdeteksi
                    //reset nilai variabel
                    continue
          if img_file = True do
                    //atur nilai varibel untuk baris selanjutnya setelah nama file foto
                    deklarasi: det_dict[img_file]
                    continue
          if numdet_flag = True do
                    //atur nilai variabel untuk jumlah wajah yang terdeteksi
                    if numdet > 0 do //jika terdapat wajah pada foto
                              //atur nilai variabel untuk anotasi koordinat
                    else do //jika tidak terdapat foto
                              //reset variabel
                    continue
          if start_det_count = True do
                    //input anotasi koordinat ke det_dict
          if det_count = numdet do //jika seluruh anotasi sudah diproses
                    //reset variabel
...
//input variable ke file JSON
```

Gambar 4.2 Sampel program format COCO

Selanjutnya, dengan penyesuaian seperti format anotasi COCO, maka terdapat atribut baru yang sebelumnya tidak ada pada anotasi orisinal WIDER-face. Anotasi

baru dari set data WIDER-face terbagi menjadi tiga bagian, yaitu *annotations* (anotasi), *categories* (kategori), dan *images* (foto). Sampel serta rincian atribut anotasi WIDER-face sesuai format COCO dapat dilihat pada gambar 4.3 serta tabel 4.2.


```
{
    "annotations": [
        {
            "area": 18178.0,
            "bbox": [
                449.0, 330.0, 122.0, 149.0
            ],
            "boxes": [
                449.0, 330.0, 122.0, 149.0, 0.0, 0.0, 0.0, 0.0, 0.0,
                0.0
            ],
            "category_id": 1,
            "id": 0,
            "image_id": 0,
            "iscrowd": 0,
            "segmentation": []
        }
        ...
    ],
    "categories": [
        {
            "id": 1,
            "name": "face"
        }
        ...
    ],
    "images": [
        {
            "file_name": "0--Parade/0_Parade_marchingband_1_849.jpg",
            "height": 1024,
            "id": 0,
            "width": 1385
        }
        ...
    ]
}
```

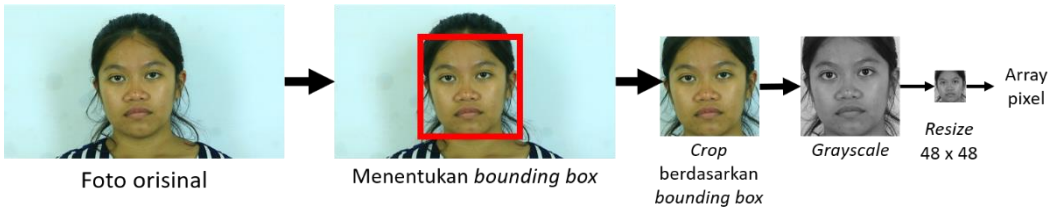

Gambar 4.3 Sampel anotasi WIDER-face sesuai format COCO

Tabel 4.2 Atribut anotasi WIDER-face sesuai format COCO

Nama atribut	**Penjelasan**
Annotations (anotasi)	
area	Luas dari objek/*bounding box*
bbox	Koordinat x_1, koordinat y_1, nilai w, dan nilai h
boxes	Seluruh atribut angka dari anotasi orisinal
category_id	Nomor id yang mengacu pada id dari kategori
id	Id anotasi
image_id	Nomor id yang mengacu pada id dari citra
iscrowd	Boolean penanda objek bergerombol
segmentation	Koordinat dari semua sudut pada objek
Categories (kategori)	
id	Id kategori
name	Nama kategori
Images (foto)	

file_name	Direktori file foto
height	Tinggi foto
id	Id foto
width	Lebar foto

4.1.2 FER-2013 dan IMED



Gambar 4.4 Praproses IMED

Sebelum set data FER-2013 dan IMED digabung, set data IMED harus disesuaikan dahulu agar sesuai dengan format FER-2013. Tahap praproses set data IMED diilustrasikan pada gambar 4.4 diawali dengan menentukan letak wajah dari foto orisinal menggunakan *bounding box*. Penentuan *bounding box* dilakukan menggunakan model YuNet. Dengan angka koordinat *bounding box* yang sudah ditentukan, kemudian foto akan di-crop, agar foto terdiri dari piksel wajah saja. Selanjutnya, foto diubah menjadi *grayscale* untuk mengurangi dimensi *channel* warna. Kemudian, resolusi foto diubah, yang sebelumnya berukuran 220 x 220 menjadi 48 x 48. Terakhir, foto dikonversi menjadi bentuk *array* piksel, agar bisa disimpan di file CSV set data FER-2013. Gambar 4.5 merupakan *pseudocode* dari tahap praproses yang dilakukan dan gambar 4.6 merupakan sampel set data FER-2013 dan IMED.

```
import cv2
// inisialisasi variabel & path foto
output_csv = []
foto_orisinal = "<path foto orisinal>"

bbox_deteksi = YuNet(foto_orisinal) //deteksi foto menggunakan YuNet

// mengubah nama variabel
x1 = bbox_deteksi[0]
y1 = bbox_deteksi[1]
width = bbox_deteksi[2]
height = bbox_deteksi[3]

foto_crop = foto_orisinal[y1 : height, x1 : width] //crop foto berdasarkan bounding box
foto_grayscale = cv2.cvtColor(croped, cv2.COLOR_BGR2GRAY) //mengubah foto menjadi grayscale
foto_resize = cv2.resize(img, (48,48), interpolation = cv2.INTER_AREA) //mengubah ukuran foto menjadi 48 x 48

1_channel = foto_resize[:,:,:1] //mengambil piksel dari 1 channel saja
array_piksel = 1_channel.flatten()//nested list 1_channel, diubah menjadi array 1 dimensi

output.append(array_piksel) //menyimpan array

// menyimpan array piksel dan label ekspresi (didapat dari nama file) ke dalam file .csv
```

Gambar 4.5 Pseudocode praproses IMED

Marah (0) Muak (1) Takut (2) Senang (3) Sedih (4) Kaget (5) Netral (6)

Gambar 4.6 Sampel set data FER-2013 dan IMED

Set data kombinasi, terdiri dari tiga bagian, yaitu *emotion, pixels,* dan *Usage. Emotion* merupakan angka yang merepresentasikan label ekspresi, tabel 4.3 adalah rincian label ekspresi. *Pixels* merupakan foto yang diubah menjadi angka piksel. *Usage* merupakan label pembagian penggunaan set data, dengan rincian yaitu 'Training' untuk *training,* 'PrivateTest' untuk *validation*, dan 'PublicTest' untuk *testing.* Gambar 4.7 merupakan sampel set data FER-2013 dan IMED.

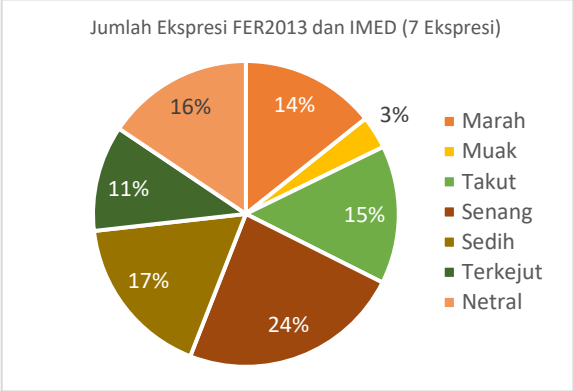
Tabel 4.3 Angka label ekspresi set data FER-2013 dan IMED (7 ekspresi)

Angka	**Label Ekspresi**
0	Marah
1	Muak
2	Takut
3	Senang
4	Sedih
5	Kaget
6	Netral

```
emotion, pixels, Usage
5, 109 105 82 68 50 29 15 13 ... 2304-th pixel, PublicTest
```

Gambar 4.7 Sampel set data FER-2013 dan IMED

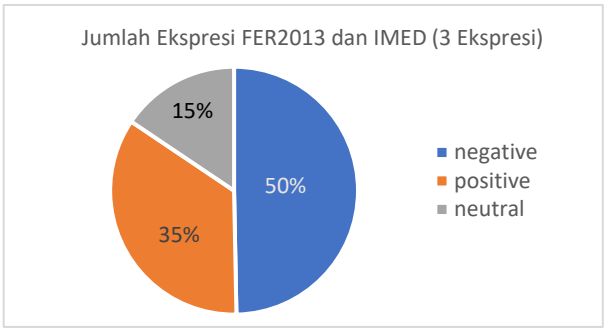
Set data IMED yang telah selesai melewati tahap praproses kemudian dikombinasikan dengan set data FER-2013. Gabungan set data FER-2013 dan IMED dibagi menjadi tiga bagian, dengan persentase *training* (80%), *validation* (10%), dan *testing* (10%), pembagian ini disamakan dengan pembagian FER-2013.



Gambar 4.8 Bagan jumlah ekspresi FER-2013 dan IMED (7 Ekspresi)

Dengan demikian, set data kombinasi FER-2013 dan IMED memiliki total data berjumlah 41.842 foto dan ditemukan bahwa terjadi *imbalance* pada kelas muak dengan persentase sebesar 3.4% . Ilustrasi dari jumlah *imbalance* yang terjadi pada set data FER-2013 dan IMED dapat dilihat pada gambar 4.8.

Selanjutnya, dilakukan praproses untuk mengubah set data menjadi 3 jenis ekspresi saja sesuai dengan pembagian pada tabel 3.1. Dengan pembagian dari 7 jenis ekspresi menjadi 3 jenis ekspresi, didapatkan hasil seperti pada gambar 4.9 berikut. Untuk mengatasi *imbalance* yang terjadi pada kedua jenis ekspresi, maka kombinasi set data menggunakan *treatment* weightedRandomSampler. Untuk penyesuaian praproses menjadi 3 jenis ekspresi, maka pelabelan juga ikut disesuaikan (tabel 4.9).



Gambar 4.9 Bagan jumlah ekspresi FER-2013 dan IMED (3 Ekspresi)

Tabel 4.4 Angka label ekspresi set data FER-2013 dan IMED (3 ekspresi)

Angka	**Label Ekspresi**
0	Negatif
1	Positif

2	Netral

4.2 Model Deteksi Wajah

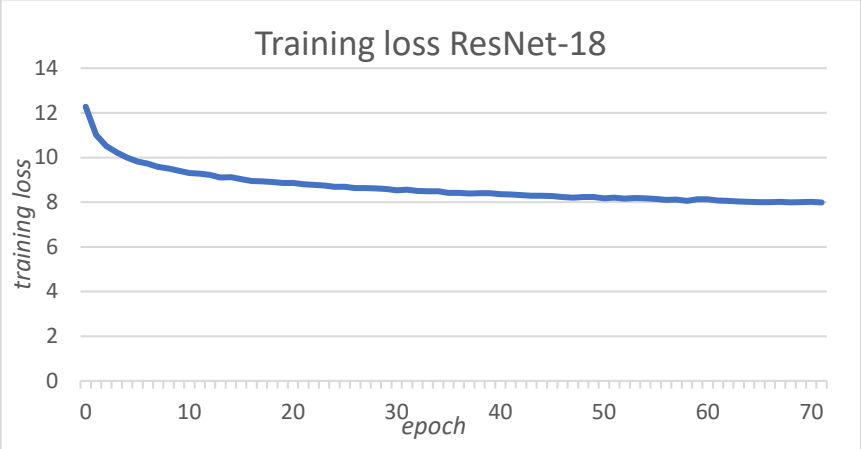
Eksperimen dimulai dengan pengembangan model deteksi wajah. Metode yang digunakan sebagai eksperimen yaitu RT-DETR dengan *backbone* ResNet-18 dan RT-DETR dengan *backbone* LCNet menggunakan set data WIDER. Sampel data *ground truth* WIDER dapat dilihat pada gambar 4.10.

Gambar 4.10 Sampel foto *ground truth*

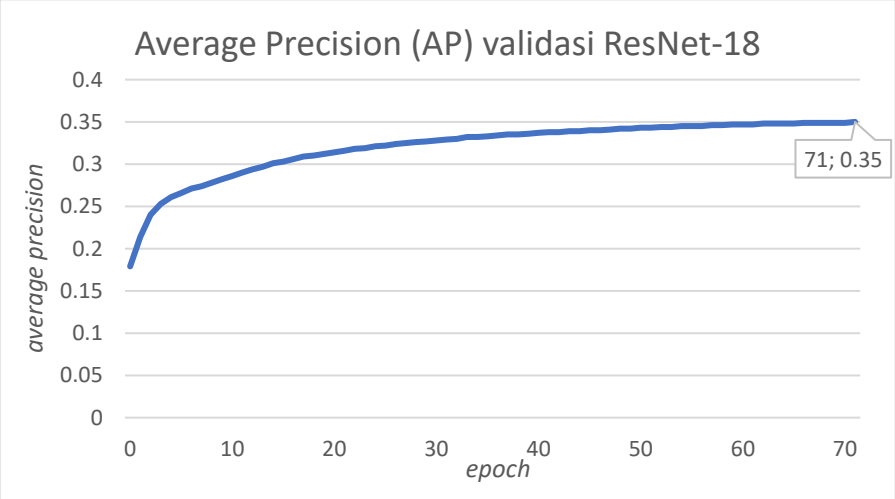
4.2.1 RT-DETR (ResNet-18)

Proses *training* pada model ini dilakukan dengan menggunakan model *pretrained* yang telah dikembangkan pada set data ImageNet. Model *pretrained* tersebut juga digunakan ketika *training* pada penelitian RT-DETR menggunakan set data COCO. Kemudian, model ini dikonfigurasikan sebanyak 72 *epoch.*

Pada akhir eksperimen, model ini memiliki parameter sebesar 20083028 atau 20M dan membutuhkan waktu *training* sebanyak 8 hari 5 jam. Model ini berukuran sebesar 76 MB dalam bentuk format ONNX. Dapat dilihat pada gambar 4.11, *training loss* menurun secara stabil.



Gambar 4.11 *Training* loss ResNet-18



Gambar 4.12 *Average Precision* validasi ResNet-18

RT-DETR dengan *backbone* LCNet-0.25 memiliki nilai AP yang cenderung meningkat (gambar 4.12). Berdasarkan nilai IoU dan AP *threshold* 0.50 hingga 0.95, maka model RT-DETR dengan *backbone* ResNet-18 memiliki performa terbaik pada *epoch* ke-71. Hasil metrik evaluasi model, dapat dilihat pada tabel 4.5. Model ini juga memiliki performa *inference* sebesar 6.64 FPS.

Tabel 4.5 Metrik evaluasi *training* dan validasi *epoch* ke-71

Jenis Metrik	**Nilai**
IoU (coco_eval_bbox)	0.3496
AP dengan IoU=0.50:0.95 (metrik COCO)	35.0
AP dengan IoU=0.50 (metrik PASCAL VOC)	61.7
AP dengan IoU=0.75	35.6

Beberapa perbandingan sampel data *ground truth* dengan hasil prediksi dapat dilihat pada gambar 4.10 dan 4.13. Walaupun terdapat prediksi yang tidak sesuai dengan *ground truth,* tetapi hasil prediksi RT-DETR ResNet-18 dapat dinilai baik.

Gambar 4.13 Sampel foto prediksi dengan RT-DETR (ResNet-18)

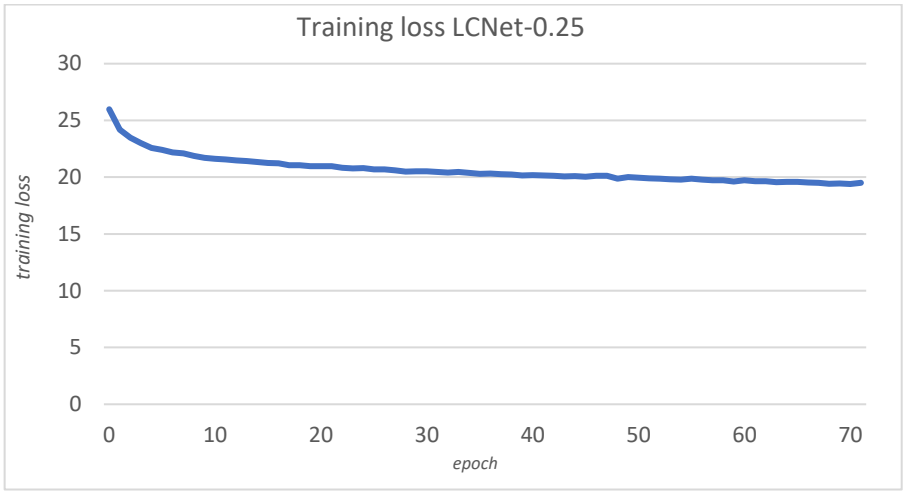
4.2.2 RT-DETR (LCNet-0.25)

Proses *training* pada model ini dilakukan dengan menggunakan model *pretrained* yang dikembangkan dengan set data ImageNet. Model pretrained tersebut terdapat pada *repository* Github PP-LCNet (Cui dkk., 2021). Kemudian, model ini dikonfigurasikan sebanyak 72 *epoch*. *Backbone* LCNet-0.25 memiliki penyesuaian konfigurasi pada file YAML. Konfigurasi dilakukan pada ukuran in_channels dan feat_strides. Besaran kedua parameter tersebut dapat dilihat pada gambar 4.14.

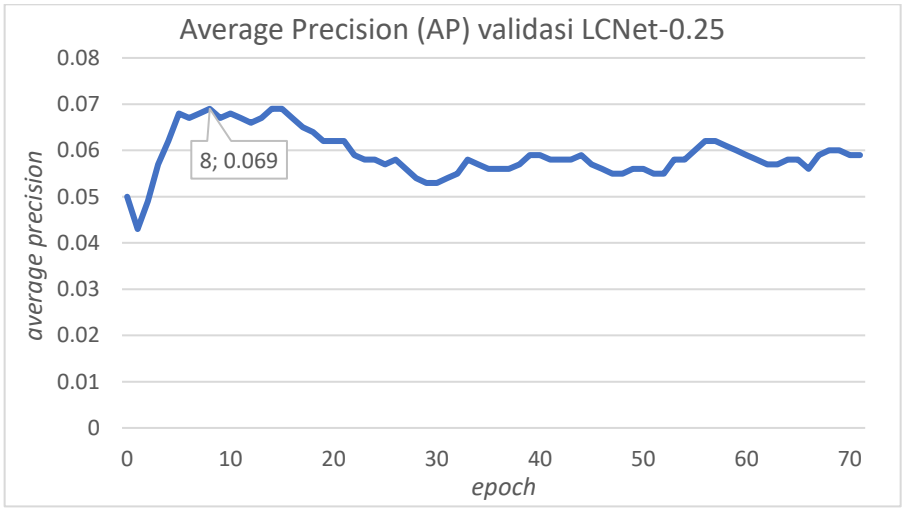
```
HybridEncoder:
  in_channels: [32, 64, 128]
  feat_strides: [8, 16, 32]
```

Gambar 4.14 Konfigurasi backbone LCNet-0.25

Pada akhir eksperimen, model ini memiliki parameter sebesar 19973379 atau 19M dan membutuhkan waktu *training* sebanyak 3 hari 11 jam. Model ini berukuran sebesar 68 MB dalam bentuk format ONNX. Dapat dilihat pada gambar 4.15, *training loss* menurun secara stabil.



Gambar 4.15 *Training* loss LCNet-0.25



Gambar 4.16 Average Precision validasi LCNet-0.25

RT-DETR dengan *backbone* LCNet-0.25 memiliki nilai AP yang fluktuatif (gambar 4.16). Berdasarkan nilai IoU dan AP *threshold* 0.50 hingga 0.95, maka model RT-DETR dengan *backbone* LCNet-0.25 memiliki performa terbaik pada *epoch* ke-8. Hasil metrik evaluasi model, dapat dilihat pada tabel 4.6. Model ini juga memiliki performa *inference* sebesar 5.29 FPS.

Tabel 4.6 Metrik evaluasi training dan validasi *epoch* ke-8

Jenis Metrik	**Nilai**
IoU (coco_eval_bbox)	0.069
AP dengan IoU=0.50:0.95 (metrik COCO)	6.9

AP dengan IoU=0.50 (metrik PASCAL VOC)	14.1
AP dengan IoU=0.75	6.0

Beberapa perbandingan sampel data *ground truth* dengan hasil prediksi dapat dilihat pada gambar 4.10 dan 4.17. Dapat dilihat bahwa model ini kesulitan untuk mendeteksi wajah, bahkan pada wajah yang memiliki tampak depan yang jelas. Maka dari itu, hasil prediksi RT-DETR LCNet-0.25 dapat dinilai buruk.

Gambar 4.17 Sampel foto prediksi dengan RT-DETR LCNet-0.25

4.2.3 Pemilihan Bobot Terbaik (Deteksi Wajah)

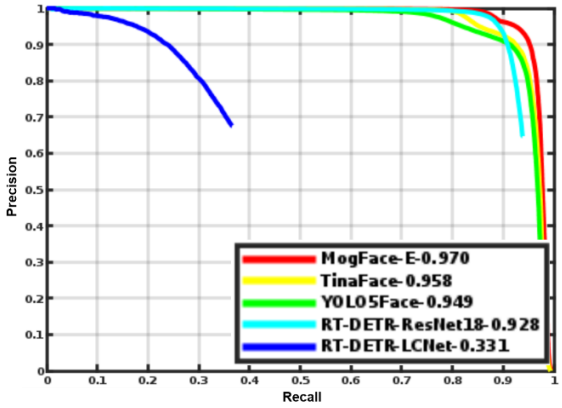
Pemilihan bobot terbaik untuk model deteksi wajah antara RT-DETR dengan *backbone* ResNet-18 dan RT-DETR dengan *backbone* LCNet-0.25 dilakukan berdasarkan hasil evaluasi pengujian set data WIDER, yaitu *average precision* (AP), FPS, dan ukuran model (parameter). Tabel 4.7 berikut merupakan hasil pengujian yang didapatkan dari tahap *training* dan validasi.

Tabel 4.7 Hasil pengujian RT-DETR dengan ResNet-18 dan LCNet-0.25

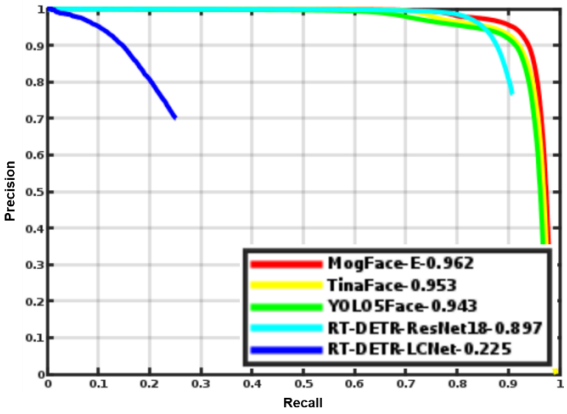
	ResNet-18	**LCNet-0.25**
Epoch terbaik	71	8
Waktu *training*	8 hari 5 jam	3 hari 11 jam
Parameter	20083028	19973379
Ukuran model (MB)	76 MB	68 MB
AP (COCO)	35.0 AP	6.9 AP
AP (PASCAL VOC)	61.7 AP	14.1 AP
FPS	6.64 FPS	5.29 FPS

Berikut merupakan hasil evaluasi test set WIDER yaitu *average precision* (AP) sesuai standar metrik set data COCO dan terbagi menjadi tiga jenis, yaitu *easy*,

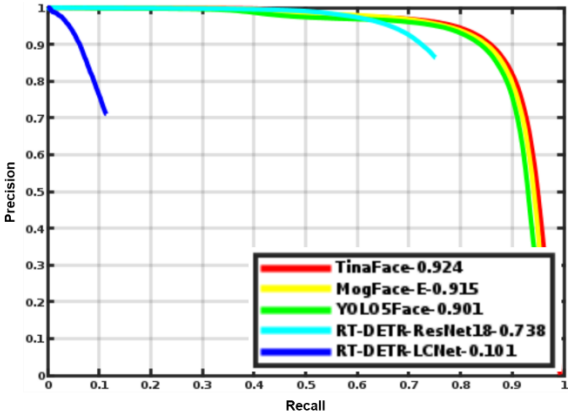
medium, dan *hard*. Evaluasi *test set* WIDER dapat dilihat pada gambar 4.18, 4.19, dan 4.20 serta tabel 4.8 untuk ringkasan hasil evaluasi. Hasil evaluasi dibandingkan dengan dua model terbaik WIDER, yaitu MogFace-E (Liu dkk., 2022) dan TinaFace (Zhu dkk., 2020) serta model YOLO5Face (Qi dkk., 2023) yang merupakan varian model YOLO dengan fokus deteksi wajah.



Gambar 4.18 Evaluasi easy-set WIDER



Gambar 4.19 Evaluasi medium-set WIDER



Gambar 4.20 Evaluasi hard-set WIDER

Tabel 4.8 Hasil evaluasi test-set WIDER

Evaluasi WIDER (AP)	**ResNet-18**	**LCNet-0.25**
Easy	92.8 AP	33.1 AP
Medium	89.7 AP	22.5 AP
Hard	73.8 AP	10.1 AP

Berdasarkan hasil evaluasi WIDER tersebut, kedua model memiliki performa terbaik pada *easy-set* dan performa terburuk pada *hard-set.* Selanjutnya, secara umum *backbone* ResNet-18 memiliki hasil yang lebih baik dibandingkan dengan *backbone* LCNet-0.25. Meskipun model *backbone* LCNet-0.25 memiliki ukuran yang lebih kecil dibandingkan dengan model *backbone* ResNet-18, tetapi performa model dalam pendeteksian juga tetap harus diperhatikan. Hal ini disebabkan karena walaupun model *backbone* ResNet-18 memiliki ukuran model yang lebih besar, performa yang dimiliki lebih baik dibandingkan dengan model *backbone* LCNet-0.25. Maka dari itu, model deteksi wajah akan menggunakan model RT-DETR dengan *backbone* ResNet-18.

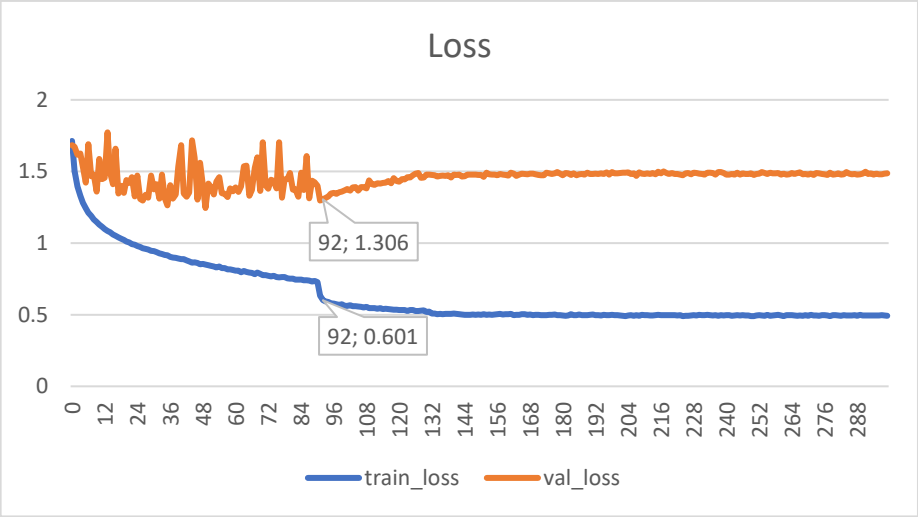
4.3 Model Klasifikasi Ekspresi

Eksperimen dilanjutkan dengan pengembangan model klasifikasi ekspresi. Metode yang digunakan sebagai eksperimen adalah Real Time-CNN menggunakan set data FER-2013 dan IMED dengan 2 tipe jenis ekspresi, yaitu 7 jenis ekspresi dan 3 jenis ekspresi.

Sehubungan dengan *imbalance* yang dialami oleh set data FER-2013 dan IMED (gambar 4.8 dan gambar 4.9), model Real-Time CNN diberikan fungsi tambahan untuk mengatasi *imbalance* tersebut. *Treatment* yang dilakukan adalah dengan menambahkan fungsi WeightedRandomSampler yang bekerja dengan cara memberikan *weights* pada masing-masing jenis kelas. *Weights* tersebut digunakan untuk menentukan probabilitas digunakannya jenis kelas tersebut ketika *sampling.* Model ini memiliki konfigurasi 300 *epoch*, membutuhkan waktu *training* sebanyak 4 jam, serta memiliki ukuran sebesar 261 KB.

4.3.1 Real Time-CNN (7 Ekspresi)

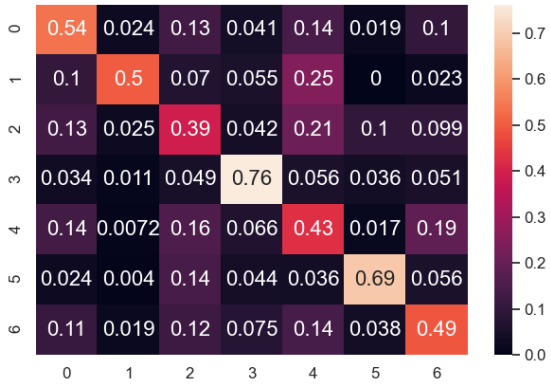
Pada model ini, jenis ekspresi yang digunakan berjumlah 7 yang terdiri dari Marah, Muak, Takut, Senang, Sedih, Kaget, dan Netral. Dapat dilihat pada gambar 4.21 bahwa model Real Time-CNN memiliki loss validasi yang fluktuatif dan loss *training* yang relatif menurun, hal ini dapat diartikan bahwa model mengalami set data validasi yang kurang representatif. Namun, terlihat mulai dari *epoch* ke-92 bahwa loss *training* menurun dan loss validasi meningkat, hal ini menandakan bahwa model mengalami *overfit.*



Gambar 4.21 Real Time-CNN (7 ekspresi) *training* dan *validation* loss

Berdasarkan grafik di atas, model ini memiliki performa terbaik, yaitu loss *training* dan *validation* terendah pada *epoch* ke-92, dengan loss *validation* sebesar 1.306 dan loss *training* sebesar 0.601. Model ini juga memiliki performa *inference* sebesar 48.6 FPS.

Selanjutnya, dapat dilihat *confusion matrix* Real Time-CNN dari *epoch* ke-92 pada gambar 4.21 dan label kelas yang digunakan merupakan label sama yang digunakan pada set data FER-2013 dan IMED (tabel 4.3).



Gambar 4.22 *Confusion matrix* Real Time-CNN (7 ekspresi) *epoch* ke-92

Secara umum, model Real Time-CNN (7 ekspresi) masih mengalami i*mbalance* walaupun telah diberikan *treatment* WeightedRandomSampler. Sebagai contoh, kesalahan prediksi paling sering terjadi pada kelas Muak yang diprediksi menjadi kelas Sedih (baris 1; kolom 4; 0,25).

Model Real Time-CNN (7 ekspresi), memiliki akurasi sebesar 0.487, *precision* 0.437, dan *recall* 0.459. Kemudian hasil evaluasi F1-score per kelas dan rata-rata dapat dilihat pada tabel 4.9.

Tabel 4.9 Evaluasi F1-score model Real Time-CNN (7 ekspresi)

Ekspresi	**F1-score**	***Average* F1-score**	
		Micro	Macro
Marah	0.408	0.487	0.444
Muak	0.292		
Takut	0.265		
Senang	0.722		
Sedih	0.355		
Kaget	0.655		
Netral	0.415		

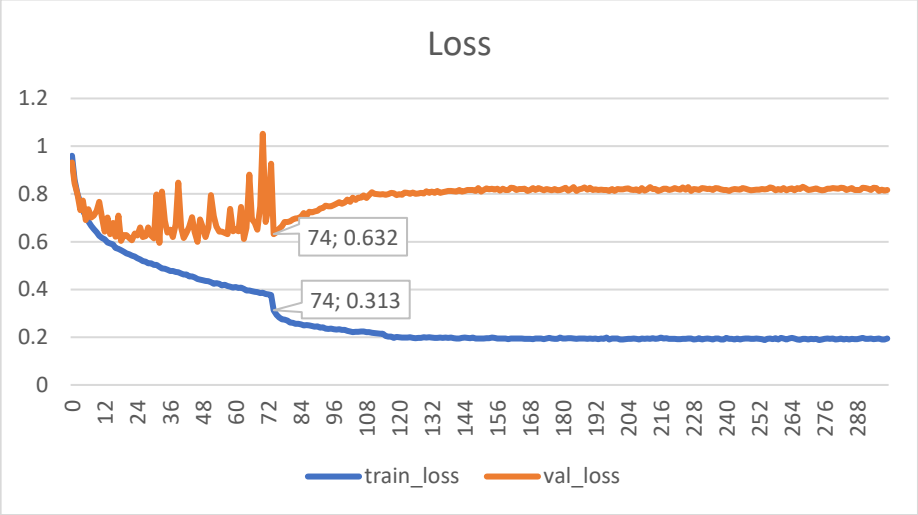
Berdasarkan tabel di atas, model Real Time-CNN (7 ekspresi) memiliki performa yang buruk. Real Time-CNN memiliki performa klasifikasi keseluruhan sebesar 0.487 (*micro average* F1-score) dan performa pada masing-masing kelas sebesar 0.444 (*macro average* F1-score). Selain itu, dengan rata-rata performa per kelas sebesar 0.4, kelas ekspresi Muak, Takut, dan Sedih memiliki performa yang buruk karena memiliki besar F1-score di bawah rata-rata. Sampel dari hasil prediksi masing-masing jenis prediksi, dapat dilihat pada tabel 4.10.

Tabel 4.10 Sampel *ground truth* dan prediksi (7 ekspresi)

Foto	***Ground truth***	**Prediksi**
	Marah	Marah
	Muak	Marah
	Takut	Takut
	Senang	Senang
	Sedih	Sedih
	Kaget	Kaget
	Netral	Netral

4.3.2 Real Time-CNN (3 Ekspresi)

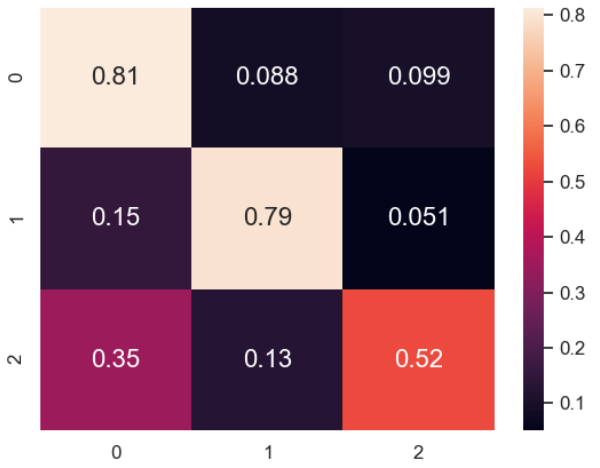
Pada model ini, jenis ekspresi yang digunakan berjumlah 3 yang terdiri dari Positif, Negatif, dan Netral. Dapat dilihat pada gambar 4.23 bahwa model Real Time-CNN memiliki loss validasi yang fluktuatif dan loss *training* yang relatif menurun, hal ini dapat diartikan bahwa model mengalami set data validasi yang kurang representatif. Namun, terlihat mulai dari *epoch* ke-74 bahwa loss *training* menurun dan loss validasi meningkat, hal ini menandakan bahwa model mengalami *overfit.*



Gambar 4.23 Real Time-CNN (3 ekspresi) training dan validation loss

Berdasarkan grafik di atas, model ini memiliki performa terbaik, yaitu loss *training* dan *validation* terendah pada *epoch* ke-74, dengan loss *validation* sebesar 0.632 dan loss *training* sebesar 0.313. Model ini juga memiliki performa *inference* sebesar 48.1 FPS.

Selanjutnya, dapat dilihat *confusion matrix* Real Time-CNN dari *epoch* ke-74 pada gambar 4.24 dan label kelas yang digunakan merupakan label sama yang digunakan pada set data FER-2013 dan IMED (tabel 4.4).



Gambar 4.24 *Confusion matrix* Real Time-CNN (3 ekspresi)

Walaupun tidak seburuk performa *confusion matrix* pada Real Time-CNN (7 ekspresi), model Real Time-CNN (3 ekspresi) masih mengalami i*mbalance* walaupun telah diberikan *treatment* WeightedRandomSampler. Sebagai contoh, kesalahan prediksi paling sering terjadi pada kelas Netral yang diprediksi menjadi kelas Positif (baris 2; kolom 0; 0.35).

Model Real Time-CNN (3 ekspresi), memiliki akurasi sebesar 0.684, *precision* 0.638, dan *recall* 0.653. Kemudian hasil evaluasi F1-score per kelas dan rata-rata dapat dilihat pada tabel 4.11.

Tabel 4.11 Evaluasi F1-score model Real Time-CNN (3 ekspresi)

Ekspresi	**F1-score**	***Average* F1-score**	
		Micro	Macro
Negatif	0.708	0.684	0.643
Positif	0.755		
Netral	0.464		

Berdasarkan tabel di atas, model Real Time-CNN (3 ekspresi) memiliki performa yang baik. Real Time-CNN memiliki performa klasifikasi keseluruhan sebesar 0.684 (*micro average* F1-score) dan performa pada masing-masing kelas sebesar 0.643 (*macro average* F1-score). Sampel dari hasil prediksi masing-masing jenis prediksi, dapat dilihat pada tabel 4.12.

Tabel 4.12 Sampel *ground truth* dan prediksi (3 ekspresi)

Foto	***Ground truth***	**Prediksi**
	Negatif	Negatif
	Positif	Positif
	Netral	Netral

4.3.3 Pemilihan Bobot Terbaik (Klasifikasi Ekspresi)

Pemilihan bobot terbaik untuk model klasifikasi antara Real Time-CNN dengan 7 ekspresi dan 3 ekspresi dilakukan berdasarkan hasil evaluasi pengujian set data FER-2013 dan IMED, yaitu *accuracy*, *macro average* F1-score, dan FPS. Tabel 4.13 berikut merupakan hasil pengujian yang didapatkan dari tahap pengujian. Dapat disimpulkan bahwa model Real Time-CNN dengan 3 jenis ekspresi memiliki performa terbaik dengan *macro-average* F1-score sebesar 0.643 dan *accuracy* sebesar 0.684.

Tabel 4.13 Evaluasi F1-score model Real Time-CNN

	7 Ekspresi	**3 Ekspresi**
Epoch terbaik	92	74
Accuracy	0.487	0.684
Precision	0.437	0.638
Recall	0.45	0.653
F1-score (*macro*)	0.444	0.643
F1-score (*micro*)	0.487	0.684
FPS	48.6	48.1

4.4 Praproses Set Data Pengujian

Sebelum set data demonstrasi pengujian digunakan untuk model integrasi deteksi wajah dan klasifikasi ekspresi, set data harus melalui tahap praproses. Tahap praproses terdiri dari penentuan ROI, konversi video ke *frame* berdasarkan nilai FPS model deteksi wajah terbaik, serta penentuan *ground truth* secara otomatis menggunakan YuNet untuk *bounding box* dan secara manual untuk label klasifikasi.

Gambar 4.25 Sampel *frame* video dengan ROI (garis biru muda)

Dengan *frame* yang memiliki resolusi 1920 x 1080, ROI dimulai dari piksel $x = 960$, $y = 50$, $lebar = 497$, dan $tinggi = 460$. Gambar 4.25 adalah sampel frame yang telah dibatasi oleh ROI (garis biru muda). Selanjutnya, video diubah menjadi bentuk *frame* agar dapat diberikan anotasi *bounding box* dan label klasifikasi. Dengan adanya ROI, wajah yang terdapat pada *frame* berjumlah 1 hingga 3 wajah.

Set data demonstrasi pengujian memiliki *frame rate* sebesar 60 FPS, tetapi karena model deteksi wajah (RT-DETR dengan ResNet-18) memiliki *frame rate* sebesar 6.64 FPS. Maka dari itu, untuk menyesuaikan dengan *frame rate* model, maka *frame rate* set data demonstrasi pengujian diubah menjadi 6 FPS. *Frame rate* tersebut akan digunakan sebagai basis dalam konversi video ke *frame* foto. Total *frame* foto setelah dikonversi adalah sebanyak 1.695 *frame.*

```
7 336 226 90 111
4 103 111 120 176
```

(a) *Frame* (b) Anotasi

Gambar 4.26 Sampel *ground truth*

Selanjutnya, masing-masing *frame* diberikan anotasi *ground truth* dan label klasifikasi ekspresi. Sampel *frame* dan anotasi *ground truth* dapat dilihat pada gambar 4.26. Setiap *frame ground truth* (.jpeg) memiliki 1 file anotasi dengan nama yang sama, yaitu (.txt). Pada tiap file anotasi memiliki jumlah baris yang sama dengan jumlah wajah yang ada pada *frame ground truth.* Masing-masing baris memiliki 5 angka yang dipisahkan oleh spasi. Angka pertama menandakan label ekspresi sesuai tabel 4.3 dan angka kedua hingga kelima menandakan *bounding box* dengan format koordinat sumbu-X, koordinat sumbu-Y, *width,* dan *height.* Namun

karena model klasifikasi ekspresi memiliki performa terbaik pada model dengan 3 jenis ekspresi, maka anotasi label ekspresi disesuaikan kembali sesuai tabel 4.4.

4.5 Integrasi Model Deteksi dan Klasifikasi

Model pengenalan ekspresi wajah terdiri dari model deteksi wajah dengan RT-DETR *backbone* ResNet-18 dan model klasifikasi ekspresi dengan Real Time-CNN (3 ekspresi) dan diuji menggunakan set data demonstrasi pengujian. Integrasi model deteksi dan klasifikasi dievaluasi menggunakan fungsi COCO eval (Lin dkk., 2014) yang juga digunakan untuk mengevaluasi RT-DETR.

Model pengenalan ekspresi wajah memiliki performa *frame rate* sebesar 5.94 FPS dan 77.5% AP untuk *bounding box* saja serta 4.7% AP untuk keseluruhan. Tabel 4.14 merupakan hasil evaluasi integrasi model. Berdasarkan hasil tersebut, model ini memiliki performa baik, jika mendeteksi *bounding box* wajah saja, tetapi memiliki performa buruk, jika harus mengklasifikasikan jenis ekspresi dan mendeteksi wajah.

Tabel 4.14 Hasil evaluasi model integrasi

	Bounding box	*Bounding box* dan label ekspresi
AP | IoU=0.50:0.95 (COCO)	77.5	4.7
AP | IoU=0.50 (PASCAL VOC)	98.1	6.1
AP | IoU=0.75	94.8	5.8

Berikut pada gambar 4.27 merupakan sampel *ground truth* dan prediksi dari model pengenalan ekspresi wajah. Dapat dilihat bahwa model mengalami kesalahan klasifikasi pada Netral dan memprediksi menjadi Positif serta sebaliknya.

a) *Ground truth* b) Prediksi

Gambar 4.27 Sampel gambar *ground truth* dan prediksi dengan anotasi

4.6 Pembahasan

Berdasarkan performa RT-DETR dengan *backbone* ResNet-18 pada eksperimen ini, dapat disimpulkan bahwa model yang dilatih pada set data WIDER memiliki performa yang lebih buruk dibandingkan model yang dilatih pada set data COCO. Performa model RT-DETR dengan *backbone* ResNet-18 dapat dilihat pada tabel berikut.

Tabel 4.15 Performa RT-DETR (ResNet-18) pada seluruh set data

RT-DETR (ResNet-18)	AP | IoU=0.50:0.95 (COCO)	AP | IoU=0.50 (PASCAL VOC)	Ukuran set data
WIDER	35.0	61.7	32 ribu foto
COCO	46.4	63.7	123 ribu foto
COCO + Objects365	49.0	66.5	1.923 ribu foto

Penyebab model RT-DETR (ResNet-18) memiliki performa yang lebih buruk dibandingkan model lain dapat disebabkan oleh besar set data yang lebih kecil dibandingkan set data yang lain. Hal tersebut sesuai dengan pernyataan pada survey penelitian mengenai *deep learning* oleh Alom dkk. (2019, hlm. 7) yang menyatakan bahwa performa model *deep learning* meningkat seiring bertambahnya ukuran set data. Dapat dilihat pada tabel 4.15, model dengan set data WIDER (ukuran terkecil) memiliki performa terburuk.

Kemudian, pada tabel 4.5 dan 4.6 dapat dilihat bahwa RT-DETR (LCNet-0.25) memiliki performa buruk dibandingkan RT-DETR (ResNet-18). Hal ini dapat disebabkan karena LCNet merupakan *backbone* dengan kompleksitas ringan. Dugaan ini diperkuat dengan penelitian survey *backbone* pada ranah *computer*

vision oleh Goldblum dkk. (2023, hlm. 24) yang dapat dilihat pada tabel 4.16. Secara umum, dapat dikatakan bahwa kompleksitas model atau ukuran parameter memengaruhi performa model.

Tabel 4.16 Tabel evaluasi parameter dan performa

Backbone	**Parameter (M)**	**AP**
ResNet-50	82 M	46.6
ResNet-101	101M	47.7
ViT-Small	84M	48.2
ConvNeXt-Tiny	86M	49.9
SwinV2-Tiny	86M	50.2
ViT-Base	155M	51.3
SwinV2-Base-w8	145M	52.4
SwinV2-Base-w24	145M	52.9
ConvNeXt -Base	146M	52.9

Selanjutnya, pada model klasifikasi ekspresi, model Real-Time CNN mengalami *training* loss yang cenderung menurun dan *validation* loss yang fluktuatif. Dapat dikatakan hal tersebut terjadi karena model Real Time-CNN mengalami *underrepresentative validation set* atau set data validasi yang digunakan kurang representatif. Untuk mengatasi hal tersebut, penelitian selanjutnya dapat dilakukan *treatment* augmentasi dataset seperti SMOTE atau mengubah rasio pembagian set data.

Kemudian, karena terdapat set data yang digunakan pada Real Time-CNN, mengalami simplifikasi, yaitu pengurangan menjadi 3 ekspresi saja, model Real Time-CNN mengalami peningkatan performa. Pada penelitian Real Time-CNN menggunakan set data FER-2013 oleh Arriaga dkk. (2017), performa yang didapatkan adalah *accuracy* sebesar 66%. Sedangkan pada penelitian ini yang menggunakan set data FER-2013 dan IMED (3 ekspresi), performa mengalami peningkatan hingga memiliki *accuracy* sebesar 68.4%.

Tabel 4.17 Sampel prediksi benar dan salah

		Prediksi Salah (image-1297)	Prediksi Benar (image-1328)
Ground truth	Anotasi	1 189 146 120 163	2 184 192 122 166
	Frame foto		
Prediksi	Anotasi	2 190 146 115 166	2 186 202 117 156
	Frame foto		

Terakhir, integrasi model deteksi wajah dan klasifikasi ekspresi memiliki performa yang buruk pada evaluasi *bounding box* dan label klasifikasi. Hal ini dikarenakan performa buruk model klasifikasi ekspresi, jika evaluasi hanya dilakukan pada *bounding box* saja, maka memiliki performa yang baik. Sampel prediksi dapat dilihat pada tabel 4.17.

Model Real Time-CNN memiliki performa yang baik pada pengujian menggunakan set data FER-2013 dan IMED, tetapi performa yang buruk pada pengujian menggunakan set data demonstrasi pengujian. Hal tersebut dapat disebabkan oleh format penentuan kelas ekspresi yang berbeda. Set data demonstrasi pengujian menggunakan format FACS oleh Ekman & Friesen (1978), sedangkan set data IMED menggunakan *facial point* sebagai deskripsi fitur dan set data FER-2013 tidak diketahui menggunakan format apa.

Gambar 4.28 Perbandingan posisi wajah set data IMED (kiri) dan demonstrasi pengujian (kanan)

Selain itu, buruknya performa model klasifikasi ekspresi juga dapat disebabkan oleh ketidakberagaman posisi wajah pada set data IMED. Dapat dilihat pada gambar 4.28 bahwa foto IMED memiliki posisi wajah, yaitu *frontal* (sisi depan), yang berbeda dengan posisi wajah set data demonstrasi pengujian, yaitu *profile* (sisi samping). Walaupun pada tahap pengembangan model klasifikasi ekspresi juga menggunakan set data FER-2013 yang memiliki kondisi yang *wild*, tetapi set data tersebut tidak memiliki karakteristik wajah orang Indonesia.

Dengan performa model integrasi yang hanya mencapai 4.7% AP, dapat dikatakan model ini masih belum siap untuk diimplementasikan di dunia nyata. Walaupun model integrasi memiliki performa yang baik pada deteksi wajah, tetapi tujuan utama dari pengembangan sistem ini adalah untuk mengevaluasi berdasarkan ekspresi wajah yang mengandalkan performa dari model klasifikasi ekspresi. Jika ekspresi wajah yang dilakukan oleh pelanggan tidak dapat diprediksi dengan baik, maka bisnis tidak dapat mendapatkan manfaat dari pengimplementasian kecerdasan buatan, yaitu tingkat akurasi serta kecepatan dari pengenalan ekspresi wajah. Hal tersebut dapat mengakibatkan efek negatif pada bisnis, seperti ketidaktepatan dalam mengidentifikasi kekurangan pada pelayanan atau produk.

BAB V
SIMPULAN DAN SARAN

Bab ini menjelaskan mengenai simpulan akhir dari hasil penelitian yang telah dilakukan serta saran dari peneliti untuk penelitian yang selanjutnya.

5.1. Simpulan

Dari penelitian yang telah dilakukan, didapatkan beberapa kesimpulan sebagai berikut:

1. Implementasi model deteksi wajah dan model klasifikasi ekspresi untuk evaluasi kepuasan pelanggan, dilakukan mulai dari tahap model deteksi wajah. Set data WIDER diberikan tahap praproses agar sesuai dengan format COCO yang dapat diolah oleh RT-DETR. Selanjutnya, set data WIDER diproses oleh model RT-DETR (ResNet-18) dan RT-DETR (LCNet-0.25) yang telah dikonfigurasikan. Kemudian, dilanjutkan dengan tahap model klasifikasi ekspresi. Set data IMED diberikan tahap praproses agar sesuai dengan format FER-2013 dan kedua set data tersebut disesuaikan dengan format 7 ekspresi atau 3 ekspresi yang dapat diolah oleh Real Time-CNN. Selanjutnya, set data FER-2013 dan IMED diproses oleh model Real Time-CNN.
2. Implementasi keseluruhan model untuk evaluasi kepuasan pelanggan, diawali dengan memilih bobot terbaik dari model deteksi wajah, yaitu RT-DETR (ResNet-18) atau RT-DETR (LCNet-0.25). Selanjutnya, model klasifikasi ekspresi juga dipilih berdasarkan bobot terbaik melalui 2 format, yaitu Real Time-CNN (7 ekspresi) atau Real Time-CNN (3 ekspresi). Bobot deteksi wajah dipilih berdasarkan hasil evaluasi menggunakan set data WIDER dengan metrik *average precision* sesuai dengan ketentuan COCO serta ukuran model dan bobot klasifikasi ekspresi dipilih dengan metrik *accuracy* dan *macro-average* F1-score menggunakan set data FER-2013 dan IMED. Dengan model deteksi wajah dan model klasifikasi ekspresi yang telah dipilih berdasarkan bobot terbaik, selanjutnya kedua model tersebut digabungkan menjadi sebuah sistem yang dapat mengevaluasi kepuasan pelanggan. Selanjutnya keseluruhan model diuji menggunakan set data demonstrasi pengujian yang telah dianotasi.

3. Performa masing-masing model untuk evaluasi kepuasan pelanggan, didapatkan berdasarkan hasil evaluasi menggunakan set data yang digunakan untuk mengembangkan masing-masing model. Model deteksi wajah yang dievaluasi menggunakan set data WIDER mendapatkan evaluasi terbaik pada model RT-DETR (ResNet-18) pada *epoch* ke-71 dengan AP (IoU=0.5:0.95) sebesar 35%. Kemudian, model klasifikasi ekspresi yang dievaluasi menggunakan set data FER-2013 dan IMED mendapatkan evaluasi terbaik dengan menggunakan 3 jenis ekspresi pada *epoch* ke-74 dengan *micro average* F1-score sebesar 68.4%.
4. Performa keseluruhan model untuk evaluasi kepuasan pelanggan, didapatkan berdasarkan hasil evaluasi menggunakan set data video demonstrasi pengujian. Model pengenalan ekspresi wajah memiliki performa sebesar 5.94 FPS dan 77.5% AP untuk bounding box saja serta 4.7% AP untuk keseluruhan.

5.2. Saran

Dari penelitian yang telah dilakukan, didapatkan beberapa saran sebagai berikut:

1. Untuk mendapatkan hasil yang lebih baik lagi pada model RT-DETR dengan *backbone* ResNet-18 dapat dilakukan dengan menggunakan set data deteksi wajah dengan ukuran yang lebih besar daripada set data WIDER.
2. Buruknya performa model RT-DETR dengan *backbone* LCNet-0.25 disebabkan oleh arsitektur LCNet yang terlalu simpel, maka penelitian selanjutnya dapat menggunakan *backbone* yang memiliki kompleksitas yang lebih tinggi.
3. Model Real Time-CNN memiliki performa buruk karena 1) *treatment imbalance* yang masih belum tepat; 2) ketidaksesuaian format penentuan jenis ekspresi pada set data pengembangan dan pengujian; dan 3) posisi wajah yang tidak beragam pada set data pengembangan dan pengujian dengan karakteristik wajah orang Indonesia. Maka dari itu, untuk saran perbaikan yang pertama dapat mencari alternatif *treatment imbalance* lain yang tidak harus mengubah persentase pembagian set data. Contohnya *treatment*

tersebut seperti augmentasi data, yaitu SMOTE. Saran perbaikan yang kedua, karena set data pengembangan dan pengujian harus menggunakan format penentuan jenis ekspresi yang sama, maka untuk penelitian selanjutnya dapat menggunakan set data dengan format penentuan jenis ekspresi yang dapat dirujuk. Hal ini agar set data pengujian juga dapat dianotasi menggunakan format penentuan jenis ekspresi yang sama. Ketiga, dengan posisi wajah pada set data pengembangan dan pengujian yang tidak beragam, maka pada penelitian selanjutnya dapat menggunakan set data yang memiliki karakteristik wajah orang Indonesia dengan posisi wajah yang beragam.

4. Untuk meningkatkan keakuratan anotasi *bounding box* video demonstrasi pengujian, dapat menggunakan model deteksi wajah dengan performa yang lebih tinggi pada evaluasi set data WIDER.
5. Agar dapat menggunakan potensi model RT-DETR dengan seutuhnya (*end-to-end)*, dapat menjadikan dua tahap pada penelitian ini menjadi satu tahap dengan menggunakan set data deteksi wajah yang telah disertai dengan kelas jenis ekspresi.
6. Disebabkan oleh tahap pengujian yang menggunakan set data dengan karakter wajah orang Indonesia, maka sebaiknya set data untuk tahap pengembangan juga menggunakan set data dengan karakter wajah orang Indonesia.

DAFTAR PUSTAKA

(IBM), I. B. M. C. (n.d.). *What is artificial intelligence (AI)?* https://www.ibm.com/topics/artificial-intelligence

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics (Switzerland)*, *8*(3), 1–67. https://doi.org/10.3390/electronics8030292

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *ArXiv Preprint ArXiv:1710.07557*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12346 LNCS*, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

Ceccacci, S., Generosi, A., Giraldi, L., & Mengoni, M. (2023). Emotional valence from facial expression as an experience audit tool: An empirical study in the context of opera performance. *Sensors*, *23*(5). https://doi.org/10.3390/s23052688

Chang, W.-J., Schmelzer, M., Kopp, F., Hsum, C.-H., Su, J.-P., Chen, L.-B., & Chen, M.-C. (2023). *A deep learning facial expression recognition based scoring system for restaurants*. 3–6.

Chetana, P., Gunasekaran, M., & Tiwari, S. K. (2022). An efficient system for classifying customer facial expressions to obtain feedback with accuracy at unmanned restaurants using improved support vector machine and flattened convolutional neural networks. *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS*

2022, 1–8. https://doi.org/10.1109/MACS56771.2022.10022836

Cui, C., Gao, T., Wei, S., Du, Y., Guo, R., Dong, S., Lu, B., Zhou, Y., Lv, X., Liu, Q., Hu, X., Yu, D., & Ma, Y. (2021). *PP-LCNet: A lightweight CPU convolutional neural network*. 1–8. http://arxiv.org/abs/2109.15099

Dai, J. (2020). Real-time and accurate object detection on edge device with TensorFlow Lite. *Journal of Physics: Conference Series*, *1651*(1). https://doi.org/10.1088/1742-6596/1651/1/012114

Darwin, C. (1898). The expression of the emotions in man and animals. *The Portable Darwin*, 364–393. https://cir.nii.ac.jp/crid/1571417124682544000

Decaro, C., Montanari, G., Molinariz, R., Gilberti, A., Bagnoli, D., Bianconi, M., & Bellanca, G. (2019). Machine learning approach for prediction of hematic parameters in hemodialysis patients. *IEEE Journal of Translational Engineering in Health and Medicine, PP*, 1. https://doi.org/10.1109/JTEHM.2019.2938951

Eikvil, L. (1993). Optical character recognition. *Citeseer. Ist. Psu. Edu/142042. Html*, *26*.

Ekman, P. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, *8*(4), 151–158.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and cue to deception. In *Psychiatry: Journal for the Study of Interpersonal Processes* (Vol. 32, Issue 1, pp. 88–106).

Ekman, P., & Friesen, W. V. (1975). Unmasking the face: A guide to recognizing emotions from facial clues. In *Unmasking the face: A guide to recognizing emotions from facial clues.* Prentice-Hall.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system* (Issue v. 1). Consulting Psychologists Press. https://books.google.co.id/books?id=08l6wgEACAAJ

Ekman, P., & Heider, K. G. (1988). The universality of a contempt expression: A replication. *Motivation and Emotion*, *12*(3), 303–308. https://doi.org/10.1007/BF00993116

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial

intelligence and business value: A literature review. *Information Systems Frontiers*, *24*(5), 1709–1734.

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A. G., & Goldstein, T. (2023). Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems*, *36*(NeurIPS), 1–29.

González-Rodríguez, M. R., Díaz-Fernández, M. C., & Pacheco Gómez, C. (2020). Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions. *Telematics and Informatics*, *51*(October 2019), 101404. https://doi.org/10.1016/j.tele.2020.101404

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., & Lee, D.-H. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, 117–124.

Hamzah, A. A. bin, & Shamsudin, M. F. (2020). Why customer satisfaction is important to business? *Journal of Undergraduate Social & Technology*, *2*(1), 2710–6918. www.jusst.abrn.asia

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Karadağ, B. (2018). Measurement of customer satisfaction through emotion analysis in the banking sector. *EasyChair Preprints*, *April*. https://doi.org/10.29007/x72l

Kelleher, J. D. (2019). *Deep learning*. MIT Press. https://books.google.co.id/books?hl=en&lr=&id=b06qDwAAQBAJ&oi=fnd&pg=PP9&dq=what+is+deep+learning&ots=_pASRJpWVO&sig=p9oL59JbaWSnwP_QR4c0Rfjmgno&redir_esc=y#v=onepage&q=what is deep learning&f=false

Lecun. (1999). *MNIST dataset (Modified National Institute of Standards and Technology)*. http://yann.lecun.com/exdb/mnist/index.html

Li, T., Wang, J., & Zhang, T. (2022). L-DETR: A light-weight detector for end-to-end object detection with transformers. *IEEE Access*, *10*(October), 105685–105692. https://doi.org/10.1109/ACCESS.2022.3208889

Liliana, D. Y., Basaruddin, T., & Oriza, I. I. D. (2018). The Indonesian Mixed Emotion Dataset (IMED): A facial expression dataset for mixed emotion recognition. *ACM International Conference Proceeding Series*, 56–60. https://doi.org/10.1145/3293663.3293671

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Y., Wang, F., Deng, J., Zhou, Z., Sun, B., & Li, H. (2022). MogFace: Towards a deeper appreciation on face detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 4083–4092. https://doi.org/10.1109/CVPR52688.2022.00406

Lv, W., Zhao, Y., Xu, S., Wei, J., Wang, G., Cui, C., Du, Y., Dang, Q., & Liu, Y. (2023). *DETRs beat YOLOs on real-time object detection*. http://arxiv.org/abs/2304.08069

Meena, G., Mohbey, K. K., & Kumar, S. (2023). Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach. *International Journal of Information Management Data Insights*, *3*(1), 100174. https://doi.org/10.1016/j.jjimei.2023.100174

Mesuga, R., & Bayanay, B. J. (2021). *A deep transfer learning approach on identifying glitch wave-form in gravitational wave data*. *May*. https://doi.org/10.48550/arXiv.2107.01863

Meyers, R. A. (2001). *Encyclopedia of physical science and technology*. https://www.sciencedirect.com/referencework/9780122274107/encyclopedia-of-physical-science-and-technology#book-info

Naqa, I. El, & Murphy, M. J. (2015). Machine learning in radiation oncology. *Machine Learning in Radiation Oncology*, 3–11. https://doi.org/10.1007/978-3-319-18305-3

Ouyang, H. (2023). *DEYOv3: DETR with YOLO for real-time object detection*. http://arxiv.org/abs/2309.11851

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1994). Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria. *Journal of Retailing*, *70*(3), 201–230. https://doi.org/10.1016/0022-4359(94)90033-7

Plastiras, G., Terzi, M., Kyrkou, C., & Theocharidcs, T. (2018). Edge intelligence: Challenges and opportunities of near-sensor machine learning applications. *Proceedings of the International Conference on Application-Specific Systems, Architectures and Processors*, *2018-July*. https://doi.org/10.1109/ASAP.2018.8445118

Pratomo, A. H., Kaswidjanti, W., & Mu'arifah, S. (2020). Implementasi algoritma region of interest (ROI) untuk meningkatkan performa algoritma deteksi dan klasifikasi kendaraan. *J. Teknol. Inf. Dan Ilmu Komput*, *7*(1), 155–162.

Qi, D., Tan, W., Yao, Q., & Liu, J. (2023). YOLO5Face: Why reinventing a face detector. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13805 LNCS*, 228–244. https://doi.org/10.1007/978-3-031-25072-9_15

Ramla, M., Sangeetha, S., & Nickolas, S. (2022). Chapter 6 - towards building an efficient deep neural network based on YOLO detector for fetal head localization from ultrasound images. In R. Sridhar, G. R. Gangadharan, M. Sheng, & R. B. T.-E.-T. in P. H. S. S. Shankaran (Eds.), *Cognitive Data Science in Sustainable Computing* (pp. 137–156). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-90585-5.00005-9

Razak, A. A., & Shamsudin, M. F. (2019). The influence of atmospheric experience on theme park tourist's satisfaction and loyalty in Malaysia. *International Journal of Innovation, Creativity and Change*, *6*(9), 10–20.

Revina, I. M., & Emmanuel, W. R. S. (2021). A survey on human face expression

recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, *33*(6), 619–628.

Ringler, C. (2021). Truth and lies: The impact of modality on customer feedback. *Journal of Business Research*, *133*(May), 376–387. https://doi.org/10.1016/j.jbusres.2021.05.014

Robinson, D. L. (2008). Brain function, emotional experience and personality. *Netherlands Journal of Psychology*, *64*(4), 152–168. https://doi.org/10.1007/bf03076418

Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., & Rodrigues, J. J. P. C. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, *68*, 817–840. https://doi.org/10.1016/j.aej.2023.01.017

Slim, M., Kachouri, R., & Atitallah, A. Ben. (2018). Customer satisfaction measuring based on the most significant facial emotion. *2018 15th International Multi-Conference on Systems, Signals and Devices, SSD 2018*, 502–507. https://doi.org/10.1109/SSD.2018.8570588

Sugianto, N., Tjondronegoro, D., & Tydd, B. (2018). Deep Residual Learning for Analyzing Customer Satisfaction using Video Surveillance. *Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 1–6. https://doi.org/10.1109/AVSS.2018.8639478

Szeliski, R. (2010). *Computer vision: Algorithms and applications*. Springer London. https://books.google.co.id/books?id=bXzAlkODwa8C

TensorFlow. (2019). *TensorFlow Lite*. Tensorflow.Org.

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. (2023). *Loss functions and metrics in deep learning*. http://arxiv.org/abs/2307.02694

Thompson, N., Greenewald, K., Lee, K., & Manso, G. F. (2023). *The computational limits of deep learning*. https://doi.org/10.21428/bf6fb269.1f033948

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR, abs/1706.0*. http://arxiv.org/abs/1706.03762

Vemou, K., Horvath, A., & Zerdick, T. (2021). Facial emotion recognition. *Lecture Notes in Mechanical Engineering*, *1*, 751–761. https://doi.org/10.1007/978-981-15-9956-9_73

Venkatesan, R., & Li, B. (2018). *Convolutional neural networks in visual computing: A concise guide*. CRC Press. https://books.google.co.id/books?id=aOtPAQAACAAJ

Vinh, T. Q., & Tran Dac Thinh, P. (2019). Advertisement system based on facial expression recognition and convolutional neural network. *Proceedings - 2019 19th International Symposium on Communications and Information Technologies, ISCIT 2019*, 476–480. https://doi.org/10.1109/ISCIT.2019.8905134

Wen, Z., Lin, W., Wang, T., & Xu, G. (2023). Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, *8*(2). https://doi.org/10.3390/biomimetics8020199

Wu, W., Peng, H., & Yu, S. (2023). YuNet: A tiny millisecond-level face detector. *Machine Intelligence Research*, *20*. https://doi.org/10.1007/s11633-023-1423-y

Yang, M. H., Kriegman, D. J., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(1), 34–58. https://doi.org/10.1109/34.982883

Yang, S., Luo, P., Loy, C.-C., & Tang, X. (2016). WIDER face: A face detection benchmark. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5533.

Zhu, Y., Cai, H., Zhang, S., Wang, C., & Xiong, Y. (2020). *TinaFace: Strong but simple baseline for face detection*. http://arxiv.org/abs/2011.13183