

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada bidang *computer vision*, dua tugas yang sangat krusial adalah *object detection* dan *instance segmentation*. *Object detection* bertujuan untuk mengenali dan melokalisasi objek dalam citra menggunakan *bounding box*, sedangkan *instance segmentation* melakukan segmentasi tiap objek secara presisi hingga ke tingkat piksel (Li et al., 2022). Kombinasi keduanya memungkinkan sistem visual memahami isi gambar secara semantik dan spasial. Teknologi ini telah banyak digunakan dalam berbagai aplikasi seperti kendaraan otonom, sistem pengawasan, augmented reality, hingga bidang medis. Meskipun begitu, implementasi *instance segmentation* dalam skenario waktu nyata (*real-time*) masih menjadi tantangan, khususnya karena beban komputasi yang tinggi dan kebutuhan akan akurasi yang tetap terjaga (Wu et al., 2022).

Seiring berkembangnya *deep learning*, arsitektur berbasis *convolutional neural networks* (CNN) seperti Faster R-CNN dan Mask R-CNN telah banyak digunakan untuk menyelesaikan masalah segmentasi dan deteksi objek (Ren et al., 2015; He et al., 2017). Namun, CNN memiliki keterbatasan dalam pemodelan konteks global karena operasi konvolusi hanya bekerja secara lokal, sehingga setiap filter hanya memiliki *receptive field* terbatas (Hatamizadeh et al., 2023). Selain itu, CNN memproses semua piksel dalam sebuah gambar secara seragam, tanpa membedakan kepentingan atau kontribusi relatif dari tiap piksel terhadap konteks tugas yang sedang dijalankan. Hal ini menyebabkan inefisiensi baik dari sisi representasi maupun komputasi (Khan et al., 2022). Untuk mengatasi keterbatasan tersebut, berbagai arsitektur berbasis *Transformer* mulai diperkenalkan dalam bidang *computer vision*, seperti DETR (*DEtection TRansformer*), *Vision Transformer* (ViT), dan *Mask2Former*. Model-model ini memanfaatkan mekanisme *self-attention* yang memungkinkan hubungan spasial antar elemen dalam citra dapat dimodelkan secara global dalam satu tahap komputasi (Carion et al., 2020; Dosovitskiy et al., 2021; Cheng et al., 2022). Tidak seperti CNN yang sangat bergantung pada struktur lokal dan pemrosesan seragam, *self-attention*

dalam Transformer secara adaptif menyesuaikan perhatian terhadap bagian penting dari input, memberikan keunggulan dalam representasi kontekstual.

Salah satu arsitektur berbasis Transformer yang menunjukkan efisiensi tinggi adalah *Real-Time Detection Transformer* (RT-DETR). RT-DETR dirancang untuk mengatasi kelemahan model DETR, khususnya dalam hal kecepatan inferensi yang rendah untuk penggunaan real-time (Lv et al., 2023). RT-DETR menggunakan *efficient hybrid encoder* yang membagi proses *Attention-based Intra-scale Feature Interaction* (AIFI) dan *CNN-based Cross-scale Feature Fusion* (CCFF), memungkinkan pemrosesan fitur multi-skala secara lebih cepat dan efisien tanpa mengorbankan kualitas (Lv et al., 2023). Selain itu, mekanisme *Uncertainty-minimal query selection*, meningkatkan kualitas *initial object queries* untuk decoder dan meningkatkan akurasi deteksi (Lv et al., 2023). Hasil eksperimen menunjukkan bahwa RT-DETR-R50 mencapai sekitar 53.1 % AP pada COCO dengan kecepatan inferensi 108 FPS pada T4 GPU, serta melampaui performa DINO-Deformable-DETR dalam hal AP dan FPS (Lv et al., 2023). Namun, RT-DETR masih terbatas pada tugas *object detection* dan belum memiliki kemampuan segmentasi piksel langsung, sehingga kurang cocok untuk aplikasi yang memerlukan pemahaman spasial mendalam seperti instance segmentation.

Sementara itu, Mask DINO yang dikembangkan oleh (Li et al., 2023) menghadirkan arsitektur *Transformer* terintegrasi yang mampu melakukan *object detection* dan *instance segmentation* secara end-to-end. Mask DINO memperluas model DINO (DETR dengan Improved DeNoising Anchor Boxes) dengan menambahkan *mask head*, serta menerapkan *denoising training* dan *hybrid matching* untuk memperkuat stabilitas dan performa selama *training* (Li et al., 2022). Dengan pendekatan *query-based mask prediction*, model ini memproyeksikan *query embeddings* ke dalam high-resolution pixel embedding map menggunakan operasi *dot-product*, menghasilkan binary masks dengan presisi tinggi (Li et al., 2022). Hasil eksperimen pada benchmark COCO menunjukkan Mask DINO mencapai Mask AP 54.5% pada COCO *instance segmentation* menggunakan backbone ResNet-50, serta kinerja unggul dalam panoptic (PQ 59.4) dan semantic segmentation (mIoU 60.8 pada ADE20K), melampaui berbagai model khusus seperti Mask2Former (Li et al., 2022). Meskipun unggul dalam hal

akurasi, Mask DINO belum dioptimalkan untuk efisiensi inferensi secara real-time dan belum menawarkan kecepatan prosesor inferensi yang ringan sehingga belum ideal digunakan pada sistem dengan batasan latency tinggi (Li et al., 2022).

Oleh sebab itu, penelitian ini bertujuan untuk meneliti dan memodifikasi RT-DETR (Lv et al., 2023) agar mampu melakukan *instance segmentation* secara real-time dengan tetap mempertahankan efisiensi inferensi yang menjadi keunggulan utamanya. Modifikasi dilakukan dengan mengadopsi tiga komponen utama dari Mask DINO (Li et al., 2022), yaitu *mask branch* yang memproyeksikan *query embeddings* ke pixel-level mask predictions, *unified denoising training* untuk mempercepat dan menstabilkan proses *training*, serta *hybrid matching* yang memungkinkan integrasi antara detection dan segmentation dalam satu mekanisme assignment. Ketiga komponen tersebut telah terbukti meningkatkan robustness dan generalisasi model pada benchmark seperti COCO dan ADE20K, bahkan dalam skenario *open vocabulary* dan *limited supervision* (Li et al., 2022). Integrasi konsep-konsep ini menghasilkan model yang diusulkan bernama Insta-RT-DETR, yang menggabungkan efisiensi dari RT-DETR dan kemampuan segmentasi presisi dari Mask DINO dalam satu arsitektur Transformer yang ringan dan serbaguna. Pendekatan ini sejalan dengan tren pengembangan *query-based instance segmentation* yang fokus pada optimalisasi trade-off antara *Average Precision (AP)* dan *Frames Per Second (FPS)*, sebagaimana ditunjukkan dalam karya-karya terbaru seperti FastInst dan Mask2Former (He et al., 2023; Cheng et al., 2022). Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem visual real-time yang akurat, efisien, dan mudah diadaptasi untuk berbagai kebutuhan praktis, termasuk pengawasan cerdas, kendaraan otonom, dan edge deployment.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, rumusan masalah dapat diformulasikan sebagai berikut:

- 1) Bagaimana cara modifikasi arsitektur *Real-Time Detection Transformer* (RT-DETR) agar mampu melakukan *instance segmentation* dengan menambahkan *mask branch* dan mekanisme dari MaskDINO?

- 2) Bagaimana pengaruh modifikasi arsitektur *Real-Time Detection Transformer* (RT-DETR) terhadap performa *object detection* dan *instance segmentation* dalam aspek akurasi (*Average Precision*)?
- 3) Bagaimana trade-off antara *Average Precision* (AP) dan *Frame Per Second* (FPS)?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dibuat, tujuan dilakukannya penelitian ini disebutkan sebagai berikut:

- 1) Modifikasi model RT-DETR dengan mengintegrasikan *mask branch*, *hybrid matching*, dan mekanisme *denoising* dari MaskDINO.
- 2) Mengevaluasi performa model hasil modifikasi dalam menjalankan *instance segmentation* dan *object detection*.
- 3) Mendapatkan trade-off antara akurasi (*Average Precision*) dan efisiensi komputasi (*Frames Per Second*/FPS) yang optimal.

1.4 Manfaat Penelitian

Adapun manfaat yang diharapkan dari peneliti sebagai berikut:

- 1) Bagi Peneliti
Peneliti diharapkan mampu memperoleh pengetahuan baru mengenai pengembangan dan modifikasi arsitektur model transformer untuk menyelesaikan permasalahan *object detection* dan *instance segmentation*.
- 2) Bagi Pihak Lain
Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem vision real-time yang menyeimbangkan akurasi dan efisiensi untuk kebutuhan praktis seperti *surveillance*, kendaraan otonom, robotika, dan aplikasi industri lainnya.

1.5 Batasan Masalah

Batasan masalah ditentukan agar penelitian yang dilakukan fokus terhadap bidang yang diteliti dan disebutkan sebagai berikut:

- 1) *Dataset* yang digunakan untuk *training* dan evaluasi pada penelitian ini adalah COCO 2017.

- 2) Model *backbone* yang digunakan adalah ResNet-50, tidak dilakukan eksplorasi terhadap *backbone* yang lebih besar atau kecil.
- 3) Evaluasi hanya mencakup dua aspek utama: *Average Precision* (AP) untuk akurasi dan *Frames Per Second* (FPS) untuk efisiensi komputasi.
- 4) Lingkup modifikasi hanya mencakup penambahan *mask branch*, *hybrid matching*, *denoising*, dan beberapa optimasi arsitektur tidak termasuk optimasi hyperparameter secara menyeluruh ataupun augmentasi data.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini terdiri dari lima bab, dengan struktur sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini berisi latar belakang topik *instance segmentation* dan *object detection*, rumusan masalah yang ingin diselesaikan, tujuan penelitian yang ingin dicapai, manfaat dari penelitian, batasan ruang lingkup penelitian, serta sistematika penulisan skripsi secara keseluruhan.

BAB II KAJIAN PUSTAKA

Pada bab ini memuat kajian pustaka terkait teknologi *object detection* dan *instance segmentation*, serta pendekatan berbasis *transformer* yang digunakan dalam penelitian ini. Penjelasan dimulai dari arsitektur *RT-DETR* sebagai *baseline model*, lalu dijabarkan pula metode dan komponen utama dari *MaskDINO* seperti *denoising training*, *hybrid matching*, dan *mask head*. Kemudian dibahas pula metrik evaluasi yang digunakan seperti *Average Precision* (AP) dan *Frames Per Second* (FPS), serta penelitian terdahulu yang relevan.

BAB III METODE PENELITIAN

Bab ini menjelaskan secara detail tahapan dalam proses penelitian. Dimulai dari perumusan masalah, desain dan modifikasi arsitektur *RT-DETR*, proses integrasi cabang segmentasi dari *MaskDINO*, pengumpulan dan pemrosesan dataset *COCO 2017*, implementasi skenario *training* dan pengujian, hingga metode evaluasi performa model terhadap aspek presisi dan kecepatan.

BAB IV TEMUAN DAN PEMBAHASAN

Pada bab ini dibahas hasil eksperimen yang telah dilakukan. Meliputi proses *training* model, evaluasi hasil *object detection* dan *instance segmentation*, perbandingan kinerja dengan model-model lain seperti *Mask DINO*, *QueryInst*, dan *Mask R-CNN*, serta analisis *trade-off* antara nilai *Average Precision* dan *FPS* pada model yang diajukan.

BAB V KESIMPULAN DAN SARAN

Bab ini menyimpulkan hasil penelitian yang telah dilakukan mengenai modifikasi arsitektur *RT-DETR* untuk mendukung *instance segmentation* secara *real-time*. Disampaikan pula saran-saran untuk pengembangan dan penelitian lanjutan yang dapat memperbaiki atau memperluas hasil penelitian ini.