

BAB 1 PENDAHULUAN

1.1 Latar Belakang Masalah

Deteksi manusia merupakan salah satu tugas penting dalam sistem cerdas berbasis *computer vision* [1]. Kemampuan untuk mendeteksi manusia secara akurat diperlukan dalam berbagai aplikasi, seperti *person counting and tracking* [2–4], *crowd detection* [5–9], serta *human activity recognition* [10, 11]. Dalam kondisi nyata, tugas ini dihadapkan pada berbagai tantangan, seperti *occlusion*, variasi pose, penampilan, ukuran, skala, orientasi, serta perubahan pencahayaan dan sudut pandang kamera, sehingga kompleksitasnya lebih tinggi dibandingkan deteksi objek umum [1, 12].

Tantangan tersebut semakin signifikan pada kondisi kerumunan padat. Pada kondisi tersebut, individu muncul dalam jumlah besar dan saling berdekatan [13, 14], sehingga menyebabkan tumpang tindih (*overlapping*) yang mempersulit pemisahan antar objek dan berdampak langsung pada penurunan nilai *precision* dan *recall* [15, 16]. Kondisi ini mendorong kebutuhan akan metode deteksi yang mampu bekerja secara cepat dan akurat, khususnya untuk kebutuhan *real-time* [12, 17].

Perkembangan *deep learning* telah meningkatkan performa deteksi objek secara signifikan, salah satunya melalui keluarga model *You Only Look Once* (YOLO) yang mampu melakukan deteksi dalam satu tahap proses inferensi [12]. Arsitektur YOLO terus berkembang untuk meningkatkan akurasi, efisiensi komputasi, dan kecepatan inferensi [18]. Salah satu model generasi terbaru, YOLOv11, menunjukkan peningkatan nilai *mean Average Precision* (mAP) serta waktu inferensi yang lebih singkat dibandingkan generasi sebelumnya [19]. Selain itu, YOLOv11 menghasilkan prediksi yang lebih stabil dengan redundansi lebih rendah dibanding YOLOv8 serta kemampuan adaptasi yang lebih baik pada skenario *transfer learning* dan *domain shift*, terutama untuk objek berukuran kecil dan kondisi kerumunan padat [19].

Dalam pengembangan model deteksi berbasis YOLO, dataset Microsoft COCO (*Common Objects in Context*) [20] lazim digunakan sebagai dasar pelatihan awal (*pretrained model*) karena memiliki variasi konteks visual yang luas serta mencakup berbagai ukuran objek [20–22]. Penggunaan *pretrained weights* berbasis COCO memungkinkan model mempelajari representasi fitur manusia dalam

berbagai kondisi lingkungan dan sering dikombinasikan dengan *transfer learning* untuk penyesuaian pada domain yang lebih spesifik [23–25]. Pendekatan ini terbukti dapat mempercepat konvergensi pelatihan serta meningkatkan mAP pada kondisi data berlabel terbatas [26, 27]. Salah satu strategi yang umum diterapkan dalam proses *fine-tuning* adalah pembekuan lapisan awal (*layer freezing*), yakni mempertahankan bobot lapisan yang telah mempelajari fitur umum dari data sumber agar adaptasi dapat difokuskan pada lapisan yang lebih spesifik terhadap domain target [27]. Namun demikian, sebagai *general-purpose dataset*, COCO tidak dirancang khusus untuk skenario kepadatan tinggi, sehingga model yang dilatih di atasnya berpotensi mengalami penurunan kinerja pada *crowded scene* yang kompleks akibat perbedaan distribusi data (*distribution shift*) [28–30].

Ketersediaan dataset yang sesuai menjadi faktor penting dalam pelatihan model deteksi manusia pada kerumunan padat. Dataset CrowdHuman [31] dirancang untuk mendukung deteksi manusia dengan tingkat *occlusion* yang tinggi dan jumlah individu yang besar dalam satu citra, sedangkan WiderPerson [32] menyediakan variasi skenario dengan tingkat kepadatan dan kondisi lingkungan yang beragam. CrowdHuman menyediakan tiga jenis anotasi, yaitu *visible box* (vbox), *full body box* (fbox), dan *head box* (hbox) [31], sementara WiderPerson menyediakan anotasi vbox dan fbox [32]. Anotasi fbox dinilai lebih representatif karena mampu menggambarkan keseluruhan tubuh manusia secara konsisten, termasuk bagian yang terhalang, sehingga lebih sesuai untuk pelatihan model pada kondisi kerumunan padat dibandingkan vbox yang cenderung menghasilkan *bounding box* kecil dan tidak konsisten [33]. Selain itu, penggunaan fbox sebagai fokus target deteksi juga telah lazim diimplementasikan pada berbagai penelitian terdahulu yang berfokus pada pengembangan model deteksi objek di lingkungan padat [34, 35].

Sejalan dengan perkembangan dataset dan arsitektur yang telah diuraikan, kajian pada [29] merupakan penelitian terdahulu yang relevan dalam konteks deteksi manusia pada kerumunan padat berbasis YOLOv11, karena secara langsung mengevaluasi strategi *fine-tuning* YOLOv11 berbasis deteksi kepala pada kondisi kerumunan padat melalui perbandingan empat skenario, yaitu *baseline pretrained COCO* tanpa adaptasi, *full fine-tuning* pada resolusi 960 piksel, *partial backbone freezing* pada resolusi 960 piksel, serta *partial freezing* pada resolusi 640 piksel. Dari sisi *transfer learning* dan *fine-tuning*, ditemukan bahwa *full fine-tuning* menghasilkan akurasi tertinggi dengan mAP@0.5 sebesar 0,682 dan MAE terendah sebesar 37,8, sementara strategi *partial backbone freezing* pada

resolusi yang sama tetap kompetitif dengan mAP@0.5 sebesar 0,667 dan MAE sebesar 39,6. Temuan tersebut mengindikasikan bahwa pembekuan lapisan awal *backbone* mampu mempertahankan kemampuan ekstraksi fitur generik dari data sumber dengan hanya penurunan akurasi yang marginal dibandingkan *full fine-tuning*. Secara keseluruhan, kedua strategi tersebut menunjukkan peningkatan yang sangat signifikan dibandingkan *baseline pretrained* COCO yang hanya mencapai mAP@0.5 sebesar 0,017 dan MAE sebesar 100,3. Lebih lanjut, perbandingan antar skenario mengungkap adanya *trade-off* antara *precision* dan *recall*, di mana model yang dilatih pada resolusi lebih rendah menghasilkan *precision* tertinggi sebesar 0,828 namun disertai penurunan *recall* yang signifikan ke angka 0,460, mengindikasikan bahwa model cenderung konservatif dan melewatkan sebagian besar individu yang terhalang. Kajian tersebut menyimpulkan bahwa *partial backbone freezing* merupakan strategi yang paling seimbang antara sensitivitas deteksi dan efisiensi pelatihan, meskipun pengaruhnya terhadap kemampuan generalisasi model pada variasi distribusi data kerumunan padat belum dikaji lebih lanjut.

Selain kajian tersebut, sejumlah penelitian terdahulu juga telah mengeksplorasi deteksi manusia pada kerumunan padat dari sudut pandang yang beragam. Pendekatan berbasis modifikasi arsitektur YOLO ditempuh melalui pengembangan model dengan strategi fusi multi-skala dan lapisan deteksi khusus untuk objek kecil yang terhalang pada YOLOv7 [36], serta model ringan berbasis YOLOv8n yang dievaluasi pada dataset CrowdHuman [31] dan WiderPerson [32, 37]. Penerapan YOLOv5 pada [9] mengidentifikasi keterbatasan model pada kondisi *occlusion* berat, sementara sistem deteksi dan penghitungan manusia berbasis YOLOv11 pada [3] menunjukkan performa kompetitif meski diuji pada kondisi kepadatan yang relatif rendah. Dalam cakupan yang lebih luas, strategi *transfer learning* dan *fine-tuning* pada arsitektur YOLO telah pula diterapkan pada domain spesifik lainnya [23, 24], dan strategi pembekuan lapisan awal dalam proses tersebut terbukti efektif mempertahankan representasi fitur generik dari data sumber sembari memfokuskan adaptasi pada lapisan yang lebih dalam [27].

Meski berbagai upaya telah dilakukan, beberapa aspek mendasar masih belum dikaji secara komprehensif. Efektivitas *pretrained weights* COCO sebagai inisialisasi *transfer learning* pada arsitektur YOLOv11 untuk deteksi kelas *person* di kondisi kerumunan padat belum dievaluasi secara empiris, mengingat perbedaan distribusi data antara dataset sumber yang bersifat umum dan dataset target

yang berdensitas tinggi berpotensi membatasi kapasitas adaptasi model secara signifikan. Selain itu, kemampuan generalisasi model yang diperoleh melalui *transfer learning* pada variasi distribusi data antar dataset kerumunan padat, yang mencakup perbedaan tingkat kepadatan, pola *occlusion*, dan karakteristik anotasi *full body*, juga belum dikaji secara sistematis.

Berdasarkan kesenjangan tersebut, dilakukan kajian empiris terhadap penerapan *transfer learning* pada model YOLOv11 untuk deteksi manusia pada kondisi kerumunan padat, dengan membandingkan tiga strategi pelatihan secara terkontrol, yaitu *training from scratch*, *full fine-tuning* berbasis *pretrained weights* COCO, dan *partial fine-tuning* dengan strategi *layer freezing*, tanpa modifikasi arsitektur maupun resolusi input, pada dataset kerumunan padat CrowdHuman [31] dan WiderPerson [32] dengan anotasi *full body box*.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut.

1. Bagaimana mengimplementasikan transfer learning pada model YOLOv11 untuk deteksi manusia pada kondisi kerumunan padat?
2. Berapa nilai mean Average Precision (mAP), precision, dan recall yang dihasilkan oleh model deteksi manusia berbasis transfer learning menggunakan YOLOv11 pada kondisi kerumunan padat?

1.3 Batasan Permasalahan

Batasan permasalahan dalam penelitian ini adalah sebagai berikut.

1. Model YOLOv11 yang digunakan terbatas pada versi *nano*.
2. Transfer learning memanfaatkan *pretrained weights* yolov11n.pt dari dataset COCO.
3. Dataset target untuk pelatihan dan evaluasi adalah CrowdHuman dan WiderPerson, dengan fokus anotasi *full body box*.
4. Evaluasi *in-domain* dilakukan pada *validation set* dari kedua dataset, karena *test set* tidak memiliki anotasi *ground truth*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Mengimplementasikan transfer learning pada model YOLOv11 untuk deteksi manusia pada kondisi kerumunan padat.
2. Mengukur nilai mean Average Precision (mAP), precision, dan recall dari model deteksi manusia berbasis transfer learning menggunakan YOLOv11 pada kondisi kerumunan padat.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut.

1. Memberikan bukti empiris mengenai efektivitas transfer learning dalam meningkatkan performa model YOLOv11 untuk deteksi manusia pada kondisi kerumunan padat.
2. Mendukung pengembangan sistem deteksi manusia yang lebih akurat untuk berbagai aplikasi, seperti sistem keamanan, pemantauan kerumunan, dan analisis perilaku manusia.
3. Membantu meningkatkan akurasi perhitungan jumlah orang (*person counting*) di area padat, seperti stadion, stasiun kereta, bandara, atau pusat perbelanjaan.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut.

1. Bab 1 PENDAHULUAN
Bab ini terdiri dari beberapa bagian, yaitu latar belakang masalah terkait deteksi manusia pada kondisi kerumunan padat, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.
2. Bab 2 LANDASAN TEORI
Bab ini berisikan landasan teori terkait dengan judul penelitian. Landasan teori yang dijelaskan pada bab ini adalah deteksi objek, arsitektur dasar

YOLO dan perkembangannya, arsitektur YOLOv11, *Transfer Learning* dan strateginya, generalisasi lintas domain (*Cross-Domain Generalization*), deteksi manusia pada kerumunan padat, oklusi, *Non-Maximum Suppression* (NMS), dataset COCO, CrowdHuman, dan WiderPerson, serta metrik evaluasi yang terdiri dari *Intersection over Union* (IoU), *Confusion Matrix*, *Precision* dan *Recall*, serta *Average Precision* (AP) dan *Mean Average Precision* (mAP).

3. Bab 3 METODOLOGI PENELITIAN

Bab ini berisikan metodologi penelitian yang digunakan pada penelitian ini. Metodologi penelitian ini terdiri dari alur penelitian, studi literatur, dataset penelitian yang mencakup pemilihan dataset dan pemilihan jenis anotasi *bounding box*, praproses data, pelatihan model yang terdiri dari *train from scratch*, *full fine-tuning*, dan *partial fine-tuning*, serta evaluasi model dan diakhiri dengan dokumentasi.

4. Bab 4 HASIL DAN DISKUSI

Bab ini berisikan hasil dan pembahasan dari ketiga strategi *transfer learning* (*train from scratch*, *full fine-tuning*, dan *partial fine-tuning*) yang telah diuji pada dataset CrowdHuman dan WiderPerson, mencakup evaluasi *in-domain* dan *cross-domain*, serta eksperimen tambahan, sesuai dengan metode penelitian pada bab 3.

5. Bab 5 SIMPULAN DAN SARAN

Bab ini berisi kesimpulan mengenai hasil *transfer learning* pada model YOLOv11 untuk deteksi manusia pada kondisi kerumunan padat, beserta saran yang dapat digunakan untuk penelitian selanjutnya.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A