# **PROPOSAL SKRIPSI**

**PERBANDINGAN METODE DETEKSI OBJEK DALAM CITRA PENCAHAYAAN RENDAH MENGGUNAKAN METODE CNN BASED DAN TRANSFORMER BASED**

****

**RIDHO PANDHU AFRIANTO**

**NIM. 162112133062**

**PROGRAM SARJANA**

**TEKNOLOGI SAINS DATA**

**DEPARTEMEN TEKNIK**

**FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN**

**UNIVERSITAS AIRLANGGA**

**2024**

# **LEMBAR PENGESAHAN**

**PERBANDINGAN METODE DETEKSI OBJEK DALAM CITRA PENCAHAYAAN RENDAH MENGGUNAKAN CNN BASED DAN TRANSFORMER BASED**

|  |  |
| --- | --- |
| Nama | : Ridho Pandhu Afrianto |
| NIM | : 162112133062 |
| Tanggal Seminar Proposal | : Kamis, 3 Oktober 2024 |

Surabaya, 24 Oktober 2024

|  |  |
| --- | --- |
| Pembimbing I  **Dr. Aziz Fajar, S.Kom., M.Kom.** | Pembimbing II  **Ratih Ardiati Ningrum, S.Si., M.Stat.** |
| **NIP.** 199410052024023101 | **NIP.** 199501262020013201 |

Mengetahui,

Koordinator Program Studi

S1 Teknologi Sains Data

**Dr. Dwi Rantini, S.Si.**

**NIP.** 199406152022033201

# **KATA PENGANTAR**

Puji syukur saya panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga saya dapat menyelesaikan proposal skripsi ini yang berjudul "Perbandingan Metode Deteksi Objek dalam Citra Pencahayaan Rendah Menggunakan CNN-*Based* dan Transformer-*Based*". Proposal ini disusun sebagai salah satu syarat untuk menempuh seminar proposal skripsi di Program Studi Teknologi Sains Data, Fakultas Sains dan Teknologi, Universitas Airlangga.

Saya menyadari bahwa dalam penyusunan proposal ini, banyak pihak yang telah memberikan bantuan dan dukungan, baik berupa bimbingan, saran, maupun dorongan moral. Oleh karena itu, saya mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Aziz Fajar dan Ibu Ratih Ardiati Ningrum, selaku dosen pembimbing, atas arahan, bimbingan, serta kesabarannya dalam membimbing saya selama penyusunan proposal ini.
2. Keluarga tercinta, yang selalu memberikan doa dan dukungan tanpa henti.
3. Rekan-rekan mahasiswa dan semua pihak yang tidak dapat saya sebutkan satu per satu, atas dukungan dan bantuan yang telah diberikan.

Saya berharap proposal ini dapat memberikan kontribusi positif, khususnya dalam pengembangan ilmu pengetahuan di bidang teknologi sains data, serta bermanfaat bagi para pembaca. Akhir kata, saya menyadari bahwa proposal ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat saya harapkan demi perbaikan dan penyempurnaan di masa mendatang.

Surabaya, 25 September 2024

Ridho Pandhu Afrianto

# **DAFTAR ISI**

[**HALAMAN JUDUL** i](#_Toc181134033)

[**LEMBAR PENGESAHAN** ii](#_Toc181134034)

[**KATA PENGANTAR** iii](#_Toc181134035)

[**DAFTAR ISI** iv](#_Toc181134036)

[**DAFTAR TABEL** v](#_Toc181134037)

[**DAFTAR GAMBAR** vi](#_Toc181134038)

[**BAB 1 PENDAHULUAN** 1](#_Toc181134039)

[1.1 Latar Belakang 1](#_Toc181134040)

[1.2 Rumusan Masalah 5](#_Toc181134041)

[1.3 Hipotesis 5](#_Toc181134042)

[1.4 Tujuan Penelitian 6](#_Toc181134043)

[1.5 Manfaat Penelitian 6](#_Toc181134044)

[1.6 Batasan Masalah 6](#_Toc181134045)

[**BAB 2 TINJAUAN PUSTAKA** 7](#_Toc181134046)

[2.1 *Computer Vision* 7](#_Toc181134047)

[2.2 Analisis *Object Detection* 8](#_Toc181134048)

[2.3 Data *Preprocessing* 8](#_Toc181134049)

[2.3.1 *Contrast Enhancement* 9](#_Toc181134050)

[2.3.2 Normalisasi Data 10](#_Toc181134051)

[2.3.3 Augmentasi Data 10](#_Toc181134052)

[2.4 *Convolutional Neural Network* 11](#_Toc181134053)

[2.5 *You Only Look Once* 13](#_Toc181134054)

[*2.*6 Transformer 15](#_Toc181134055)

[2.7 CNN-Transformer (YOLOv10) 18](#_Toc181134056)

[2.8 *Hyperparameter Tuning* 19](#_Toc181134057)

[2.9 *Loss Function* 21](#_Toc181134058)

[2.10 Metrik Evaluasi Model 24](#_Toc181134059)

[2.11 Pareto Frontier 26](#_Toc181134060)

[**BAB 3 METODE PENELITIAN** 28](#_Toc181134061)

[3.1 Lokasi dan Waktu 28](#_Toc181134062)

[3.2 Bahan dan Alat 28](#_Toc181134063)

[3.3 Cara Kerja 29](#_Toc181134064)

[3.3.1 *Data Preprocessing* 31](#_Toc181134065)

[3.3.2 Implementasi Model 34](#_Toc181134066)

[3.3.3 Metode Perbandingan Model 40](#_Toc181134067)

[**DAFTAR PUSTAKA** 42](#_Toc181134068)

# **DAFTAR TABEL**

[Tabel 3.1 Deskripsi Parameter Augmentasi 33](#_Toc180616121)

[Tabel 3.2 Tipe Model YOLOv9 35](#_Toc180616122)

[Tabel 3.3 Tipe Model RT-DETR 37](#_Toc180616123)

[Tabel 3.4 Tipe Model YOLOv10 39](#_Toc180616124)

[Tabel 3.5 Deskripsi *Hyperparameter* yang Diuji 39](#_Toc180616125)

# **DAFTAR GAMBAR**

Gambar 1. 1 1

(a) Citra berpencahayaan rendah 1

(b) Contoh hasil dari model deteksi objek 1

[Gambar *2.* 1 *Object detection* dengan satu atau lebih objek 8](#_Toc180613991)

[Gambar 2. 2 Proses Konvolusi 2D 12](#_Toc180613992)

[Gambar 2. 3 Sistem deteksi objek dengan YOLO 13](#_Toc180613993)

[Gambar 2. 4 Arsitektur YOLO 13](#_Toc180613994)

[Gambar 2. 5 Arsitektur YOLOv9 14](#_Toc180613995)

[Gambar 2. 6 Arsitektur Transformer 16](#_Toc180613996)

[Gambar 2. 7 Arsitektur RT-DETR 17](#_Toc180613997)

[Gambar 2. 8 Arsitektur YOLOv10 19](#_Toc180613998)

[Gambar 2. 9 Perbedaan *ℓ2 loss dan IoU loss* 22](#_Toc180613999)

[Gambar 3.1 Sampel *Exclusively Dark* dataset.......................................................29](#_Toc180614704)

[Gambar 3.2 *Flowchart* Keseluruhan Proses 30](#_Toc180614705)

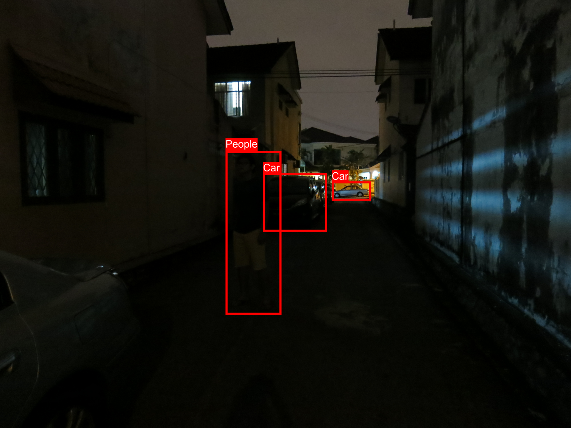
[Gambar 3. 3 Proses Inferensi Model berbasis CNN 34](#_Toc180614706)

[Gambar 3. 4 Proses Inferensi Model Berbasis Transformer (RE-DETR) 36](#_Toc180614707)

[Gambar 3. 5 Proses Inferensi Model berbasis CNN-Transformer (YOLOv10) 38](#_Toc180614708)

# **BAB 1 PENDAHULUAN**

## 1.1 Latar Belakang

 Dalam dunia keamanan modern, penggunaan CCTV telah menjadi pilihan utama dalam aktivitas di berbagai ruang publik dan privat, mulai dari pemantauan lalu lintas hingga perlindungan objek vital seperti bandara dan gedung pemerintahan. Namun, meskipun teknologi CCTV telah berkembang, masalah mendasar terkait deteksi objek dalam kondisi pencahayaan rendah tetap menjadi kendala signifikan. Pada malam hari atau di area yang minim penerangan, CCTV konvensional sering kali gagal menangkap gerakan mencurigakan dengan jelas, sehingga memberikan celah bagi tindakan kejahatan, termasuk terorisme. Kasus-kasus seperti pemboman dan serangan di ruang publik yang terekam CCTV menunjukkan bahwa saat ini identifikasi objek yang ada di dalamnya masih mengalami situasi yang sulit dan lambat​ (Putri, 2021). Keterbatasan ini menyoroti perlunya pengembangan lebih lanjut dalam teknologi pengawasan yang mampu beradaptasi dengan berbagai kondisi cahaya untuk memastikan keamanan maksimal.

(b)

(a)

(b)



(a)

(a)

(a)

Gambar 1.1 (a) Citra berpencahayaan rendah, (b) Contoh hasil dari model deteksi objek

Sumber: Loh & Chan (2018)

Salah satu permasalahan dari perangkat yang memerlukan waktu kerja 24 jam non-stop seperti CCTV, adalah kemampuannya untuk mendeteksi sebuah objek jika objek tersebut berada di pencahayaan yang rendah atau malam hari. Berdasarkan Gambar 1.1 (a) sudah dapat disimpulkan bahwa belum tentu manusia dapat mengetahui objek apa saja yang berada di gambar tersebut apalagi mengawasinya selama 24 jam. Oleh karena itu, perlu bantuan model deteksi objek seperi yang dicontohkan dalam Gambar 1.1 (b) untuk melakukan tugas demikian. Menurut Chen & Shah (2021), faktor-faktor seperti *noise, glare, lousy illumination, low contrast, shadows,* dan *reflectance* membuat deteksi objek dalam kondisi tersebut menjadi sulit. Berbagai penelitian dengan pendekatan yang beragam telah dilakukan untuk mengatasi masalah tersebut. Sebagian besar penelitian dalam peningkatan citra berfokus pada mempercantik subjek dan meningkatkan kualitas estetika gambar, dibandingkan dengan pemulihan fitur yang diperlukan untuk analisis lebih lanjut. Beberapa penelitian juga mencoba mengatasi masalah tersebut dengan menggunakan *hardware* yang lebih canggih, seperti kamera yang dilengkapi dengan sensor inframerah dan gambar termal. Namun, metode ini memiliki biaya yang tinggi dan sering kali menghasilkan citra yang kurang realistis (Chen & Shah, 2021). Salah satu pendekatan untuk meningkatkan kecerahan dalam kondisi pencahayaan rendah adalah dengan menaikkan ISO atau memperpanjang *exposure time*. Namun, strategi ini masing-masing cenderung meningkatkan *noise* dan menyebabkan *motion blur* (Chen *et al.,* 2018). Pendekatan lain yang dapat diterapkan adalah memanfaatkan *software* seperti Photoshop atau Lightroom untuk melakukan penyesuaian pencahayaan secara digital. Meskipun demikian, *software* ini memerlukan keterampilan artistik dan tidak efisien untuk dataset skala besar dengan kondisi pencahayaan yang beragam (Zheng & Gupta, 2022). Selebihnya, pendekatan-pendekatan tersebut menjadi kurang layak untuk aplikasi seperti CCTV yang memerlukan operasi non-stop selama 24 jam dan menuntut kecepatan maupun ketepatan deteksi.

Menyikapi masalah tersebut, Loh & Chan (2018) membuat sebuah dataset bernama ExDark atau *Exclusively Dark* untuk memfasilitasi para peneliti dalam memahami fenomena citra di pencahayaan rendah. ExDark memiliki 7363 citra yang terdiri dari citra berpencahayaan rendah hingga citra di berpencahayaan fajar dengan 12 kelas objek di dalamnya sehingga cocok untuk riset berbasis aplikasi seperti *object detection* (Loh & Chan, 2018). Dataset ExDark ini kemudian diadopsi oleh model PE-YOLO atau *Pyramid Enhancement - You Only Look Once* yang secara khusus dirancang untuk mendeteksi objek dalam pencahayaan rendah dengan menggunakan pendekatan *Convolutional Neural Networks* (CNN) (Yin *et al.,* 2023). Kelemahan model ini adalah pada model *backbone* yang terbilang lama (YOLOv3), serta mekanisme di dalamnya yang masih memiliki keterbatasan dalam penanganan *noise* terutama pada frekuensi tinggi.  Pada penelitian lain, Cui dkk. (2022) mengembangkan model *Multitask Auto-Encoding Transformation* (MAET).  MAET menawarkan pendekatan baru dalam mendeteksi objek di kondisi pencahayaan rendah dengan memanfaatkan regularisasi ortogonal untuk memisahkan fitur degradasi pencahayaan dan deteksi objek. Model ini menggunakan pipeline *image signal processor* (ISP) yang lebih realistis untuk mensintesis gambar berpencahayaan rendah, yang memungkinkan model untuk mengoptimalkan deteksi objek di lingkungan gelap tanpa menghasilkan artefak yang sering terjadi pada metode pemulihan pencahayaan tradisional. Disamping skor mAP yang bernilai 0,74 pada dataset ExDark, terdapat beberapa kelemahan yang membatasi model ini  (Cui *et al.,* 2022). Penggunaan regularisasi ortogonal menambah kompleksitas komputasi yang dapat menghambat implementasi model ini pada perangkat dengan sumber daya terbatas. Selain itu, MAET lagi-lagi juga masih mengandalkan YOLOv3 sebagai *backbone* yang merupakan arsitektur lama. Menggunakan pendekatan yang berbeda,  Cui dkk. (2022) melakukan penelitian lain dengan mekanisme Transformer untuk mengembangkan *Illumination Adaptive Transformer* (IAT) untuk memulihkan gambar sRGB dengan pencahayaan normal dari kondisi *low-light* atau *under/overexposure*. IAT mampu mencapai skor mAP sebesar 77,2 pada dataset ExDark dengan *time inference* 0,04 detik berkat komponen *simplified image signal processor* (ISP) yang dimilikinya (Cui *et al.,* 2022). Komponen *Simplified* ISP inilah yang membantu IAT mampu merealisasikan inferensi yang sangat cepat dengan parameter yang relatif kecil yaitu sebesar 90.000 parameter. Namun, justru dari *simplified* ISP inilah IAT mengalami keterbatasan generalisasi ketika diterapkan pada dataset dengan kondisi yang beragam.

Kedua pendekatan yang populer pada *object detection*, yaitu CNN dan Transformer telah terbukti bahwa selalu ada *trade-off* masing-masing. Oleh karena itu, akan sangat menarik untuk mengeksplorasi apakah penggunaan kedua pendekatan yang berbeda pada model-model *state of the art* terbaru berpotensi menghasilkan peningkatan performa yang signifikan dan *trade-off*  yang lebih optimal dibandingkan dengan penelitian-penelitian terdahulu. Definisi optimal dalam konteks ini adalah  model yang mampu memberikan skor *mean Average Precision* (mAP) yang tinggi sembari meminimalkan waktu inferensi. Model-model *state-of-the-art* saat ini yang menggunakan pendekatan CNN, Transformer, dan Hybrid (CNN-Transformer) adalah YOLOv9 yang dirancang oleh Wang dkk. (2024) , RT-DETR  oleh Zhao dkk. (2023), dan YOLOv10 oleh peneliti lain Wang dkk. (2024). Pemilihan model-model ini didasarkan tidak hanya pada statusnya sebagai model *state-of-the-art* saat ini, tetapi juga karena kemampuan mereka dalam melakukan deteksi multi-objek secara *real-time* dengan kecepatan yang tinggi, menjadikannya ideal untuk aplikasi yang membutuhkan deteksi cepat dan akurat.

Dalam membandingkan dua jenis metrik yang berbeda, pendekatan dengan perbandingan secara langsung sangat tidak memungkinkan untuk dilakukan karena mAP diukur dalam skala skor sementara waktu inferensi diukur dalam satuan waktu (*ms*). Oleh karena itu, pendekatan yang tepat untuk mengevaluasi kedua metrik ini adalah dengan menggunakan *Pareto Frontier*. *Pareto Frontier* memungkinkan evaluasi *trade-off* antara dua atau lebih metrik, di mana sebuah solusi dianggap optimal jika tidak ada solusi lain yang lebih baik dalam kedua metrik secara bersamaan. Perbandingan antara mAP dan waktu inferensi menjadi sangat penting, karena perangkat seperti CCTV memerlukan model yang tidak hanya cepat tetapi juga akurat.

Penelitian ini ditujukan untuk memberikan kontribusi yang signifikan dalam pengembangan metode deteksi objek pada citra berpencahayaan rendah dengan mengeksplorasi dan membandingkan model berbasis CNN dan pendekatan Transformer. Dengan mengatasi tantangan yang dihadapi oleh model deteksi objek dalam kondisi pencahayaan rendah, penelitian ini dapat membantu mengarahkan untuk meningkatkan akurasi serta kecepatan deteksi dalam berbagai aplikasi dunia nyata, seperti pengawasan CCTV dan industri lainnya yang bergantung pada deteksi objek secara *real-time*. Tujuan dari penelitian ini adalah dapat membantu mengidentifikasi solusi yang lebih efisien dan efektif baik secara biaya maupun sumber daya, terutama dalam kondisi pencahayaan yang beragam. Selebihnya, optimalisasi pada penelitian ini juga diharapkan dapat menjadi solusi alternatif pengganti *thermal-cam* dan *infrared* untuk CCTV konvensional yang ada di Indonesia.

## 1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana performa metode YOLOv9 terhadap objek dengan pencahayaan rendah dan apa faktor yang menyebabkannya?
2. Bagaimana performa metode RT-DETR terhadap objek dengan pencahayaan rendah dan apa faktor yang menyebabkannya?
3. Bagaimana performa metode YOLOv10 terhadap objek dengan pencahayaan rendah dan apa faktor yang menyebabkannya?
4. Apa metode yang terbaik dari segi *mean Average Precision* (mAP) dan inferensi waktuuntuk mendeteksi objek berpencahayaan rendah?

## 1.3 Hipotesis

Metode penggabungan CNN & Transformer akan menjadi yang paling unggul dibandingkan dengan metode CNN dan Transformer secara terpisah. Hal ini didasarkan pada kemampuan CNN-Transformer untuk mengintegrasikan keunggulan dari kedua metode, yaitu kemampuan CNN dalam mendeteksi objek dengan cepat dan kemampuan Transformer dalam mengenali fitur-fitur secara semantik.

## 1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Untuk mengetahui performa metode YOLOv9 terhadap objek dengan pencahayaan rendah dan mencari faktor yang menyebabkannya.
2. Untuk mengetahui performa metode RT- DETR terhadap objek dengan pencahayaan rendah dan mencari faktor yang menyebabkannya.
3. Untuk mengetahui performa metode YOLOv10 terhadap objek dengan pencahayaan rendah dan mencari faktor yang menyebabkannya.
4. Untuk mencari metode yang terbaik dalam *mean Average Precision* (mAP) dan inferensi waktudalam mendeteksi objek dengan pencahayaan rendah.

## 1.5 Manfaat Penelitian

Manfaat yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menjadi referensi bagi penelitian selanjutnya dalam pengembangan model berbasis CNN ataupun Transformer untuk mendeteksi objek.
2. Model yang lebih optimal dalam mendeteksi objek dalam pencahayaan rendah akan bermanfaat untuk meningkatkan kapabilitas *hardware* dengan sumber daya terbatas dalam mendeteksi objek.

## 1.6 Batasan Masalah

Batasan masalah pada penelitian ini jumlah kelas yang dapat dideteksi hanya 12 kelas karena sesuai dengan data yang didapat dari ExDark.

# **BAB 2 TINJAUAN PUSTAKA**

## 2.1 *Computer Vision*

*Computer vision* adalah kemampuan komputer untuk memahami dan menganalisis data visual, terutama dari gambar atau video, dengan tujuan mengidentifikasi objek, pola, dan fitur lainnya. Ini adalah cabang ilmu komputer yang berkaitan dengan pengembangan sistem yang memungkinkan komputer untuk “melihat” dan memahami dunia sekitarnya melalui data visual yang diberikan kepadanya (Wibowo, 2016).

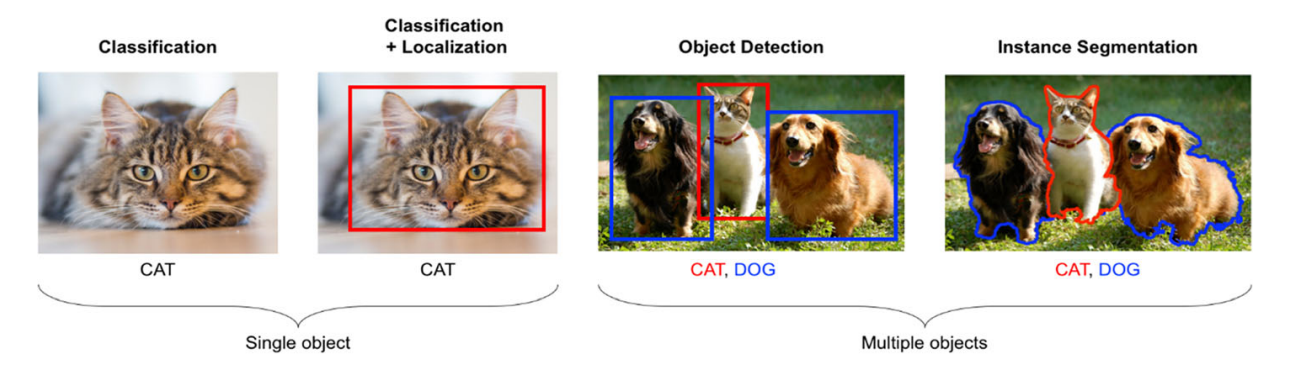
*Computer Vision* telah menjadi topik yang luas, meliputi mulai dari merekam data mentah hingga ekstraksi pola gambar dan interpretasi informasi. Ini menggabungkan konsep, teknik, dan ide dari pemrosesan gambar digital, pengenalan pola, kecerdasan buatan, dan grafika komputer. Sebagian besar tugas dalam *computer vision* berkaitan dengan proses mendapatkan informasi tentang peristiwa atau deskripsi dari gambar digital dan ekstraksi fitur. Tujuan utama *computer vision* adalah menciptakan model dan mengekstrak data serta informasi dari gambar, sementara pemrosesan gambar berkaitan dengan menerapkan transformasi komputasi pada gambar*. Computer vision* juga memiliki keterkaitan dengan *Human-Computer Interaction* (HCI), yang fokus pada desain, *interface*, dan interaksi antara manusia dan komputer (Wiley & Lucas, 2018).

Meskipun fungsinya serupa, sistem *computer vision* memiliki keterbatasan dibandingkan dengan mata manusia, terutama dalam hal sensitivitas parameter, kekuatan algoritma, dan akurasi hasil. Beberapa tantangan signifikan dalam teknik mereka termasuk pengukuran kinerja dan evaluasi algoritma untuk mencapai akurasi, kekuatan, atau skalabilitas. Ada upaya yang komprehensif untuk mengembangkan dan mengategorikan *computer vision* ke dalam berbagai aplikasi spesifik, seperti otomatisasi garis perakitan, sensor jarak jauh, robotika, komunikasi komputer dan manusia, serta alat bagi kaum difabel visual (Wiley & Lucas, 2018).

## 2.2 Analisis *Object Detection*

Deteksi objek merupakan proses penting dalam *computer vision* yang bertujuan untuk mengidentifikasi instansi dari objek-objek semantik tertentu dalam gambar digital dan video. Salah satu pendekatan umum dalam deteksi objek adalah dengan menciptakan sejumlah besar *candidate windows* yang kemudian dianalisis menggunakan fitur-fitur *Convolutional Neural Networks* (Voulodimos, 2018).

Pendeteksian objek digunakan untuk mengklasifikasikan dan menemukan lokasi setiap objek dalam kotak pembatas atau *bounding box*. Penelitian oleh Lee (2017) menghasilkan sebuah metode bernama *sliding window* yang mampu mengurangi kompleksitas waktu yang sudah menjadi teknik dasar untuk mendeteksi objek. Dalam metode ini, sebuah *window* atau jendela dengan ukuran misalnya M x N, dipilih untuk mencari di atas gambar objektif. Pada awalnya, sebuah pengklasifikasi disiapkan pada serangkaian gambar *train*, menyebar di atas objek yang dituju untuk dideteksi sebagai satu kelas dan objek tidak teratur sebagai kelas yang berbeda. Tes yang termasuk dalam objek yang dituju untuk dideteksi disebut gambar positif, sementara sampel tidak teratur disebut gambar negatif (Lee *et al.*, 2017).



Gambar 2. 1 Object detection dengan satu atau lebih objek

Sumber: Diwan, *et al.* (2022)

## 2.3 Data *Preprocessing*

Pada domain *computer vision, data preprocessing* adalah langkah penting yang bertujuan untuk meningkatkan akurasi dan efisiensi model. Dalam banyak studi terbaru, seperti yang dijelaskan oleh Rouhbakhshmeghrazi & Alizadeh (2023), teknik ini mencakup serangkaian metode seperti augmentasi gambar, peningkatan kontras, dan pengurangan *noise*, yang diterapkan untuk mempersiapkan gambar sebelum proses pelatihan model. Sesuai dasar tersebut, penelitian ini akan melakukan *data preprocessing* berupa normalisasi data dan augmentasi data.

### 2.3.1 *Contrast Enhancement*

*Contrast Enhancement* merupakan teknik dalam pemrosesan citra yang bertujuan untuk meningkatkan kualitas visual dengan memperbesar perbedaan intensitas antara objek dan latar belakangnya. Salah satu metode yang paling sering digunakan untuk meningkatkan kontras suatu citra adalah CLAHE.

*Contrast Limited Adaptive Histogram Equalization* (CLAHE) adalah teknik peningkatan gambar yang bertujuan untuk memperbaiki kontras pada area gambar dengan menggunakan histogram *equalization,* tetapi dengan batasan untuk menghindari amplifikasi *noise* secara berlebihan. Berbeda dengan *histogram* *equalization* biasa, CLAHE membagi gambar menjadi blok-blok kecil dan melakukan proses *equalization* secara lokal, yang membantu dalam mengatasi masalah *over-amplification* pada area terang atau gelap. Seperti yang dijelaskan oleh Yusuf dkk. (2024), CLAHE telah digunakan dalam berbagai aplikasi untuk meningkatkan visibilitas objek pada gambar yang memiliki variasi pencahayaan, seperti pada deteksi target menggunakan citra UAV. Teknik ini membantu meningkatkan detail gambar sambil mengurangi *noise* yang tidak diinginkan.

Selain itu, Manongga dkk. (2024) juga menunjukkan bahwa CLAHE digunakan dalam deteksi tanda jalan di kondisi pencahayaan rendah, dimana kombinasi CLAHE dan YOLOv7 membantu meningkatkan deteksi secara signifikan. CLAHE memastikan bahwa objek pada gambar yang memiliki kontras rendah tetap terdeteksi dengan jelas tanpa menghasilkan distorsi berlebih. Pendekatan ini sangat penting dalam aplikasi *real-time* yang memerlukan pengolahan gambar yang cepat dan akurat.

### 2.3.2 Normalisasi Data

Normalisasi data adalah langkah dalam *pre-processing* data yang digunakan untuk meningkatkan kinerja model *machine learning*. Proses ini melibatkan transformasi data sehingga setiap fitur memiliki skala yang seragam, yang membantu model untuk berfungsi dengan lebih baik dan menghasilkan hasil yang lebih akurat. Tanpa normalisasi, fitur-fitur dengan skala besar dapat mendominasi algoritma dan mengakibatkan model tidak mampu mempelajari data dengan efektif. Teknik normalisasi yang umum digunakan meliputi *min-max scaler*, *z-score normalization*, dan *normalization* berdasarkan distribusi. *Min-max scaler* mengubah data ke dalam rentang [0, 1] atau [-1, 1], sementara *z-score normalization* mengubah data sehingga memiliki *mean* 0 dan standar deviasi 1. Dalam penelitian ini, teknik normalisasi data yang digunakan adalah *min-max scaler* dengan persamaan yang dijelaskan sebagai berikut.

*=*  (2.1)

Penelitian terbaru menunjukkan bahwa normalisasi data sangat penting dalam berbagai aplikasi *machine learning*. Misalnya, dalam studi yang dilakukan oleh Peng dkk. (2023), normalisasi data digunakan untuk meningkatkan prediksi keletihan material dalam lingkungan ekstrem, menunjukkan bahwa data yang dinormalisasi mempermudah model untuk memprediksi dengan lebih akurat. Demikian juga, penelitian oleh Cao dkk. (2023) mengimplementasikan normalisasi data dalam pemodelan kognitif *fuzzy*, yang menunjukkan peningkatan interpretabilitas dan akurasi model. Studi ini menyoroti bahwa normalisasi tidak hanya meningkatkan kinerja model tetapi juga membantu dalam menangani data heterogen yang mungkin memiliki distribusi yang berbeda.

### 2.3.3 Augmentasi Data

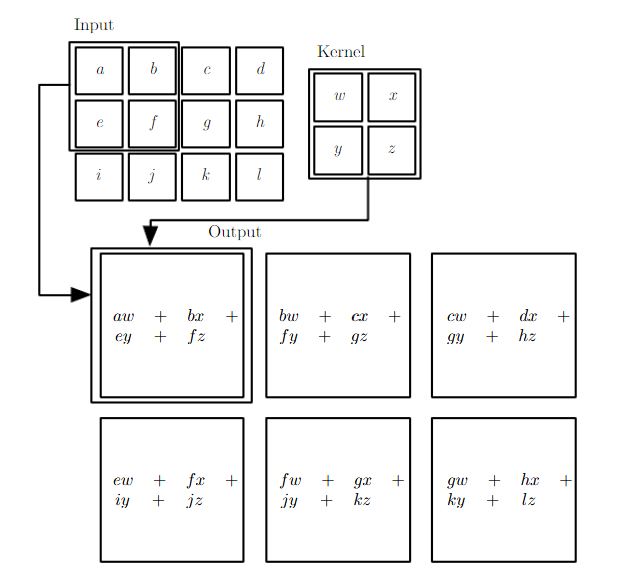
Augmentasi data adalah teknik yang digunakan dalam bidang *computer vision* untuk meningkatkan variasi data latih tanpa harus mengumpulkan data baru. Teknik ini menjadi sangat penting dalam pelatihan model *machine learning*, khususnya *deep learning*, yang membutuhkan jumlah data yang besar untuk mencapai kinerja optimal. Augmentasi data membantu mengatasi masalah *overfitting* dengan menyediakan variasi yang lebih besar dalam data latih, sehingga model dapat lebih general dan mampu menangani data baru dengan lebih baik.

Beberapa teknik umum yang digunakan dalam augmentasi data meliputi rotasi, skala, translasi, *flipping*, pemotongan, perubahan kecerahan dan kontras, serta penambahan *noise*. Teknik ini memungkinkan model untuk belajar dari berbagai variasi gambar yang mungkin ditemui di dunia nyata. Studi terbaru oleh Shorten dan Khoshgoftaar (2019) menunjukkan bahwa augmentasi data dapat secara signifikan meningkatkan kinerja model *deep learning* dalam berbagai tugas *computer vision*. Selain itu, penelitian ini mengungkapkan bahwa kombinasi berbagai teknik augmentasi seringkali memberikan hasil terbaik.

## 2.4 *Convolutional Neural Network*

Jaringan saraf konvolusional atau *Convolutional Neural Network (CNN)* merupakan sebuah konsep yang diperkenalkan oleh LeCun dkk. pada tahun 1989. Penelitian tersebut mengacu pada penggunaan metode *backpropagation* yang diaplikasikan pada *Handwritten Zip Code* untuk *U.S. Postal Service*. Data yang digunakan pada penelitian tersebut berjumlah 9298 gambar tulisan tangan untuk sebuah kode pos yang ditulis oleh banyak orang yang berbeda, menggunakan berbagai macam ukuran, gaya penulisan, dan instrumen, dengan tingkat penulisan yang bervariasi. Pada penelitian tersebut, LeCun berhasil membuat sebuah jaringan tunggal yang dapat mempelajari seluruh operasi pengenalan, mulai dari gambar karakter yang dinormalisasi hingga klasifikasi akhir dan diberi nama *convolution networks* atau *convolutional neural network*.

Sesuai namanya, jaringan saraf konvolusional menunjukkan bahwa jaringan tersebut menggunakan operasi matematika yang disebut konvolusi. Konvolusi adalah jenis operasi linier khusus. Jaringan *convolutional* hanyalah jaringan saraf yang menggunakan konvolusi sebagai pengganti perkalian matriks umum di setidaknya satu lapisannya (Goodfellow, 2016).



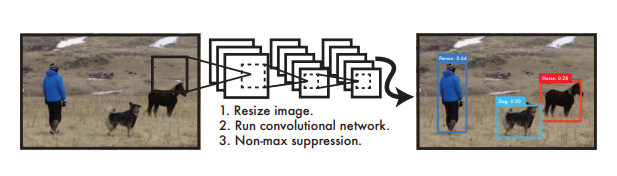
Gambar 2. 2 Proses Konvolusi 2D

Sumber: Goodfellow (2016)

Saat ini, CNN telah menjadi konsep yang sangat penting dalam bidang *computer vision* terutama dalam analisis citra. Arsitektur CNN meniru cara manusia memproses visual dengan mengadopsi konsep seperti *filter receptive field* dan penggabungan fitur hierarkis. Struktur CNN terdiri dari beberapa lapisan yang bekerja secara bersama-sama untuk mempelajari representasi fitur dari gambar input. Contohnya adalah lapisan konvolusi yang menggunakan *filter* atau *kernel* untuk memindai gambar dan mengekstrak fitur-fitur penting seperti tepi, tekstur, atau pola yang lebih kompleks. Setelah itu, lapisan aktivasi diterapkan untuk memperkenalkan non-linearitas ke dalam jaringan, yang membantu dalam pembelajaran representasi fitur yang lebih kompleks (Li *et al.*, 2021).

## 2.5 *You Only Look Once*

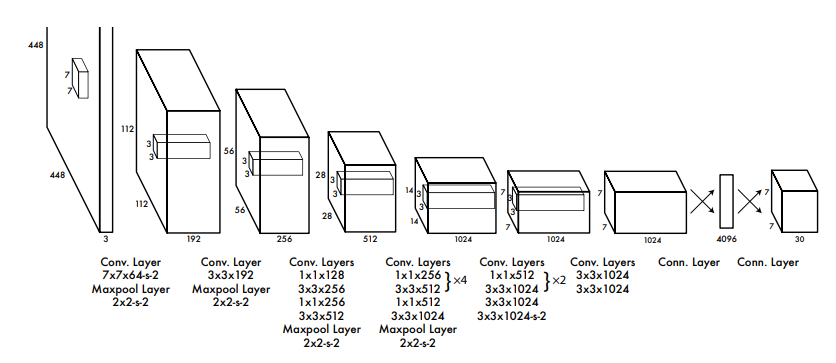
*You Only Look Once* (YOLO) adalah sebuah model deteksi objek revolusioner yang pertama kali diperkenalkan oleh Joseph Redmon dan Santosh Divvala pada tahun 2015. YOLO merombak pendekatan tradisional dalam deteksi objek dengan mengusulkan pendekatan yang lebih efisien dan cepat.



Gambar 2.3 Sistem deteksi objek dengan YOLO

Sumber: Redmon, *et al.* (2015)

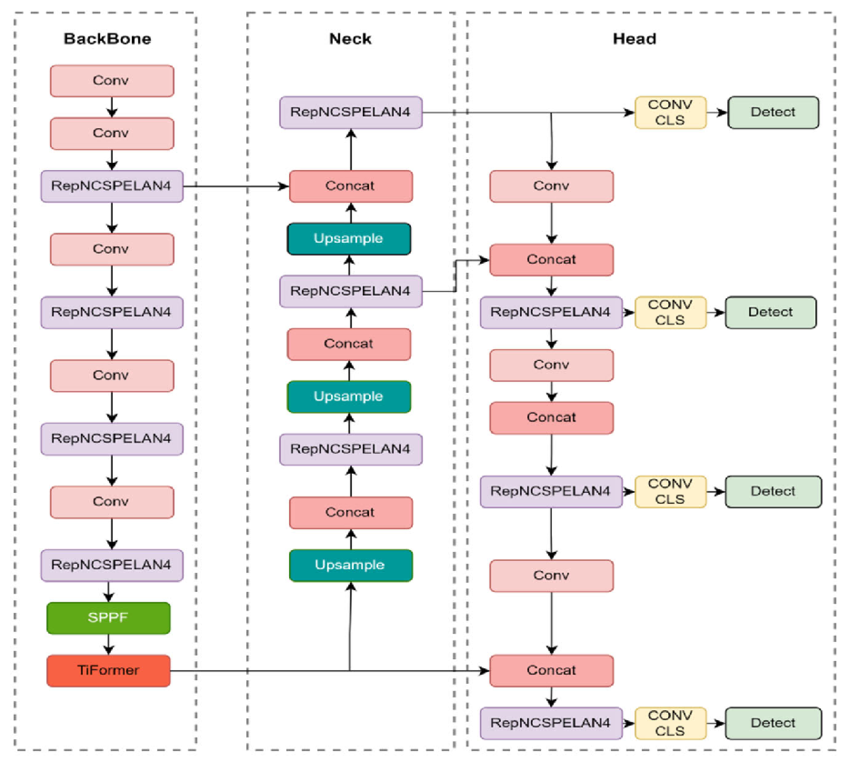
Arsitektur YOLO terdiri dari beberapa lapisan konvolusi yang menghasilkan *grid* dari prediksi *bounding box* dan probabilitas kelas. Lapisan-lapisan konvolusi ini digabungkan dengan lapisan-lapisan pooling untuk menghasilkan representasi fitur dari gambar input. Kemudian, menggunakan *fully connected layer*, prediksi untuk *bounding box* dan probabilitas kelas dilakukan secara bersamaan (Redmon *et al.,* 2019).



Gambar 2.4 Arsitektur YOLO

Sumber: Redmon, *et al.* (2015)

Salah satu keunggulan utama dari YOLO adalah kemampuannya untuk mendeteksi objek dengan cepat dalam waktu *realtime*, karena memproses seluruh gambar dalam satu langkah. Hal ini membuatnya sangat cocok untuk aplikasi seperti deteksi objek dalam video atau pada perangkat dengan komputasi yang terbatas (Wang &Liao, 2024).



Gambar 2.5 Arsitektur YOLOv9

Sumber: Li, *et al.* (2024)

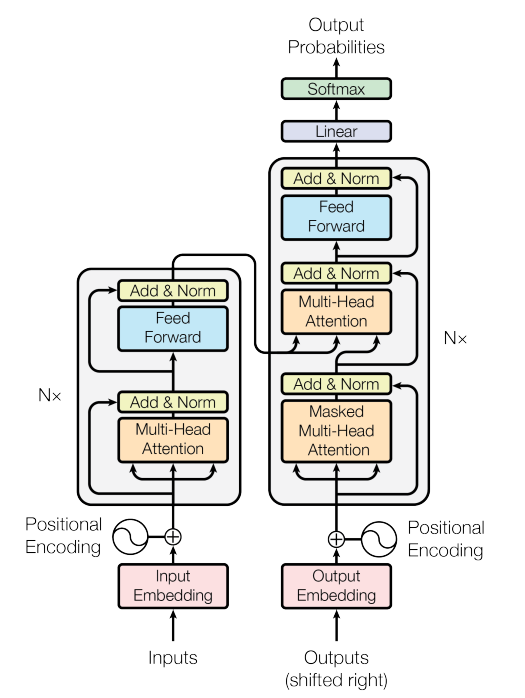
Seiring berkembangnya ilmu pengetahuan dan teknologi, muncul berbagai varian dari YOLO mulai dari YOLO9000 (YOLOv2), IA-YOLO, dan juga YOLOv9. YOLOv9 merupakan pengembangan dari sistem deteksi objek berbasis *Convolutional Neural Network* (CNN) yang mengombinasikan dua komponen utama, yaitu *Programmable Gradient Information* (PGI) dan *Generalized Efficient Layer Aggregation Network* (GELAN). Arsitektur YOLOv9 dimulai dengan pemrosesan input melalui serangkaian lapisan konvolusi pada GELAN, yang dirancang untuk memaksimalkan efisiensi jalur gradien dan parameter jaringan. PGI berperan dalam mengatasi kehilangan informasi yang sering terjadi selama proses *feedforward* dalam jaringan dalam (*information bottleneck*) dengan memperkenalkan mekanisme *reversible branch*. Mekanisme ini memastikan bahwa gradien yang diperoleh untuk memperbarui bobot jaringan optimal, sehingga jaringan dapat mempertahankan informasi penting dari fitur yang lebih dalam. Dalam YOLOv9, arsitektur GELAN menggunakan blok konvolusi yang dirancang untuk mencapai keseimbangan antara akurasi, *computational complexity*, dan *inference time*. Di tahap akhir, prediksi dilakukan menggunakan *head* deteksi yang telah dioptimalkan untuk efisiensi deteksi objek *multiscale*. Kombinasi PGI dan GELAN memungkinkan YOLOv9 untuk mencapai kinerja deteksi objek yang unggul dibandingkan dengan model sebelumnya, dengan parameter yang lebih sedikit namun akurasi yang lebih tinggi, khususnya pada dataset MS COCO (Wang *et al.,* 2024) .

Walau demikian, varian terbaru YOLO saat ini adalah YOLOv10 yang sudah menggunakan mekanisme Transformer di dalamnya. Setiap adanya kemunculan baru dari varian YOLO ini membuat YOLO menjadi lebih optimal yang secara spesifik dibutuhkan oleh pengembangnya. Jadi, bukan berarti karena YOLOv10 adalah versi terbaru, varian tersebut akan secara pasti mengungguli varian-varian lainnya dalam seluruh aspek maupun tugas yang dikerjakan, melainkan varian-varian dari YOLO ini akan menyesuaikan antara *trade-off* yang dibutuhkan. (Wang & Liao, 2024).

## *2.*6Transformer

Transformer merupakan sebuah arsitektur *neural network* yang menjadi *State of The Art* pada bidang *natural language processing* (NLP) sejak diperkenalkan oleh Vaswani dkk. (2017). Arsitektur Transformer menyajikan pendekatan yang berbeda dengan model *recurrent* atau CNN tradisional yang sebelumnya dominan dalam NLP.

Pusat dari Transformer adalah mekanisme *attention*, yang memungkinkan model untuk memberikan perhatian yang berbeda pada setiap bagian dari input *sequence*, terlepas dari jaraknya. Ini memungkinkan Transformer untuk memahami konteks semantik secara global dari sebuah *sequence*, yang merupakan keuntungan besar dalam NLP.

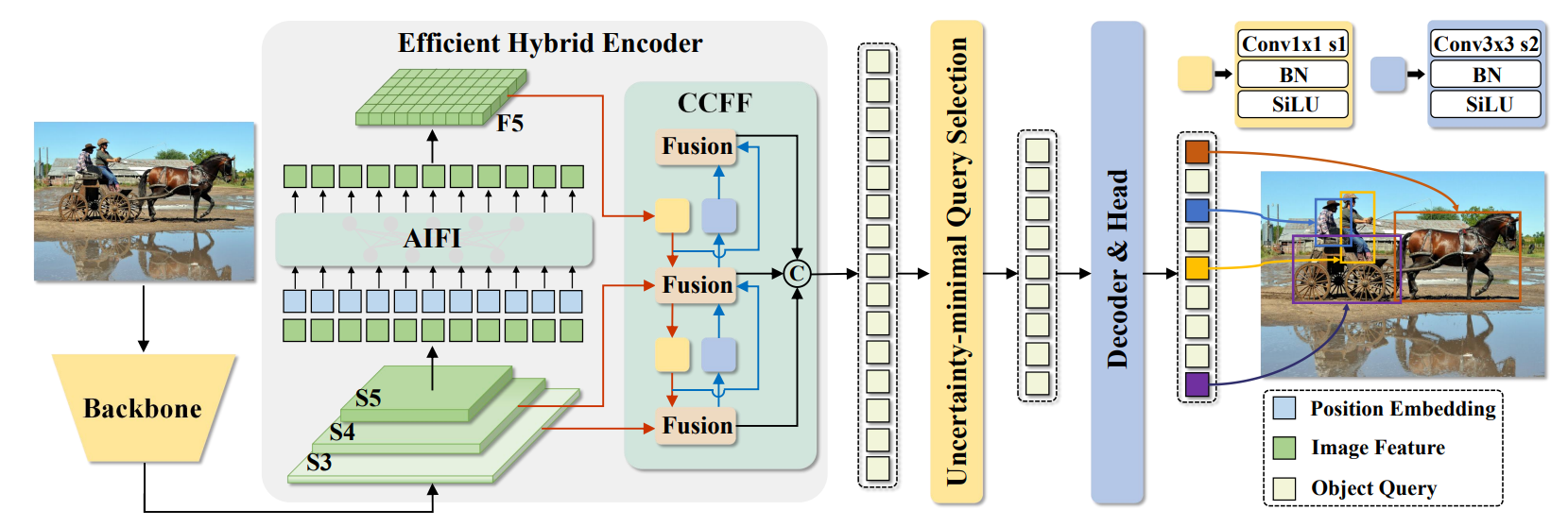


Gambar 2.6 Arsitektur Transformer

Sumber: Vaswani, *et al.* (2017)

Transformer terdiri dari dua komponen utama yaitu *encoder* dan *decoder*. *Encoder* bertugas untuk mengubah *input* *sequence* menjadi representasi fitur yang semantik, sedangkan *decoder* bertanggung jawab untuk menghasilkan *output* *sequence* dari representasi yang diberikan oleh *encoder*. Masing-masing *encoder* dan *decoder* terdiri dari beberapa *layer* yang terdiri dari *sub-layer*. *Sub-layer* dalam *encoder* terdiri dari *multi-head self-attention layer*, diikuti oleh *feed-forward neural network layer*. Di sisi lain, *sub-layer* dalam *decoder* memiliki *multi-head self-attention layer*, diikuti oleh *masked multi-head self-attention layer* dan *feed-forward neural network layer*. *Sub-layer* ini bekerja bersama-sama untuk memperoleh representasi fitur yang semantik dan menghasilkan *output* yang akurat (Vaswani *et al. ,* 2017).

Selain diterapkan dalam NLP, Transformer juga telah digunakan secara luas dalam bidang *computer vision*, khususnya dalam tugas-tugas seperti segmentasi gambar, pemrosesan gambar, dan pengenalan objek. Penggunaan Transformer dalam domain non-NLP menunjukkan fleksibilitas dan kemampuan adaptasi yang luar biasa dari arsitektur tersebut (Zhang *et al.*, 2021). Perluasan Transformer ke dalam domain *computer vision* ini salah satunya dipelopori oleh Carion dkk. di tahun 2020. Carion menciptakan suatu rancangan transformer yang dapat memahami konteks secara spasial dan diberi nama *DEtection TRansformer* atau DETR. DETR ini dapat menyaingi model *Faster* R-CNN secara signifikan dalam objek-objek besar berkat mekanisme *self-attention*-nya (Carion *et al.,* 2020). DETR ini kemudian dikembangkan oleh Zhao di tahun 2024 yang memungkinkan penggunaannya menjadi *realtime* dengan beberapa modifikasi seperti penghapusan mekanisme NMS, penerapan metode *Efficient Hybrid Encoder*, *Uncertainty-Minimal Query Selection*, dan menciptakan *Flexible Speed Tuning* yang dapat mengatur jumlah *decoder layers* sesuai kebutuhan tanpa perlu *retraining*. Varian *realtime* dari DETR yang dikembangkan oleh Zhao ini diberi nama *RealTime*-DETR atau RT-DETR (Zhao *et al.,* 2024)



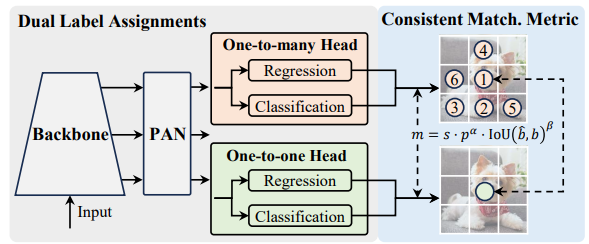
Gambar 2.7 Arsitektur RT-DETR

Sumber: Zhao, *et al.* (2024)

RT-DETR merupakan salah satu pengembangan dari arsitektur Transformer untuk deteksi objek secara *realtime*, dirancang untuk mengatasi keterbatasan model YOLO yang memerlukan *Non-Maximum Suppression* (NMS) dalam tahap *post-processing*. Arsitektur RT-DETR terdiri dari tiga komponen utama, yaitu *backbone*, *efficient hybrid encoder*, dan *decoder Transformer*. Fitur *multiscale* dari *backbone* diproses oleh *hybrid encoder* yang terdiri dari dua modul, yaitu *Attention-based Intra-scale Feature Interaction* (AIFI) untuk interaksi fitur *intrascale*, dan *CNN-based Cross-scale Feature Fusion* (CCFF) untuk *cross-scale fusion*. RT-DETR menggunakan mekanisme *uncertainty-minimal query selection* untuk memilih fitur *encoder* dengan ketepatan yang tinggi sebagai *query* awal untuk *decoder*, yang kemudian secara iteratif mengoptimalkan prediksi objek. Keunggulan utama dari RT-DETR adalah kemampuannya untuk menyelaraskan kecepatan dan akurasi, menghilangkan kebutuhan akan NMS, serta mendukung *tuning* kecepatan dengan menyesuaikan jumlah lapisan *decoder* tanpa perlu pelatihan ulang, menjadikannya solusi yang fleksibel untuk berbagai skenario deteksi *realtime*​ (Zhao *et al.,* 2024).

## 2.7 CNN-Transformer (YOLOv10)

Sesuai namanya, metode berbasis CNN-Transformer menggabungkan *Convolutional Neural Network* (CNN) dengan mekanisme Transformer. CNN dikenal unggul dalam mengekstraksi fitur spasial dari citra melalui jaringan konvolusi bertumpuk, sementara Transformer, yang awalnya digunakan dalam bidang *Natural Language Processing* (NLP), mampu menangkap hubungan jangka panjang dengan memanfaatkan mekanisme *self-attention*. Kombinasi ini bertujuan untuk mengatasi keterbatasan masing-masing metode, yaitu CNN yang terbatas dalam menangkap dependensi jarak jauh dan Transformer yang kurang optimal untuk pengolahan citra tanpa bantuan lapisan konvolusi. Metode *hybrid* ini memungkinkan sistem untuk melakukan deteksi objek secara lebih akurat, terutama dalam lingkungan yang kompleks seperti pencahayaan rendah atau perspektif yang variatif.



Gambar 2.8 Arsitektur YOLOv10

Sumber: Wang, *et al.* (2024)

YOLOv10 memperkenalkan serangkaian peningkatan signifikan dalam arsitektur deteksi objek berbasis CNN yang bertujuan untuk meningkatkan efisiensi dan akurasi. Arsitektur YOLOv10 mengimplementasikan strategi *dual label assignment*s untuk pelatihan tanpa *Non-Maximum Suppression* (NMS), yang menghilangkan kebutuhan *post-processing* dalam inferensi. Ini memungkinkan YOLOv10 mencapai deteksi *real-time* yang lebih efisien dengan menggabungkan kekuatan dari penugasan *one-to-many* dan *one-to-one*. Selain itu, desain arsitektur yang mengedepankan efisiensi-akurasi mencakup penggunaan *lightweight classification head*, *downsampling* yang terpisah antara *spatial* dan *channel*, serta *rank-guided block* untuk mengurangi redundansi komputasi tanpa mengorbankan kinerja. YOLOv10 juga memanfaatkan large-kernel convolution dan modul *partial self-attention* untuk memperluas *receptive field* dan meningkatkan kemampuan representasi global. Hasilnya, YOLOv10 berhasil melampaui model YOLO sebelumnya dan model *state-of-the-art* lainnya dalam hal latensi dan jumlah parameter, dengan tetap mempertahankan atau meningkatkan akurasi pada dataset MS COCO​ (Wang *et al.,* 2024).

## 2.8 *Hyperparameter Tuning*

*Hyperparameter* adalah parameter yang digunakan untuk mengontrol proses pelatihan model dan tidak dipelajari langsung dari data (Goodfellow *et al*., 2016). Berbeda dengan parameter model yang di-*update* selama pelatihan, seperti bobot dalam jaringan saraf, *hyperparameter* ditentukan sebelum proses pelatihan dimulai dan berperan penting dalam menentukan performa model. Beberapa contoh *hyperparameter* yang umum meliputi *Learning Rate*, *Batch Size*, dan *Epochs*.

*Hyperparameter* *Tuning* adalah proses mengoptimalkan nilai *hyperparameter* agar model dapat mencapai performa terbaik pada data tertentu (Bergstra & Bengio, 2012). Proses *tuning* ini biasanya melibatkan eksperimen dengan berbagai kombinasi nilai *hyperparameter*, menggunakan teknik seperti *grid search* atau *random search*, untuk menemukan konfigurasi yang paling sesuai dengan data dan tugas yang dihadapi. *Hyperparameter tuning* sangat penting karena *hyperparameter* yang tidak tepat dapat menyebabkan model *underfitting* ataupun *overfitting*, dimana model tidak mampu menangkap pola yang kompleks dari data atau justru terlalu menyesuaikan diri dengan data latih sehingga kinerjanya buruk pada data uji. Salah satu cara untuk mengoptimalkan performa model dengan *hyperparameter tuning* adalah mengkonfigurasi parameter *learning rate.*

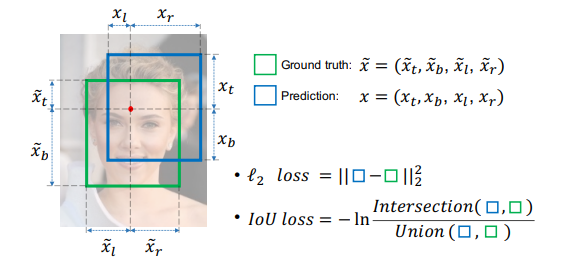
*Learning Rate* menentukan seberapa besar perubahan bobot dalam jaringan saraf terjadi setiap kali ada *update* berdasarkan *error* yang dihitung (Smith, 2017). *Learning Rate* yang terlalu tinggi dapat menyebabkan model menjadi tidak stabil karena langkah-langkah perbaikan yang terlalu besar sehingga tidak pernah mencapai titik minimum *error*. Sebaliknya, *Learning Rate* yang terlalu rendah dapat menyebabkan pelatihan model menjadi sangat lambat, memerlukan lebih banyak waktu dan sumber daya untuk mencapai konvergensi, dan bahkan bisa terjebak pada *local minimum* (Goodfellow *et al*., 2016).

*Batch Size* adalah *hyperparameter* lain yang memiliki pengaruh signifikan terhadap proses pelatihan model. *Batch Size* menentukan jumlah contoh data yang diproses sebelum bobot model diperbarui (Masters & Luschi, 2018). Pada umumnya, *Batch Size* yang lebih besar memungkinkan pelatihan yang lebih cepat dan stabil karena bisa lebih efisien dalam pemrosesan komputasi, terutama saat menggunakan GPU. Namun, *Batch Size* yang besar juga bisa mengurangi kemampuan model untuk keluar dari *local minimum*, yang dapat menyebabkan kinerja yang buruk pada generalisasi (Keskar *et al.,* 2017). Sebaliknya, *Batch Size* yang kecil seringkali menghasilkan variabilitas yang lebih tinggi dalam pembaruan bobot, yang bisa membantu model menghindari *local minimum*, meskipun dengan risiko pelatihan yang lebih lama.

## 2.9 *Loss Function*

*Loss function* merupakan komponen penting dalam pelatihan model terutama untuk *deep learning* dalam mendeteksi objek. *Loss function* digunakan untuk mengukur seberapa baik model memprediksi hasil yang diinginkan selama proses pelatihan. Secara umum, *object detection* memerlukan dua *loss function*, yaitu *loss function* untuk klasifikasi objek itu sendiri, dan yang *loss function* untuk *bounding box regression*. Salah satu *loss function* yang paling umum digunakan untuk *bounding box regression* adalah *Intersection over Union* (IoU), yang mengukur tumpang tindih antara prediksi *box* dengan *ground truth*. IoU loss ini pertama kali diperkenalkan oleh Yu (2016) untuk mengatasi permasalahan yang ada pada *loss function* sebelumnya yaitu ℓ2 loss (Yu *et al.,* 2016).

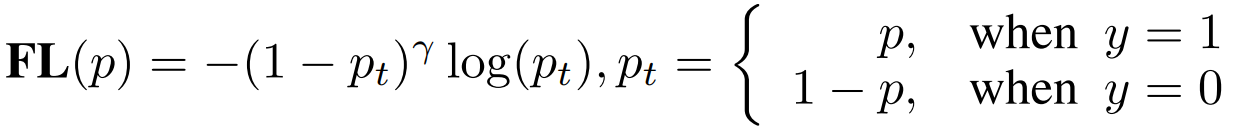
Kelemahan utama dari ℓ2 *loss* terletak pada koordinat sebuah *bounding box* yang dioptimalkan sebagai empat variabel yang independen. Asumsi ini mengabaikan fakta bahwa batas-batas suatu objek sangat berkorelasi. Hal ini dapat mengakibatkan beberapa kasus kegagalan di mana satu atau dua batas dari *bounding box* yang diprediksi sangat dekat dengan *ground truth*, tetapi keseluruhan *bounding box* tidak dapat diterima. IoU menyelesaikan masalah ini dengan menganggap satu *bounding box* adalah sebagai satu unit. Dengan begitu IoU *loss* dapat memberikan prediksi yang lebih akurat daripada ℓ2 *loss* (Yu *et al.* 2016). Secara umum, perbedaan ℓ2 *loss* dan IoU *loss* dapat dilihat pada Gambar 2.9.



Gambar 2.9 Perbedaan ℓ2 loss dan IoU loss

Sumber: Yu, *et al.* (2016)

Namun, IoU *loss* ini hanya digunakan untuk memprediksi *bounding box* pada *ground truth* dan belum dapat mengklasifikasikan kelas objek. Agar dapat mengklasifikasikan kelas objek, terdapat sebuah *loss function* bernama *Focal Loss* (FL) yang menerapkan konsep modulasi pada *cross entropy loss* untuk memfokuskan pembelajaran pada contoh-contoh yang sulit dan mengurangi bobot dari *easy negatives* (Lin *et al.,* 2017). Persamaan *Focal Loss* ini dijelaskan sebagai berikut.



(2.2)

Dimana *y* ∈ {1, 0} menentukan kelas *ground-truth* dan *p* ∈ [1, 0] menunjukkan estimasi probabilitas untuk kelas dengan label *y = 1* , γ adalah parameter fokus yang dapat dilakukan *tuning*. Secara spesifik, FL terdiri dari bagian *cross entropy − log(pt)* dan bagian *dynamically scaling factor (1 − pt) γ* , dimana *scaling factor (1 − pt) γ* secara otomatis menurunkan bobot kontribusi dari *easy examples* selama pelatihan dan dengan cepat memfokuskan model pada *hard examples*.

*Focal Loss* ini kembali dikembangkan menjadi *Generalized Focal Loss* (GFL) yang menggeneralisasi *Focal Loss* dari diskrit {1,0} ke versi kontinu. GFL dapat dikhususkan menjadi *Quality Focal loss* (QFL) dan *Distribution Focal Loss* (DFL), di mana QFL mendorong untuk mempelajari representasi gabungan yang lebih baik dari kualitas klasifikasi dan lokalisasi, dan DFL memberikan estimasi *bounding box* yang lebih informatif dan tepat dengan memodelkan lokasinya sebagai *General Distributions* (Li *et al.,* 2020).

IoU juga dikembangkan menjadi *Generalized IoU* (GIoU) *loss*. GIoU adalah perbaikan dari IoU *loss* yang digunakan untuk menghitung seberapa baik *bounding box* prediksi sesuai dengan *ground truth*. IoU biasa hanya mengukur tumpang tindih antara dua kotak, GIoU memperbaiki kelemahan IoU ketika tidak ada tumpang tindih antara prediksi dan *ground truth*. GIoU memperkenalkan area pembungkus minimal yang mencakup kedua kotak tersebut dan menghitung rasio antara area pembungkus dan selisih area antara kotak prediksi dengan *ground truth*. Dengan pendekatan ini, GIoU memberikan nilai yang lebih representatif bahkan ketika *bounding box* prediksi dan *ground truth* tidak bertumpang tindih, yang sering terjadi pada awal pelatihan model deteksi objek (Rezatofighi *et al*., 2019). Persamaan GIoU ini dijelaskan sebagai berikut.

(2.3)

Dimana *A* dan *B* merupakan dua bentuk *convex* sembarang dan *C* adalah bentuk *convex* terkecil yang dapat menyelimuti kedua bentuk A dan B secara keseluruhan.

Seiring dengan perkembangan *loss function* untuk prediksi *bounding box*, terdapat pula kemajuan dalam *loss function* untuk *object classification* yang dikenal sebagai *Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss)*. *BCEWithLogitsLoss* menggabungkan dua operasi penting yaitu *sigmoid activation* dan *binary cross-entropy loss* dalam sat langkah yang lebih efisien. Pada dasarnya, *loss* ini digunakan untuk mengukur kesalahan antara prediksi model (*logits*) dan target yang diinginkan (label biner). Pada fungsi BCE biasa, tahap yang dilakukan pertama kali adalah aktivasi sigmoid pada *output* model untuk mengubah nilai *logit* menjadi probabilitas antara 0 dan 1. Kemudian, probabilitas ini dibandingkan dengan target menggunakan binary *cross-entropy loss*, yang menghitung seberapa jauh perbedaan antara distribusi prediksi dan target. Namun, dalam *BCEWithLogitsLoss*, proses aktivasi sigmoid dilakukan secara internal, sehingga mengurangi risiko masalah numerik seperti *gradient vanishing* yang sering menggunakan sigmoid secara terpisah (Ruby & Yendapalli, 2020). Persamaan *BCEWithLogitsLoss* ini dijelaskan sebagai berikut.

(4)

Dimana *y* adalah label sebenarnya (0 atau 1) , *x* adalah *logit* atau *output* dari model, dan adalah fungsi sigmoid dari *logit* tersebut.

## 2.10 Metrik Evaluasi Model

Metrik evaluasi model digunakan untuk mengukur kinerja model. Dalam penelitian yang berkaitan dengan deteksi objek pada kondisi pencahayaan rendah, pemilihan metrik evaluasi model menjadi elemen krusial dalam menentukan performa model secara akurat. *F1-Score*, *Precision*, *Recall*, *Accuracy*, dan *Mean Average Precision* (mAP) adalah beberapa metrik yang sering digunakan dalam berbagai penelitian untuk menilai efektivitas model deteksi objek. *F1-Score* menggabungkan *Precision* dan *Recall* dalam satu ukuran, memberikan bobot yang seimbang antara *false positives* dan *false negatives*, yang sangat penting dalam kasus data yang tidak seimbang (Chicco & Jurman, 2020). *Precision* menggambarkan proporsi dari prediksi positif yang benar-benar akurat, sementara *Recall* mengukur sejauh mana model dapat mendeteksi semua contoh positif yang ada, keduanya adalah metrik yang esensial dalam konteks pengenalan objek yang sering kali mengalami masalah *misalignment* atau *noise* pada citra (Powers, 2020). Persamaan untuk metrik-metrik tersebut dijelaskan sebagai berikut,

, (2.5)

, (2.6)

, (2.7)

*accuracy =*  (2.8)

Dimana *TP* adalah *True Positive*, yaitu jumlah data di kelas positif yang berhasil diprediksi dengan benar sebagai positif, dan *TN* adalah *True Negative*, yaitu jumlah data di kelas negatif yang juga diprediksi dengan benar sebagai negatif. Sebaliknya, *FN* (*False Negative*) mengacu pada kasus di mana data yang seharusnya berada di kelas positif justru diprediksi sebagai negatif, dan *FP* (*False* *Positive*) terjadi ketika data dari kelas negatif salah diklasifikasikan sebagai positif.

Metrik-metrik dasar ini sangat penting dalam mengevaluasi performa model karena memberikan informasi rinci tentang jenis kesalahan yang dibuat oleh model. *Precision*, misalnya, sangat sensitif terhadap *False* *Positives* dan berguna dalam skenario di mana kesalahan deteksi positif (misalnya, mendeteksi objek yang sebenarnya tidak ada) harus diminimalkan. *Recall*, di sisi lain, lebih relevan dalam situasi di mana tidak terlewatkan satu pun deteksi benar sangat penting, seperti pada sistem keamanan atau aplikasi medis, di mana *False* *Negatives* (objek yang ada tetapi tidak terdeteksi) lebih berbahaya.

*Accuracy* sering kali digunakan sebagai metrik dasar dalam evaluasi model, tetapi metrik ini bisa menyesatkan terutama pada dataset yang tidak seimbang, di mana jumlah contoh dari satu kelas jauh lebih besar dibandingkan kelas lainnya (Saito *et al*., 2015). Oleh karena itu, untuk deteksi objek dalam citra, *mean Average Precision* (mAP) dianggap sebagai metrik yang lebih komprehensif karena memperhitungkan akurasi deteksi untuk setiap kelas objek, dan memberikan gambaran keseluruhan performa model dalam mendeteksi berbagai jenis objek (Zhao *et al*., 2019). Penggunaan mAP dalam tugas deteksi objek telah menjadi standar industri, terutama dalam kompetisi seperti COCO dan PASCAL VOC, yang mengharuskan model tidak hanya mendeteksi objek dengan benar, tetapi juga menentukan lokasi dengan presisi tinggi (Everingham *et al*., 2015). Persamaan mAP ini dijelaskan sebagai berikut,

*mAP=*  (2.9)

Dimana *n* adalah jumlah kelas yang ada dan adalah rerata nilai *precision* dari kelas *k*.

## 2.11 Pareto Frontier

Pareto Frontier, atau juga dikenal dengan Pareto Optimum, merupakan suatu *framework* yang digunakan dalam *multi-objective optimization* dan berasal dari teori ekonomi. Dalam domain *machine learning*, Pareto Frontier juga digunakan untuk mengidentifikasi solusi optimal ketika terdapat lebih dari satu tujuan yang saling bertentangan, seperti akurasi model dan waktu inferensi.

Solusi pada Pareto Frontier adalah solusi yang dinamakan Pareto Optimal, yang berarti tidak ada solusi lain yang dapat meningkatkan satu tujuan tanpa merugikan tujuan lainnya. Dengan demikian, Pareto Frontier memungkinkan identifikasi model yang mencapai keseimbangan optimal antara berbagai metrik (Benmeziane *et al*., 2022).

Dalam *multi-objective optimization*, tujuan yang sering dicapai adalah untuk meminimalkan atau memaksimalkan beberapa fungsi objektif secara bersamaan. Dalam optimasi dua *objective*, dan , solusi dianggap Pareto Optimal jika tidak ada solusi yang memenuhi:

*dan*  *dengan*  *atau*  (2.10)

Dengan kata lain, solusi adalah optimal jika setiap perbaikan pada salah satu tujuan akan menyebabkan kerugian pada tujuan lainnya. Sekumpulan solusi Pareto optimal ini membentuk Pareto Frontier, yang menggambarkan *trade-off* antara tujuan-tujuan yang saling bertentangan tersebut (Deb, 2001). Konsep Pareto Frontier telah terbukti sangat berguna dalam berbagai aplikasi termasuk *machine learning*. Salah satu contoh yang menonjol adalah dalam *Neural Architecture Search* (NAS), dimana Pareto Frontier digunakan untuk mengidentifikasi arsitektur yang optimal dalam hal latensi dan akurasi inferensi (Ye *et al.*, 2024). Selain itu, penelitian juga menunjukkan potensi Pareto Frontier dalam mengoptimalkan beberapa tujuan secara simultan, seperti *fairness*, *privacy*, dan *utility* dalam *modeling* (Yaghini et al., 2023).

# 

# **BAB 3 METODE PENELITIAN**

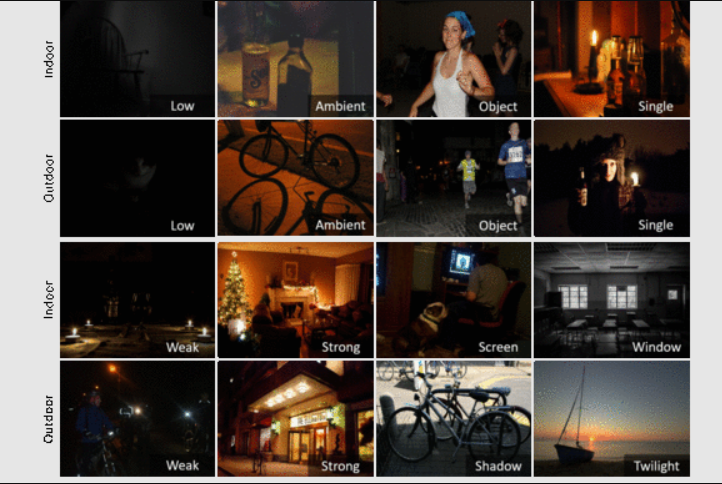
## 3.1 Lokasi dan Waktu

Penelitian dilakukan di Laboratorium MPL, Gedung Kuliah Bersama, Kampus C Universitas Airlangga yang beralamat di Jalan Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, Jawa Timur. Penelitian ini berlangsung dari bulan Juli 2024 hingga bulan Desember 2024, sesuai dengan jadwal yang telah ditentukan. Dalam periode tersebut, peneliti melakukan perencanaan, pengumpulan data, dan penyusunan skripsi.

Penelitian ini dilaksanakan dengan menggunakan *hardware* yang terdiri dari 12.7 GB RAM, GPU NVIDIA Tesla T4, dan CPU Intel(R) Xeon(R). Penggunaan platform perangkat lunak melibatkan torch-2.1.0+cu121 dan Python-3.10.12.

## 3.2 Bahan dan Alat

Data yang digunakan pada penelitian ini merupakan data sekunder *The Exclusively Dark* atau *ExDark* yang dibuat oleh Loh, Y. P., & Chan, C. S. (2019). Data ini merupakan kumpulan dari 7363 citra dengan pencahayaan sangat rendah hingga pencahayaan fajar dengan 12 kelas objek. Objek-objek tersebut adalah *table, people, motorbike, dog, cup, chair, bicycle, boat, bottle, bus, car,* dan *cat.* Sebagian besar data citra ini diperoleh dari *website* dan *search engine* seperti *Flickr.com, Photo-bucket.com, Imgur.com, Deviantart.com, Gettyimages.com* dan *Google Search* dengan kata kunci *dark, low-light, nighttime,* dan lain-lain. ExDark juga mengambil sampel dari dataset-dataset besar seperti PASCAL VOC, ImageNet, dan Microsoft COCO serta menambah variasi data dengan mengekstraksi setiap *frame* berpencahayaan rendah dalam film-film dan mengambil gambar secara manual menggunakan beragam model *smartphone* dan kamera.



Gambar 3.1 Sampel Exclusively Dark dataset

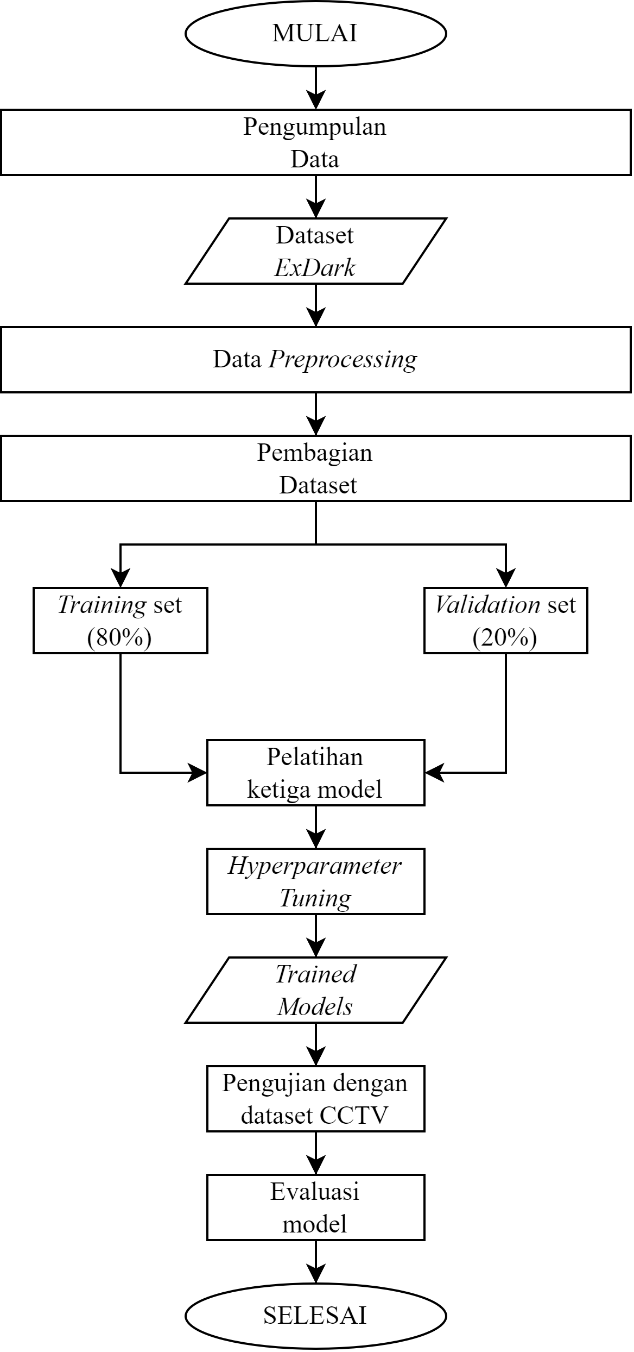
Sumber: Loh, Y. P., & Chan, C. S. (2019)

## 3.3 Cara Kerja

Penelitian ini tidak secara langsung menguji model pada perangkat CCTV atau sistem dengan sumber daya terbatas. Namun, salah satu kekuatan utama dari model deteksi objek modern adalah kemampuannya untuk melakukan generalisasi dari berbagai jenis dataset, termasuk dari dataset ExDark yang mengandung gambar dalam kondisi pencahayaan rendah yang sangat mirip dengan gambar yang dihasilkan oleh kamera CCTV di malam hari atau area yang gelap dengan berbagai macam sudut kamera.

Metode penelitian ini terdiri dari beberapa tahapan. Tahapan penelitian ini dimulai dengan data *preprocessing*, dimana dilakukan beberapa tahapan utama seperti pengubahan format label *bounding box* yang disesuaikan untuk kebutuhan model YOLO dan RT-DETR. Data kemudian ditingkatkan kualitasnya dengan teknik *contrast enhancement* menggunakan CLAHE. Selanjutnya, dilakukan normalisasi data menggunakan *min-max scaler* dan pembagian dataset menjadi *training* sebesar 80% dan *validation* sebesar 20% menggunakan teknik *stratified random sampling* yang stratanya dibagi berdasarkan kelas objeknya.

Setelah tahap *preprocessing* dan pembagian data, *train set* dan *validation set* digunakan untuk melatih tiga model (YOLOv9, RT-DETR, dan YOLOv10) dengan proses *hyperparameter tuning*. Setelah pelatihan selesai, model terlatih (*trained* *models*) akan dievaluasi berdasarkan performanya, termasuk *mean Average Precision* (mAP) dan waktu inferensi, kemudian dibandingkan menggunakan analisis *Pareto Frontier*. Model terlatih ini kemudian diuji menggunakan *test set* yang berupa dataset rekaman CCTV asli. Rekaman CCTV diperoleh dari sumber daring dengan kata kunci "*CCTV footage of a [subject*]", dimana *subject* mengacu pada jenis objek yang terdapat dalam dataset ExDark. Durasi rekaman bervariasi antara 5 hingga 10 detik, dengan 1 *frame* per detik.



Gambar 3.2 Flowchart Keseluruhan Proses

### 3.3.1 *Data Preprocessing*

Pada tahap ini dilakukan *data preprocessing* yang meliputi:

3.3.1.1 Pengubahan format label *bounding box*

Pada penelitian ini, dataset asli dari ExDark menggunakan format anotasi yang berbeda dengan format yang dibutuhkan oleh YOLO & RT-DETR untuk pelatihan model deteksi objek. Setiap anotasi di dataset ExDark mencakup beberapa informasi, antara lain 16 karakter pertama yang berisi data dari alat anotasi, kolom pertama yang berisi nama kelas objek, serta kolom kedua hingga kelima yang menyatakan koordinat *bounding box* dalam bentuk [l t w h], di mana l adalah jarak dari tepi kiri gambar, t adalah jarak dari tepi atas gambar, w adalah lebar *bounding box*, dan h adalah tinggi *bounding box*. Selain itu, anotasi ini juga memiliki informasi tambahan mengenai oklusi dan orientasi yang tidak akan digunakan.

Untuk mengubah format tersebut menjadi format yang dapat diterima oleh seluruh model, dilakukan proses konversi yang melibatkan beberapa langkah penting. Pertama, nama kelas objek dikonversi menjadi ID kelas yang sesuai. Selanjutnya, koordinat *bounding box* yang awalnya dinyatakan dalam bentuk [l t w h] dikonversi menjadi [center\_x, center\_y, width, height] di mana center\_x dan center\_y adalah koordinat pusat *bounding box*, sedangkan *width* dan *height* tetap merupakan dimensi dari *bounding box*. Semua nilai ini kemudian dinormalisasi dengan membagi nilaix*,* y*,* width*,* dan heightdengan lebar dan tinggi gambar. Hasil akhir dari proses ini dituliskan ke dalam *file* dengan format .yaml yang sesuai untuk digunakan dalam pelatihan model.

3.3.1.2 *Contrast Enhancement*

Setelah proses konversi format label *bounding box*, tahap selanjutnya dalam metodologi ini adalah penerapan teknik normalisasi data menggunakan *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Teknik ini dipilih karena kemampuannya dalam meningkatkan kontras gambar, khususnya pada kondisi pencahayaan rendah, seperti yang terdapat pada dataset ExDark. CLAHE bekerja dengan membagi gambar menjadi blok-blok kecil, kemudian melakukan *equalization* pada histogram setiap blok secara adaptif. Langkah ini memastikan bahwa kontras lokal pada setiap bagian gambar dapat ditingkatkan secara proporsional, tanpa menyebabkan amplifikasi *noise* yang berlebihan.

Dalam penelitian ini, CLAHE diterapkan untuk meningkatkan visibilitas objek pada gambar, dengan tujuan memaksimalkan akurasi deteksi objek pada kondisi pencahayaan yang beragam. Proses ini penting karena peningkatan kontras yang dihasilkan dari CLAHE membantu model deteksi objek dalam mengidentifikasi objek secara lebih efektif, terutama dalam skenario dengan pencahayaan yang tidak merata.

3.3.1.3 Normalisasi Data dengan *min-max scaler*

Pada tahap ini, seluruh data gambar dalam dataset dinormalisasi menggunakan teknik *min-max scaler*. Proses ini dilakukan dengan mengubah nilai piksel gambar ke dalam rentang 0 hingga 1, sehingga seluruh data memiliki skala yang seragam. Teknik *min-max scaler* dipilih karena sesuai dengan karakteristik dataset citra yang memiliki nilai piksel dalam rentang yang terbatas, dan teknik ini menjaga distribusi relatif antar nilai piksel. Hal ini penting untuk memastikan model deteksi objek dapat memproses gambar secara optimal, tanpa distorsi skala yang berlebihan. *min-max scaler* cocok untuk digunakan pada model seperti CNN dan Transformer yang sensitif terhadap perbedaan skala antar *input*, terutama pada tugas deteksi objek dalam citra pencahayaan rendah, di mana perbedaan kontras dapat mempengaruhi performa model.

3.3.1.4 Pembagian data

Tahap selanjutnya dalam *preprocessing* ini adalah membagi dataset ke dalam dua set, yaitu *train* (80%), *validation* (20%), dan akan diuji pada *test* set berupa rekaman asli CCTV dari dunia nyata. Pada tahap ini, pemilihan *train* set dan *validation* set akan menggunakan teknik *Stratified Random Sampling*. Teknik ini dipilih karena setiap gambar dalam dataset terbagi ke dalam 12 kelas yang berbeda, dan tujuan dari stratifikasi adalah memastikan bahwa distribusi kelas dalam set *validation* tetap seimbang, sehingga mewakili proporsi yang sama dengan dataset asli. Teknik ini penting untuk menghindari bias dalam proses validasi model dan memastikan bahwa semua kelas terwakili secara proporsional selama evaluasi.

3.3.1.5 *Data* *Augmentation*

Setelah membagi dataset ke dalam *train set* dan juga *validation set*, tahap terakhir adalah melakukan augmentasi data pada *train set* yang sudah dibuat. Proses augmentasi data ini dilakukan dengan menggunakan bantuan dari modul *Ray Tune* yang dapat mengotomasi proses augmentasi dalam bentuk parameter. Metode *Ray Tune* ini menggunakan sistem pengambilan sampel dari distribusi uniform yang akan mengambil satu nilai dari batas *lower* dan *upper* yang telah ditentukan. Parameter augmentasi dengan hasil terbaik kemudian akan disimpan dan dibandingkan dengan hasil parameter augmentasi pada iterasi selanjutnya. Iterasi dari pengambilan nilai acak ini akan dilakukan sebanyak lima kali. Dengan memanfaatkan prosedur ini *computational cost* yang dibutuhkan diharapkan berkurang secara signifikan. Parameter yang akan dimasukkan ke dalam otomasi augmentasi *Ray Tune* dijelaskan dalam Tabel 3.1.

Tabel 3.1 Deskripsi Parameter Augmentasi

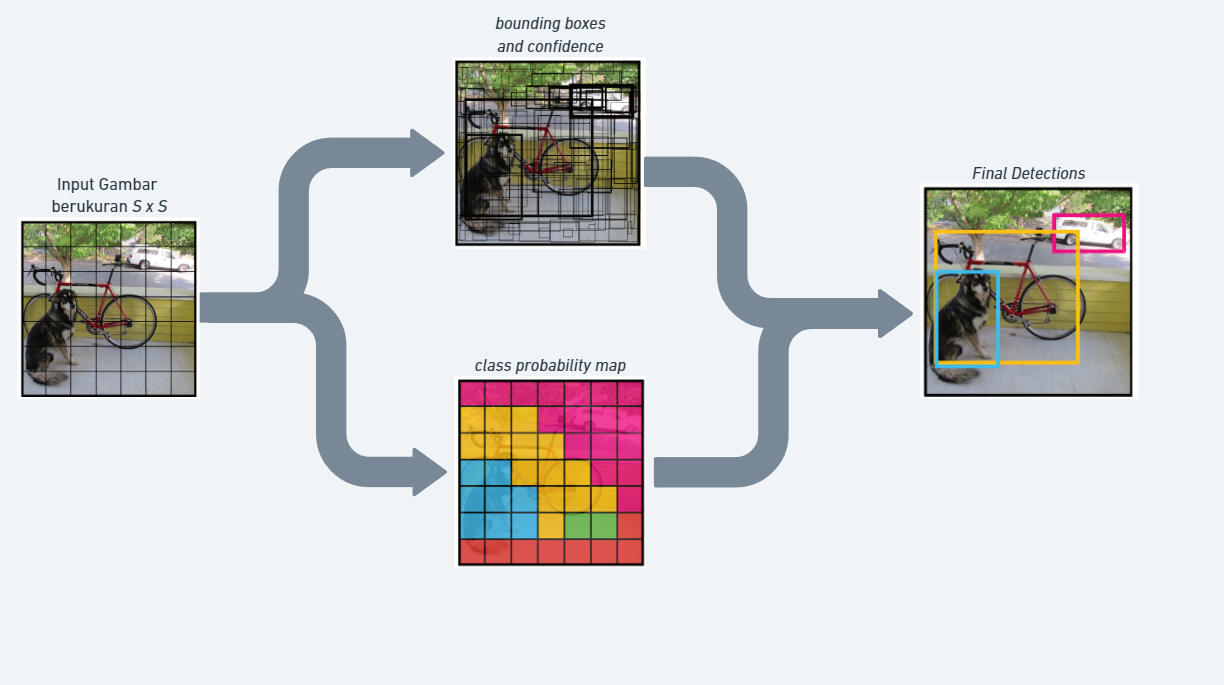
| ***Parameter*** | ***Value Range*** |
| --- | --- |
| *Rotation augmentation range* | 0,0 ; 45,0 |
| *Translation augmentation range* | 0,0 ; 0,9 |
| *Scaling augmentation range* | 0,0 ; 0,9 |
| *Shear augmentation range* | 0,0 ; 10,0 |
| *Perspective augmentation range* | 0,0 ; 0,001 |
| *Horizontal flip augmentation probability* | 0,0 ;1,0 |
| *Mosaic augmentation probability* | 0,0 ; 1,0 |

### 3.3.2 Implementasi Model

Pada tahap ini model berbasis CNN, Transformer, dan CNN-Transformer akan dilatih menggunakan *training set* serta akan divalidasi dengan *validation set*. Setelah itu akan dilakukan proses *hyperparameter tuning* untuk mendapatkan kombinasi parameter optimal.

3.3.2.1 Model berbasis CNN

Proses inferensi model berbasis CNN yaitu YOLOv9 yang digunakan pada penelitian ini ditunjukkan pada Gambar 3.3. Proses dimulai dengan sebuah gambar *input* berukuran S×S. dimana setiap sel grid memiliki tugas untuk mendeteksi objek yang pusatnya jatuh dalam sel tersebut.



Gambar 3.3 Proses Inferensi Model berbasis CNN

YOLO bekerja dengan dua cabang utama. Cabang pertama bertugas memprediksi *bounding boxes* dan *confidence scores*. Setiap sel pada grid memprediksi beberapa *bounding boxes*, yang masing-masing berisi informasi tentang koordinat (x, y), lebar (*width*), tinggi (*height*), dan nilai kepercayaan yang mencerminkan seberapa yakin model bahwa kotak tersebut mengandung objek dan seberapa tepat prediksi tersebut dibandingkan dengan objek sebenarnya.

Cabang kedua dari arsitektur ini menghasilkan *class probability map*. Setiap sel *grid* juga memprediksi probabilitas kelas untuk berbagai jenis objek yang mungkin berada dalam sel tersebut. Probabilitas ini menunjukkan seberapa besar kemungkinan objek dalam sel tersebut termasuk dalam setiap kelas yang ada.

Hasil dari kedua cabang ini digabungkan untuk menghasilkan deteksi *final*. *Bounding boxes* yang memiliki nilai kepercayaan tinggi dan probabilitas kelas yang signifikan akan dipilih sebagai hasil deteksi akhir, ditampilkan sebagai kotak yang mengelilingi objek yang terdeteksi dalam gambar.

YOLOv9 memiliki lima tipe model yang dibedakan berdasarkan jumlah parameternya. Namun, karena keterbatasan sumber daya, penelitian ini hanya akan menggunakan tiga tipe model yang berada di tengah-tengah spektrum jumlah parameter. Pemilihan ketiga tipe ini dilakukan dengan tujuan untuk memprediksi tren kinerja, yang diharapkan dapat mewakili performa dua tipe model lainnya yang tidak digunakan dalam penelitian ini. Tipe-tipe model YOLOv9 yang akan diuji pada penelitian ini dijelaskan dalam Tabel 3.2.

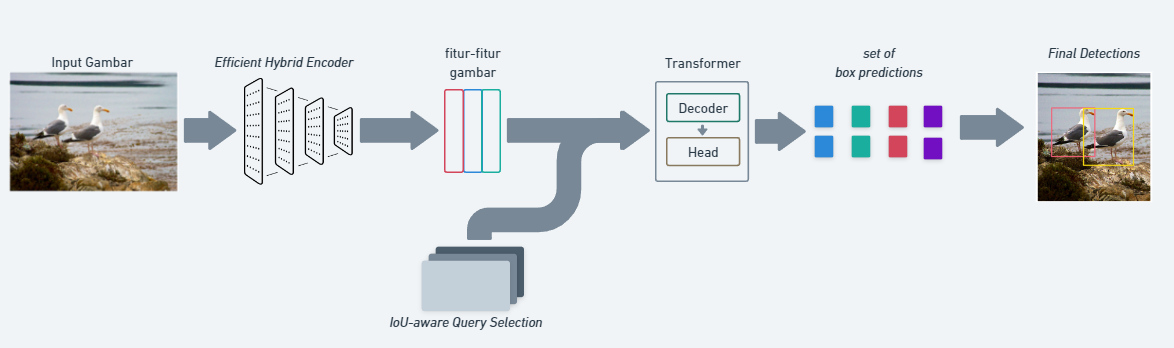
Tabel 3.2 Tipe Model YOLOv9

|  |  |
| --- | --- |
| **Tipe** | **Jumlah Parameter**  **(dalam juta)** |
| *YOLOv9s* | 7,2 |
| *YOLOv9m* | 20,1 |
| *YOLOv9c* | 25,5 |

Proses inferensi model berbasis CNN, yaitu YOLOv9, yang digunakan dalam penelitian ini akan bersumber dari library *Ultralytics*. Library ini menyediakan implementasi YOLO yang telah dioptimalkan untuk berbagai aplikasi deteksi objek secara *real-time*. Dalam implementasinya, *Ultralytics* menggunakan beberapa *loss function* yang penting untuk meningkatkan kinerja deteksi objek pada YOLOv9, yaitu *box-loss, cls-loss,* dan *dfl-loss*. Berdasarkan dokumentasi *Ultralytics*,algoritma yang digunakan oleh *box-loss* adalah *Intersection over Union* *loss*, untuk *cls-loss* adalah *BCEWithLogitsLoss*, dan *dfl-loss* adalah *Distribution Focal Loss*.

3.3.2.2 Model berbasis Transformer

Proses inferensi model berbasis Transformer yang menggunakan model *Real-Time DEtection TRansformer* (RT-DETR) yang digunakan pada penelitian ini ditunjukkan pada Gambar 3.4. Proses dimulai dengan memasukkan gambar *input* ke dalam *convolutional neural network* (CNN). CNN berfungsi untuk mengekstraksi fitur-fitur penting dari gambar, menghasilkan representasi fitur yang lebih terperinci dan kompak.



Gambar 3.4 Proses Inferensi Model Berbasis Transformer (RE-DETR)

Fitur-fitur gambar yang dihasilkan dari CNN kemudian diberi pengkodean posisi (*positional encoding*) untuk mempertahankan informasi spasial dari gambar asli. *Positional encoding* ini penting karena Transformer memerlukan informasi posisi untuk memproses urutan data secara efektif.

Selanjutnya, fitur-fitur gambar yang telah diberi pengkodean posisi ini dimasukkan ke dalam Transformer. Transformer terdiri dari dua komponen utama, *Encoder* dan *Decoder*. *Encoder* bertugas untuk memproses dan memahami representasi fitur gambar secara menyeluruh, sementara *Decoder* menghasilkan prediksi *bounding boxes* dan kelas objek dari fitur-fitur yang telah diproses.

Set *box prediction* yang dihasilkan oleh *Decoder Transformer* mencakup informasi koordinat dan kategori objek yang terdeteksi. Pada tahap akhir, prediksi-prediksi ini digabungkan untuk menghasilkan deteksi akhir, di mana objek-objek yang terdeteksi ditandai dengan *bounding boxes*  pada gambar asli.

Tipe-tipe model RT-DETR yang akan diuji pada penelitian ini dijelaskan dalam Tabel 3.3.

Tabel 3.3 Tipe Model RT-DETR

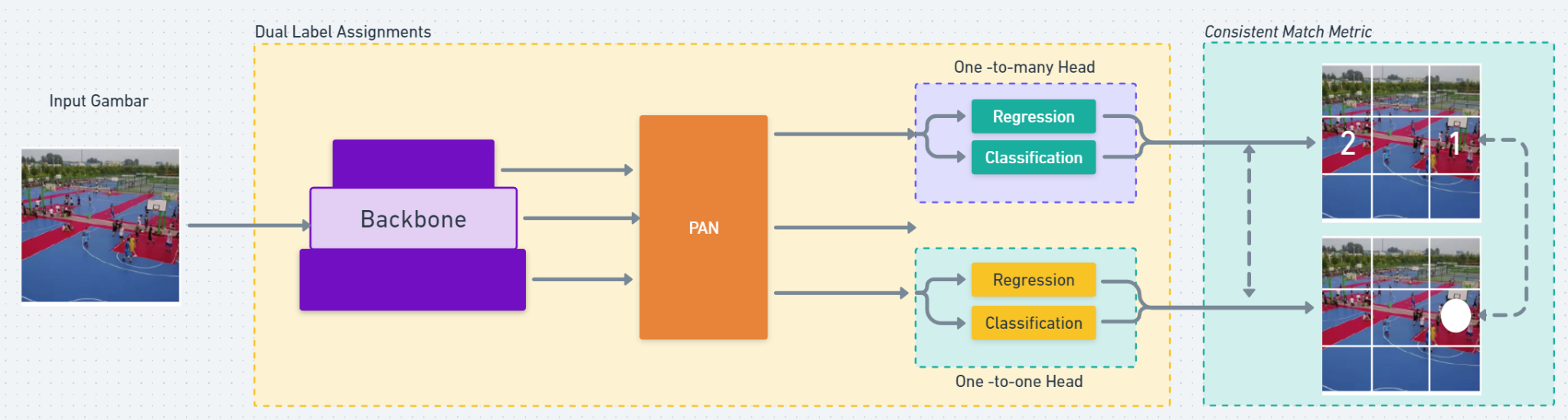
|  |  |
| --- | --- |
| **Tipe** | **Jumlah Parameter**  **(dalam juta)** |
| *RT-DETR-L* | 32,9 |
| *RT-DETR-X* | 67,4 |

Proses inferensi model berbasis Transformer, yaitu RT-DETR, yang digunakan dalam penelitian ini akan bersumber dari library *Ultralytics*. Dalam implementasinya, *Ultralytics* menggunakan beberapa *loss function* yang penting untuk meningkatkan kinerja deteksi objek pada RT-DETR, yaitu *GIoU-loss* dan *cls-loss*. Berdasarkan dokumentasi *Ultralytics*,algoritma yang digunakan oleh *GIoU-loss* adalah *Generalized Intersection over Union* *loss* dan untuk *cls-loss* adalah *BCEWithLogitsLoss*.

3.3.2.3 Model berbasis CNN-Transformer

Proses inferensi model berbasis CNN-Transformer yaitu model YOLOv10 yang digunakan pada penelitian ini ditunjukkan pada Gambar 3.5. YOLOv10 mengimplementasikan *Dual Label Assignments* untuk meningkatkan kinerja deteksi objek, yaitu melalui pendekatan *one-to-one* dan *one-to-many*. Pada pendekatan *one-to-one*, setiap prediksi dicocokkan dengan satu *ground truth*. Strategi ini menghindari penggunaan proses *post-processing* seperti *Non-Maximum Suppression* (NMS) yang biasa digunakan untuk menangani prediksi *overlapping predictions*. Namun, kelemahan dari *one-to-one* adalah kurangnya pengawasan yang optimal, yang dapat mempengaruhi akurasi dan kecepatan konvergensi selama pelatihan.

Sebaliknya, pendekatan *one-to-many* memungkinkan satu prediksi dicocokkan dengan beberapa *ground truth*. Meskipun pendekatan ini memberikan *supervised signal* yang lebih baik, metode ini memerlukan NMS untuk mengurangi *overlapping predictions*. YOLOv10 menggabungkan kedua metode ini melalui *dual label assignments*. Dalam tahap pelatihan, dua "*head*" model yang mewakili kedua pendekatan ini dioptimalkan secara bersamaan. Namun, pada tahap inferensi, hanya *one-to-one* *head* yang digunakan untuk membuat prediksi, sehingga memungkinkan penerapan YOLOv10 secara *end-to-end* tanpa *cost* tambahan pada waktu inferensi.



Gambar 3.5 Proses Inferensi Model berbasis CNN-Transformer (YOLOv10)

Selama proses *label assignment*, baik pendekatan *one-to-one* maupun *one-to-many* menggunakan *Consistent Matching Metri*c untuk mengevaluasi kesesuaian antara prediksi dengan *instance* objek. Metrik yang digunakan oleh YOLOv10 adalah fungsi yang mempertimbangkan skor klasifikasi , IoU, dan prior spasial, yaitu apakah titik prediksi berada di dalam *instance* objek.

YOLOv10 memiliki enam tipe model yang dibedakan berdasarkan jumlah parameternya. Namun, karena keterbatasan sumber daya, penelitian ini hanya akan menggunakan tiga tipe model yang berada di tengah-tengah spektrum jumlah parameter. Pemilihan ketiga tipe ini dilakukan dengan tujuan untuk memprediksi tren kinerja, yang diharapkan dapat mewakili performa tiga tipe model lainnya yang tidak digunakan dalam penelitian ini. Tipe-tipe model YOLOv10 yang akan diuji pada penelitian ini dijelaskan dalam Tabel 3.4.

Tabel 3.4 Tipe Model YOLOv10

|  |  |
| --- | --- |
| **Tipe** | **Jumlah Parameter**  **(dalam juta)** |
| *YOLOv10-S* | 7,2 |
| *YOLOv10-M* | 15,4 |
| *YOLOv10-B* | 20,5 |

Proses inferensi model berbasis CNN-Transformer, yaitu YOLOv10, yang digunakan dalam penelitian ini akan bersumber dari library *Ultralytics*. Sama halnya seperti YOLOv9, *Ultralytics* menggunakan beberapa *loss function* pada YOLOv10, yaitu *box-loss, cls-loss,* dan *dfl-loss*. Algoritma yang digunakan oleh *box-loss* adalah *Intersection over Union* *loss*, untuk *cls-loss* adalah *BCEWithLogitsLoss*, dan *dfl-loss* adalah *Distribution Focal Loss*.

3.3.2.4 *Hyperparameter Tuning*

Pada ketiga model dilakukan *hyperparameter tuning* dengan parameter yang ditunjukkan pada Tabel 3.5.

Tabel 3.5 Deskripsi Hyperparameter yang Diuji

|  |  |
| --- | --- |
| **Parameter** | ***Value*** |
| *Learning Rate* | 0,001 ; 0,01 |
| *Batch Size* | 8 ; 16 |

Dalam penelitian ini,konfigurasi *Learning rate & Batch Size* menjadi esensial karena model deteksi objek harus dapat menangani kompleksitas citra dengan pencahayaan rendah. Sementara itu, Epochs tidak dijadikan sebagai parameter yang di-*tuning* karena penggunaan *callbacks* seperti *early stopping* dan *patience* sudah cukup efektif dalam menghindari *overfitting*, dengan cara menghentikan pelatihan ketika tidak ada perbaikan pada performa model dalam beberapa iterasi (Prechelt, 1998). Epochs yang akan digunakan adalah 500 dengan *patience 50.* Pendekatan ini memastikan bahwa model tidak dilatih terlalu lama sehingga tetap efisien dan optimal dalam performa.

### 3.3.3 Metode Perbandingan Model

Dalam penelitian ini, perbandingan kinerja antara ketiga model, yaitu YOLOv9, RT-DETR, dan YOLOv10, akan dilakukan dengan mempertimbangkan akurasi dan *time complexity* menggunakan Pareto Frontier. Prosedur perbandingan model ini adalah sebagai berikut:

1. Pengumpulan Data Waktu Inferensi dan Nilai *mean Average Precision* (mAP)

Setiap model akan diuji untuk mengumpulkan data mAP dan waktu inferensi pada dataset yang sama. Selain itu, *cls\_loss* (*BCEWithLogitsLoss)* akan digunakan sebagai alat ukur *(loss function)* untuk ketiga model dikarenakan *loss function* ini dapat digunakan untuk ketiga model yang bersangkutan, yakni YOLOv9, RT-DETR, dan YOLOv10.

1. Konstruksi *Pareto Frontier*

Berdasarkan hasil pengujian, nilai-nilai akurasi dan waktu inferensi dari tiap model akan dibentuk ke dalam plot untuk menghasilkan *Pareto Frontier*. *Pareto Frontier* mengidentifikasi model yang optimal secara relatif, yaitu model yang menawarkan kombinasi terbaik dari mAP tinggi dan waktu inferensi rendah.

1. Analisis Perbandingan

Model-model yang terletak pada *Pareto Frontier* dikategorikan sebagai solusi Pareto optimal. Jika dua model memiliki tingkat mAP yang setara, maka model dengan waktu inferensi yang lebih singkat akan lebih diprioritaskan. Sebaliknya, jika waktu inferensi kedua model serupa, model dengan akurasi yang lebih tinggi akan dianggap lebih unggul. Model-model yang berada di luar Pareto Frontier dinilai kurang efisien karena terdapat model lain yang dapat memberikan kinerja yang lebih bagus, baik dalam mAP maupun waktu inferensi.

# 

# **DAFTAR PUSTAKA**

Alif, M. A. R., & Hussain, M. (2024). YOLOv1 to YOLOv10: *A comprehensive review of YOLO variants and their application in the agricultural domain*. *arXiv preprint arXiv:2406.10139*.

Benmeziane, H., Niar, S., Ouarnoughi, H., & El Maghraoui, K. (2022, May). *Pareto* *rank surrogate model for hardware-aware neural architecture search.* In 2022 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) (pp. 267-276). IEEE.

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization.* Journal of machine learning research, 13(2).

Cao, X., Zhang, L., & Liu, Y. (2023). *Information Flow-Based Fuzzy Cognitive Maps with Enhanced Interpretability*. Springer.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). *End-to-end object detection with transformers. In European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing.

Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). *Learning to see in the dark*. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3291-3300).

Chen, W., & Shah, T. (2021). *Exploring low-light object detection techniques*. *arXiv preprint arXiv:2107.14382*.

Chicco, D., & Jurman, G. (2020). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* BMC genomics, 21, 1-13.

Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., ... & Harada, T. (2022*). You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction*. *arXiv preprint arXiv:2205.14871*.

Cui, Z., Qi, G. J., Gu, L., You, S., Zhang, Z., & Harada, T. (2021*). Multitask aet with orthogonal tangent regularity for dark object detection*. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2553-2562).

Deb, K. (2001). *Multiobjective Optimization Using Evolutionary Algorithms*. Wiley, New York.

Dhillon, A., & Verma, G. K. (2019). *Convolutional neural network: a review of models, methodologies and applications to object detection*. Progress in Artificial Intelligence, *9*(2), 85–112. https://doi.org/10.1007/s13748-019-00203-0.

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2022). *Object detection using YOLO: challenges, architectural successors, datasets and applications*. Multimedia Tools and Applications, *82*(6), 9243–9275. <https://doi.org/10.1007/s11042-022-13644-y>.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). *The pascal visual object classes (voc) challenge*. International journal of computer vision, 8*8*, 303-338.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. In *MIT Press eBooks*. https://dl.acm.org/citation.cfm?id=3086952.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). *On large-batch training for deep learning: Generalization gap and sharp minima*. arXiv preprint arXiv:1609.04836.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). *Backpropagation applied to handwritten zip code recognition*. Neural computation, *1*(4), 541-551.

Lee, J., Bang, J., & Yang, S. (2017). *Object detection with sliding window in images including multiple similar object*s. 2017 International Conference on Information and Communication Technology Convergence *(*ICTC), Pp. 803–806. <https://doi.org/10.1109/ictc.2017.8190786>.

Li, J., Feng, Y., Shao, Y., & Liu, F. (2024). IDP-YOLOV9: *Improvement of Object Detection Model in Severe Weather Scenarios from Drone Perspective*. Applied Sciences, *14*(12), 5277.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020). *Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection*. Advances in Neural Information Processing Systems, *33*, 21002-21012.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). *A survey of convolutional neural networks: analysis, applications, and prospects*. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). *Focal loss for dense object detection.* In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

Loh, Y. P., & Chan, C. S. (2019*). Getting to know low-light images with the exclusively dark dataset*. Computer Vision and Image Understanding, *178*, 30-42.

Manongga, W. E., Chen, R. C., & Jiang, X. *Enhancing road marking sign detection in low-light conditions with YOLOv7 and contrast enhancement techniques*.

Masters, D., & Luschi, C. (2018). *Revisiting small batch training for deep neural networks*. arXiv preprint arXiv:1804.07612.

Peng, X., Li, J., & Wang, Y. (2023). *Data-driven machine learning approach for predicting dwell fatigue.* Scripta Materialia.

Powers, D. M. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv preprint arXiv:2010.16061.

Putri, G. R. C. K. B. (2021). *Peran Kamera Pengawas Closed-Circuit Television (CCTV) dalam Kontra Terorisme*. *Jurnal Lemhannas RI*, *9*(4), 100–116. https://doi.org/10.55960/jlri.v9i4.418

Prechelt, L. (2002). *Early stopping-but when?. In Neural Networks: Tricks of the trade* (pp. 55-69). Berlin, Heidelberg: Springer Berlin Heidelberg.

Redmon, J., & Farhadi, A. (2018). YOLOV3*: an incremental improvement*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1804.02767.

Redmon, J., Santosh, D. H. H., Ross, G., & Farhadi, A. (2015). *You only look once: Unified, Real-Time Object Detection.* arXiv (Cornell University*)*. <https://doi.org/10.48550/arxiv.1506.02640>.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). *Generalized intersection over union: A metric and a loss for bounding box regression*. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658-666).

Rouhbakhshmeghrazi, A., & Alizadeh, G. *Instance Segmentation of Messier Objects: YOLO vs. Mask R-CNN*.

Ruby, U., & Yendapalli, V. (2020). *Binary cross entropy with deep learning technique for image classification*. *Int. J. Adv. Trends Comput. Sci. Eng*, *9*(10).

Saito, T., & Rehmsmeier, M. (2015). *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PloS one, *10*(3), e0118432.

Shorten, C., & Khoshgoftaar, T. M. (2019). *A survey on Image Data Augmentation for Deep Learning*. Journal of Big Data*, 6*(1), 60.

Smith, L. N. (2017, March). *Cyclical learning rates for training neural networks.* In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 464-472). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5.

Voulodimos, A., Doulamis, A., & Protopapadakis, E. (2018). *Deep Learning for Computer Vision: A Brief review.* Computational Intelligence and Neuroscience, 2018, 1–13. <https://doi.org/10.1155/2018/7068349>.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). Yolov10: *Real-time end-to-end object detection*. arXiv preprint arXiv:2405.14458..

Wang, C. Y., & Liao, H. Y. M. (2024). YOLOv1 to YOLOv10: *The fastest and most accurate real-time object detection systems.* arXiv preprint arXiv:2408.09332.

Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2024). *Yolov9: Learning what you want to learn using programmable gradient information*. arXiv preprint arXiv:2402.13616.

Wiley, V., & Lucas, T. W. (2018). *Computer Vision and Image Processing: A paper review*. International Journal of Artificial Intelligence Research, 2(1), 22. <https://doi.org/10.29099/ijair.v2i1.42>.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., & Huang, T. (2016, October). *Unitbox: An advanced object detection network.* In Proceedings of the 24th ACM international conference on Multimedia (pp. 516-520).

Yaghini, M., Liu, P., Boenisch, F., & Papernot, N. (2023). *Learning with impartiality to walk on the pareto frontier of fairness, privacy, and utility*. arXiv preprint arXiv:2302.09183.

Ye, R., Chen, L., Liao, W., Zhang, J., & Ishibuchi, H. (2024). *Data-Driven Preference Sampling for Pareto Front Learning.* arXiv.

Yin, X., Yu, Z., Fei, Z., Lv, W., & Gao, X. (2023, September). *Pe-yolo: Pyramid enhancement network for dark object detection*. In *International Conference on Artificial Neural Networks* (pp. 163-174). Cham: Springer Nature Switzerland.

Yusuf, M. O., Hanzla, M., Al Mudawi, N., Sadiq, T., Alabdullah, B., Rahman, H., & Algarni, A. (2024). *Target Detection and Classification via EfficientDet and CNN over Unmanned Aerial Vehicles*. Frontiers in Neurorobotics, 18, 1448538.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). *Detrs beat yolos on real-time object detection.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16965-16974).

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). *Object detection with deep learning: A review*. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

Zheng, S., & Gupta, G. (2022). *Semantic-guided zero-shot learning for low-light image/video enhancement*. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision* (pp. 581-590).