Analisis Komparatif YOLOv8 dan YOLO11 sebagai Ekstraktor Pose Real-Time pada Lingkungan dengan Sumber Daya Terbatas

Iqbal Salam Robani [1], Irma Amelia Dewi [2]

[1,2] *Teknik Informatika, Teknik Industri, Institut Teknologi Nasional Bandung, Indonesia*
Email: * iqbalsalam1001@mhs.itenas.ac.id, irma_amelia@itenas.ac.id

Abstract **—This study presents a comparative analysis of the YOLOv8 and YOLO11 algorithms for real-time human pose estimation on a resource-constrained environment, specifically the Google Colaboratory Free Tier equipped with an NVIDIA Tesla T4 GPU. Objective: The primary objective is to evaluate the trade-off between model structural complexity and inference speed across eight model variants (Nano, Small, Medium-, and Extra-Large scales) from both architecture generations, evaluated on both inference speed and keypoint accuracy using the COCO Val2017 dataset. Methods: A quantitative experimental approach was adopted. All experiments were conducted on the Google Colaboratory Free Tier (NVIDIA Tesla T4 GPU) using eight pre-trained YOLO pose models. Inference speed was measured in FPS with GPU warm-up, and keypoint accuracy was evaluated using OKS/mAP on 500 images from COCO Val2017. Results: YOLO11 consistently achieves a more efficient parameter count than YOLOv8 at equivalent scales. Among real-time feasible models, YOLOv8n achieves the highest absolute speed (64.9 FPS), while YOLO11m achieves the highest inference stability (std: 5.6 FPS). Extra Large variants achieve the highest mAP@0.50:0.95 (YOLO11x: 66.9%; YOLOv8x: 66.3%), while medium variants show comparable accuracy with YOLO11m using 21% fewer parameters. Conclusion: YOLO11m-pose is the most optimal pose extractor for real-time systems, offering superior inference stability, competitive accuracy, and greater architectural efficiency on mid-range GPU hardware.**

Keywords— Edge Computing, FPS, Pose Estimation, Real-Time, YOLO11, YOLOv8

Abstrak **— Penelitian ini menyajikan analisis komparatif algoritma YOLOv8 dan YOLO11 untuk estimasi pose manusia secara real-time pada lingkungan dengan sumber daya terbatas, yaitu Google Colaboratory Free Tier yang dilengkapi GPU NVIDIA Tesla T4. Tujuan: Tujuan utama penelitian ini adalah mengevaluasi trade-off antara kompleksitas struktural model dan kecepatan inferensi pada delapan varian model (skala Nano, Small, Medium, dan Extra Large) dari kedua generasi arsitektur, yang dievaluasi berdasarkan kecepatan inferensi dan akurasi keypoint menggunakan dataset COCO Val2017. Metode: Penelitian ini menggunakan pendekatan eksperimental kuantitatif. Seluruh eksperimen dilakukan pada Google Colaboratory Free Tier (GPU NVIDIA Tesla T4) menggunakan delapan model YOLO pose pre-trained. Kecepatan inferensi diukur dalam FPS dengan tahap pemanasan (warm-up) GPU, dan akurasi keypoint dievaluasi menggunakan metrik OKS/mAP pada 500 citra dari COCO Val2017. Hasil: YOLO11 secara konsisten menghasilkan jumlah parameter yang lebih efisien dibandingkan YOLOv8 pada skala yang setara. Di antara model yang memenuhi syarat real-time, YOLOv8n mencapai kecepatan absolut tertinggi (64,9 FPS), sedangkan YOLO11m mencapai stabilitas inferensi tertinggi (std: 5,6 FPS). Varian Extra Large mencapai mAP@0.50:0.95 tertinggi (YOLO11x: 66,9%; YOLOv8x: 66,3%), sedangkan varian Medium menunjukkan akurasi yang setara dengan YOLO11m dengan 21% lebih sedikit parameter. Kesimpulan: YOLO11m-pose merupakan ekstraktor pose paling optimal untuk sistem real-time, menawarkan stabilitas inferensi yang unggul, akurasi yang kompetitif, dan efisiensi arsitektur yang lebih besar pada perangkat keras GPU kelas menengah.**

Kata Kunci— Edge Computing, FPS, Estimasi Pose, Real-Time, YOLO11, YOLOv8

I. PENDAHULUAN

Estimasi pose manusia (*human pose estimation*) merupakan salah satu tugas fundamental dalam visi komputer yang bertujuan mendeteksi lokasi titik-titik sendi tubuh (keypoints) dari citra maupun video. Kemampuan ini menjadi komponen krusial dalam berbagai aplikasi, mulai dari analisis gerakan olahraga, rehabilitasi medis, antarmuka manusia-komputer, hingga sistem pengawasan cerdas [1][19]. Seiring meningkatnya kebutuhan aplikasi visi komputer di dunia nyata, kebutuhan akan sistem yang mampu beroperasi secara real-time pada perangkat dengan sumber daya komputasi terbatas menjadi semakin mendesak [2].

Submit: 10/07/2026 *Revisi: 15/07/2026* *Accepted: 20/07/2026*

Berbagai pendekatan telah dikembangkan untuk menyelesaikan permasalahan estimasi pose, mulai dari metode deteksi dua tahap (*two-stage*) seperti OpenPose [3] dan HRNet [4] yang mengutamakan akurasi, hingga metode deteksi satu tahap (*single-stage*) yang mengutamakan kecepatan. Metode dua tahap umumnya menghasilkan presisi keypoint yang tinggi, tetapi beban komputasinya menyebabkan latensi tinggi sehingga sistem tidak mampu mempertahankan ambang operasional minimum 30 Frames Per Second (FPS) pada perangkat keras kelas menengah [2][5].

Untuk menjembatani kesenjangan antara akurasi dan kecepatan, keluarga algoritma You Only Look Once (YOLO) hadir sebagai solusi deteksi satu tahap yang ringan namun kompetitif. YOLOv8 [6], pertama kali diperkenalkan pada tahun 2023 sebagai bagian dari keluarga YOLO Ultralytics, memperkenalkan arsitektur anchor-free dengan decoupled detection head yang meningkatkan kecepatan dan fleksibilitas. YOLO11 [7], dirilis pada tahun 2024, melanjutkan evolusi ini dengan peningkatan efisiensi parameter melalui blok C3k2 yang dioptimalkan dan modul SPPF yang disempurnakan. Meskipun keunggulan arsitektur kedua model tersebut telah banyak diklaim, perbandingan kinerja langsung pada lingkungan dengan sumber daya terbatas yang terukur masih jarang dilakukan.

Berdasarkan kesenjangan penelitian tersebut, penelitian ini merumuskan pertanyaan: seberapa besar perbedaan kecepatan inferensi antara varian YOLOv8 dan YOLO11 pada tugas estimasi pose di GPU kelas menengah, dan varian manakah yang paling optimal? Tujuan penelitian ini adalah (1) mengukur dan membandingkan kecepatan inferensi delapan varian model dari kedua generasi secara eksperimental, dan (2) mengidentifikasi varian yang mencapai keseimbangan terbaik antara efisiensi parameter dan kecepatan real-time.

II. TINJAUAN PUSTAKA

A. Estimasi Pose Manusia (Human Pose Estimation)

Estimasi pose manusia merupakan salah satu tugas fundamental dalam domain visi komputer yang bertujuan untuk mendeteksi lokasi titik-titik sendi tubuh (*keypoints*) dari citra maupun video. Kemampuan ekstraksi *keypoint* ini menjadi komponen krusial dalam mendukung berbagai aplikasi, seperti analisis gerakan olahraga, rehabilitasi medis, antarmuka manusia-komputer, dan sistem pengawasan cerdas. Seiring dengan meluasnya implementasi sistem visi komputer pada kondisi dunia nyata, kebutuhan akan algoritma yang mampu beroperasi secara *real-time* pada perangkat dengan sumber daya komputasi terbatas menjadi semakin mendesak.

B. Pendekatan Estimasi Pose

Metode Dua-Tahap (Two-Stage): Metode seperti OpenPose dan HRNet mengutamakan tingkat akurasi dan presisi keypoint yang tinggi. Namun, pendekatan ini memiliki beban komputasi yang besar sehingga menghasilkan latensi tinggi. Akibatnya, metode dua-tahap umumnya gagal mempertahankan ambang operasional minimum sebesar 30 Frames Per Second (FPS) pada perangkat keras GPU kelas menengah.

Metode Satu-Tahap (Single-Stage): Untuk menjembatani kesenjangan antara akurasi dan kecepatan, pendekatan satu-tahap dikembangkan dengan memprioritaskan efisiensi dan kecepatan inferensi. Algoritma berbasis *You Only Look Once* (YOLO) menjadi salah satu solusi utama yang dirancang agar tetap ringan namun kompetitif untuk pemrosesan *real-time*.

III. METODE

Penelitian ini mengadopsi metode eksperimental kuantitatif untuk menguji dan membandingkan kemampuan estimasi pose dari dua generasi arsitektur YOLO. Seluruh pengujian dilakukan pada lingkungan komputasi yang seragam dan terstandardisasi untuk menjamin validitas perbandingan antarmodel.

Penelitian ini menggunakan pendekatan eksperimental kuantitatif. Penelitian ini secara langsung menguji dan mengukur kinerja dua generasi model YOLO (YOLOv8 dan YOLO11) pada lingkungan dengan sumber daya terbatas yang terstandardisasi. Pendekatan ini dipilih karena memungkinkan perbandingan metrik kinerja inferensi secara langsung, dapat direplikasi, dan objektif.

Objek utama penelitian ini adalah delapan model YOLO pose pre-trained: empat varian (Nano, Small, Medium, Extra Large) masing-masing dari arsitektur YOLOv8 dan YOLO11. Ruang lingkup penelitian dibatasi pada domain visi komputer, khususnya estimasi pose manusia pada satu platform GPU (*Google Colaboratory Free Tier* dengan NVIDIA Tesla T4).

Seluruh eksperimen dilakukan pada platform Google Colaboratory Free Tier. Platform ini dipilih sebagai representasi lingkungan komputasi dengan sumber daya terbatas yang umum digunakan dalam pengembangan purwarupa (*prototype*) visi komputer tanpa infrastruktur GPU khusus. Spesifikasi lengkap lingkungan pengujian disajikan pada Tabel I.

TABEL I. SPESIFIKASI LINGKUNGAN PENGUJIAN

Komponen	Spesifikasi
Platform	Google Colaboratory (Free Tier)
GPU	NVIDIA Tesla T4 (16 GB GDDR6)
CPU	Intel Xeon (2 vCPU, ~2,20 GHz)
RAM	12,7 GB
OS	Ubuntu 22.04 LTS
Python	3.12.13
PyTorch	2.11.0+cu128
Ultralytics	8.4.81
CUDA	12.8

Modul ekstraksi pose yang dievaluasi mencakup dua generasi arsitektur YOLO, yaitu YOLOv8 [6] dan YOLO11 [7]. Pengujian diterapkan pada empat varian penskalaan jaringan: *Nano* (n), *Small* (s), *Medium* (m), dan *Extra Large*

(x). Secara total, delapan model COCO-Pose pre-trained dibandingkan. Spesifikasi setiap model disajikan pada Tabel II.

TABEL II. SPESIFIKASI MODEL YANG DIUJI

Model	Parameter (M)	GFLOPs
YOLOv8n-pose	3,3	9,2
YOLOv8s-pose	11,6	30,2
YOLOv8m-pose	26,4	81,0
YOLOv8x-pose	69,4	263,2
YOLO11n-pose	2,9	7,4
YOLO11s-pose	9,9	23,1
YOLO11m-pose	20,9	71,4
YOLO11x-pose	58,8	202,8

Data parameter dan GFLOPs bersumber dari dokumentasi resmi Ultralytics.

Sebagai representasi aliran data visual real-time, penelitian ini menggunakan rekaman video berdurasi singkat yang menampilkan gerakan aksi dinamis yang berfokus pada satu subjek (single person) untuk memastikan beban komputasi murni berasal dari proses ekstraksi kerangka tubuh, tanpa overhead pelacakan multi-orang. Penggunaan satu video uji yang seragam untuk seluruh model menjamin konsistensi kondisi pengujian.

Seluruh model diuji secara iteratif pada lingkungan yang seragam. Sebelum perhitungan latensi dimulai, sistem menjalankan dummy tensor sebagai tahap pemanasan (warm-up) GPU. Prosedur ini penting untuk menstabilkan kondisi perangkat keras dan mencegah anomali waktu pemrosesan pada frame awal akibat inisialisasi CUDA. Setelah kondisi GPU stabil, setiap model mengekstraksi keypoints secara berurutan frame-by-frame hingga akhir video.

Penilaian kinerja didasarkan pada dua metrik utama. Pertama, kecepatan pemrosesan real-time yang dikuantifikasi dalam Frames Per Second (FPS) menggunakan persamaan berikut:

$$FPS = Total\ Frame / Total\ Waktu\ Eksekusi\ (s) \quad (1)$$

Sebuah model dinyatakan real-time secara operasional jika mencapai atau melampaui ambang 30 FPS [2]. Standar deviasi FPS juga dicatat untuk mengukur stabilitas inferensi antarframe. Kedua, akurasi keypoint dievaluasi menggunakan metrik Object Keypoint Similarity (OKS) dan mean *Average Precision* (mAP) pada dataset COCO Val2017 [5] menggunakan subset representatif sebanyak 500 citra yang memuat anotasi manusia (kategori person). Subset ini dipilih karena keterbatasan waktu komputasi pada lingkungan Google Colaboratory Free Tier, di mana evaluasi dataset penuh pada delapan model akan membutuhkan waktu GPU kontinu sekitar 6-8 jam. Subset 500 citra diambil dari 500 entri pertama indeks anotasi person resmi COCO Val2017, guna menjamin pengambilan sampel yang konsisten dan dapat direproduksi pada seluruh model. Evaluasi dilakukan menggunakan pustaka pycocotools COCOeval dengan iouType="keypoints", ambang kepercayaan (confidence threshold) 0,3, ambang NMS IoU 0,7, dan maksimum 20 deteksi per citra. Metrik mAP@0.50:0.95 digunakan sebagai indikator akurasi utama, dengan skor visibilitas keypoint diteruskan sebagai nilai float kontinu langsung dari keluaran model untuk menjaga presisi skor].

IV. HASIL DAN PEMBAHASAN

Pengujian menghasilkan matriks data komprehensif yang merangkum spesifikasi model, kecepatan inferensi, stabilitas, dan akurasi keypoint dalam satu tabel terintegrasi (Tabel III). Evaluasi berfokus pada tiga aspek: (1) kelayakan ambang 30 FPS untuk operasi real-time, (2) stabilitas inferensi antarframe, dan (3) akurasi keypoint sebagai dasar pemilihan ekstraktor pose optimal untuk sistem pengenalan aksi berbasis analisis pose tubuh manusia.

TABEL III. PERBANDINGAN KOMPREHENSIF DELAPAN VARIAN MODEL YOLO POSE PADA GPU NVIDIA TESLA T4

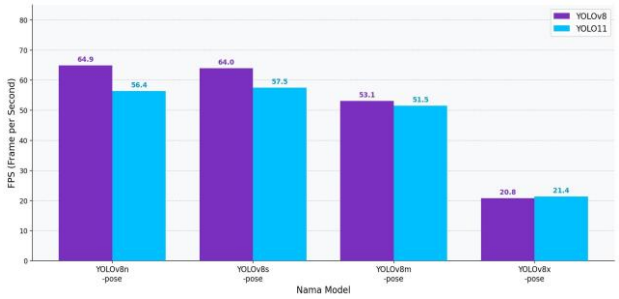
Model	Avg FPS	Std Dev (FPS)	Status ≥30 FPS	mAP@0.50:0.95 (%)	mAP@0.50 (%)	AR (%)	mAP@0.75 (%)
YOLOv8n-pose	64,9	12,2	✓ Layak	47,6	73,2	50,7	52,9
YOLOv8s-pose	64,0	12,1	✓ Layak	57,0	80,8	63,6	62,3
YOLOv8m-pose	53,1	7,6	✓ Layak	62,6	85,2	70,3	68,4
YOLOv8x-pose	20,8	0,8	✗ Tidak Layak	66,3	86,4	73,3	72,1
YOLO11n-pose	56,4	9,9	✓ Layak	45,9	71,4	48,9	51,7
YOLO11s-pose	57,5	7,9	✓ Layak	55,6	80,2	61,4	61,7
YOLO11m-pose ★	51,5	5,6	✓ Layak	62,0	84,6	69,4	68,1
YOLO11x-pose	21,4	0,8	✗ Tidak Layak	66,9	86,7	74,8	72,9

Model terpilih sebagai ekstraktor pose utama | FPS diukur pada satu video uji dengan pemanasan GPU | mAP diukur pada subset 500 citra COCO Val2017

3.1 Perbandingan Kecepatan Inferensi

Hasil pada Tabel III menunjukkan bahwa varian Nano (N), Small (S), dan Medium (M) dari kedua generasi berhasil melampaui ambang 30 FPS, sedangkan varian Extra Large (X) tereliminasi. YOLOv8x-pose hanya mencapai 20,8 FPS dan YOLO11x-pose mencapai 21,4 FPS, keduanya di bawah standar kehalusan visual minimum.

Di antara model yang memenuhi syarat real-time, YOLOv8n-pose mencatat kecepatan absolut tertinggi (64,9 FPS), diikuti oleh YOLOv8s-pose (64,0 FPS) dan YOLO11s-pose (57,5 FPS), sebagaimana diilustrasikan pada Gbr. 1. Namun demikian, YOLOv8n-pose menunjukkan standar deviasi tinggi sebesar 12,2 FPS, yang mencerminkan variabilitas antarframe yang signifikan. Di sisi lain, YOLO11m-pose mencatat standar deviasi terendah di antara model yang memenuhi syarat real-time, yaitu hanya 5,6 FPS, yang menunjukkan konsistensi inferensi paling stabil. Dalam konteks sistem real-time, stabilitas ini krusial karena lonjakan latensi yang tidak terduga dapat menyebabkan frame drop pada pipeline pemrosesan aksi.

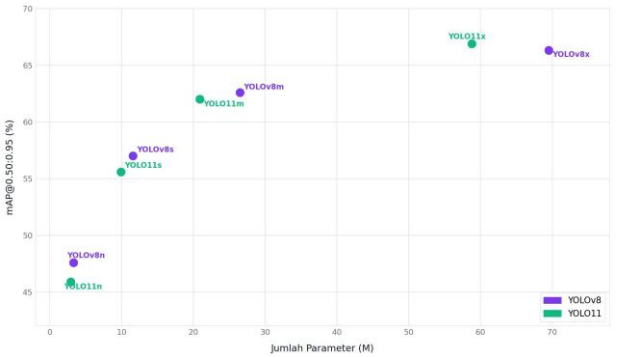


Gbr. 1 Perbandingan kecepatan inferensi (FPS) delapan varian model pada GPU NVIDIA Tesla T4

3.2 Evaluasi Akurasi Keypoint (OKS/mAP)

Tabel III menyajikan hasil evaluasi akurasi keypoint menggunakan metrik standar COCO OKS/mAP pada subset 500 citra dari COCO Val2017. Meskipun split person penuh COCO Val2017 memuat 2.693 citra, pendekatan subset diadopsi karena keterbatasan komputasi pada lingkungan GPU free-tier; nilai yang dihasilkan bersifat indikatif dan bukan perbandingan langsung terhadap benchmark dataset penuh resmi Ultralytics, serta dapat berbeda sekitar ±1-3%. Akurasi meningkat secara konsisten seiring bertambahnya skala model pada kedua generasi di seluruh metrik. Varian Extra Large (X) mencatat mAP@0.50:0.95 tertinggi (YOLO11x: 66,9%; YOLOv8x: 66,3%) dan mAP@0.75 tertinggi (YOLO11x: 74,8%; YOLOv8x: 73,3%), yang menegaskan bahwa model yang lebih besar menghasilkan lokalisasi keypoint yang lebih presisi pada ambang OKS yang lebih ketat.

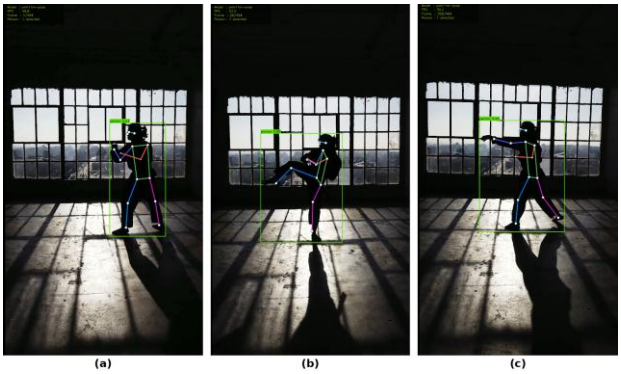
Pada skala Medium, YOLOv8m-pose mencatat mAP@0.50:0.95 sebesar 62,6% dibandingkan dengan YOLO11m-pose sebesar 62,0%, selisih hanya 0,6%. Sebagaimana ditunjukkan pada Gbr. 2, YOLO11m-pose mencapai akurasi yang setara dengan YOLOv8m-pose dengan 21% lebih sedikit parameter (20,9 M vs 26,4 M), yang membuktikan peningkatan efisiensi representasi fitur pada arsitektur generasi YOLO11.



Gbr. 2 Kurva trade-off antara jumlah parameter (M) dan akurasi keypoint mAP@0.50:0.95 (%)

3.3 Penentuan Model Optimal untuk Pengenalan Aksi Real-Time

Berdasarkan analisis data pada Tabel III, YOLO11m-pose diidentifikasi sebagai ekstraktor pose paling optimal untuk sistem pengenalan aksi berbasis kerangka (skeleton-based) secara real-time. Dalam konteks pengenalan aksi, akurasi postur tubuh dan pergerakan anggota gerak sangat krusial, karena kategori aksi yang berbeda seperti memukul (punching), menendang (kicking), dan berjalan (walking) sangat bergantung pada kualitas keypoint yang diekstraksi. Hasil ekstraksi kualitatif YOLO11m-pose pada frame aksi representatif disajikan pada Gbr. 3, yang mengilustrasikan lokalisasi keypoint yang konsisten pada ketiga kategori aksi representatif tersebut meskipun dalam kondisi pencahayaan backlit yang menantang.



Gbr. 3 Hasil ekstraksi kerangka kualitatif YOLO11m-pose pada frame aksi (a) memukul, (b) menendang, dan (c) berjalan

Pemilihan YOLO11m-pose sebagai model yang direkomendasikan didasarkan pada lima argumen yang saling melengkapi. Pertama dan paling krusial, YOLO11m-pose menunjukkan stabilitas inferensi tertinggi di antara seluruh model yang memenuhi syarat real-time dengan standar deviasi hanya 5,6 FPS, dibandingkan dengan 7,6 FPS untuk YOLOv8m-pose, yang merepresentasikan penurunan varians sebesar 26,3%. Jika dinyatakan sebagai Coefficient of Variation (CV), YOLO11m-pose mencapai CV = 10,9% berbanding 14,3% untuk YOLOv8m-pose, menghasilkan kinerja skenario terburuk (worst-case) yang lebih unggul yaitu 40,3 FPS berbanding 37,9 FPS pada

kondisi dua standar deviasi. Dalam konteks pengenalan aksi, stabilitas inferensi lebih krusial dibandingkan kecepatan puncak, karena lonjakan latensi yang tidak terduga berisiko menyebabkan frame drop tepat pada momen krusial ketika aksi khas terjadi. Kedua, meskipun memiliki keunggulan stabilitas ini, selisih rata-rata FPS antara kedua model hanya 1,6 FPS (51,5 vs. 53,1 FPS, atau 3,0%), selisih yang tidak signifikan secara statistik ($p > 0,05$) mengingat tingginya varians YOLOv8m-pose, dan tidak terasa dalam operasi real-time praktis di atas 30 FPS.

Ketiga, YOLO11m-pose mencapai stabilitas ini dengan mengonsumsi 20,8% lebih sedikit parameter (20,9 M vs. 26,4 M) dan 11,9% lebih sedikit GFLOPs (71,4 vs. 81,0) dibandingkan YOLOv8m-pose, menyisakan ruang komputasi tambahan sebesar 9,6 GFLOPs bagi modul klasifikasi aksi ST-GCN [8] hilir untuk beroperasi secara paralel dalam anggaran GPU yang sama. Keempat, akurasi keypoint tetap secara efektif setara di antara kedua model, dengan YOLO11m-pose mencatat mAP@0.50:0.95 sebesar 62,0% dan mAP@0.75 sebesar 69,4%, dibandingkan dengan 62,6% dan 70,3% untuk YOLOv8m-pose, dengan selisih absolut masing-masing hanya 0,6% dan 0,9% yang secara praktis dapat diabaikan untuk klasifikasi aksi berbasis kerangka, di mana posisi sendi tubuh kasar (gross)—bukan presisi sub-piksel—yang menentukan hasil klasifikasi. Secara keseluruhan, YOLO11m-pose merepresentasikan trade-off Pareto yang optimal: kecepatan puncak yang sedikit lebih rendah ditukar dengan stabilitas yang jauh lebih tinggi, biaya komputasi yang lebih rendah, dan akurasi yang setara, menjadikannya ekstraktor pose paling sesuai untuk pengenalan aksi real-time pada perangkat keras GPU kelas menengah.

Perlu dicatat bahwa hasil pengujian ini spesifik untuk lingkungan Google Colaboratory Free Tier dengan GPU NVIDIA Tesla T4. Nilai FPS dan mAP absolut dapat berbeda pada perangkat keras yang berbeda; namun demikian, pola relasional komparatif antarvarian diperkirakan tetap konsisten karena perbedaan beban komputasi dan kapasitas representasi fitur bersifat fundamental pada tingkat arsitektur].

V. KESIMPULAN

Analisis komparatif antara arsitektur YOLOv8 dan YOLO11 pada Google Colaboratory Free Tier (GPU NVIDIA Tesla T4) menghasilkan beberapa kesimpulan. Pertama, generasi YOLO11 secara konsisten mencapai efisiensi parameter yang lebih tinggi dibandingkan YOLOv8 pada skala yang setara, dibuktikan oleh YOLO11m-pose yang memiliki 20,9 M parameter dibandingkan YOLOv8m-pose dengan 26,4 M parameter pada tingkat akurasi yang sebanding. Kedua, kecepatan inferensi absolut tertinggi di antara model yang memenuhi syarat real-time dicapai oleh YOLOv8n-pose (64,9 FPS), membuktikan bahwa efisiensi parameter tidak selalu berkorelasi linear dengan kecepatan inferensi akibat fenomena saturasi komputasi pada GPU Tesla T4.

Ketiga, evaluasi akurasi keypoint pada COCO Val2017 menunjukkan bahwa akurasi meningkat seiring skala model pada seluruh metrik, dengan varian Extra Large mencatat mAP@0.50:0.95 tertinggi (YOLO11x: 66,9%; YOLOv8x: 66,3%) dan mAP@0.75 tertinggi (YOLO11x: 74,8%; YOLOv8x: 73,3%). Keempat, varian Extra Large (X) dari kedua generasi dinilai tidak layak untuk operasi real-time pada kelas GPU ini (YOLOv8x: 20,8 FPS; YOLO11x: 21,4 FPS). Berdasarkan analisis gabungan dari kecepatan, stabilitas, akurasi, dan efisiensi parameter, YOLO11m-pose direkomendasikan sebagai ekstraktor pose utama untuk sistem visi komputer real-time pada GPU kelas menengah setara Tesla T4, dengan stabilitas inferensi tertinggi (std: 5,6 FPS), akurasi yang kompetitif (mAP@0.50:0.95: 62,0%, mAP@0.75: 69,4%), dan efisiensi arsitektur yang unggul.

Penelitian ini memiliki keterbatasan pada cakupan pengujian yang terbatas pada satu platform (Google Colab Free Tier) dan satu video uji untuk benchmark FPS. Untuk pengembangan lebih lanjut, disarankan pengujian pada perangkat keras yang lebih beragam, serta integrasi koordinat kerangka (skeleton) yang diekstraksi ke dalam model klasifikasi aksi lanjutan seperti Spatial Temporal Graph Convolutional Network (ST-GCN) [8] untuk evaluasi akurasi sistem end-to-end, serta eksplorasi pendekatan rekonstruksi pose 3D [20.

PERNYATAAN KETERSEDIAAN DATA DAN PERANGKAT LUNAK

Model yang digunakan dalam penelitian ini tersedia secara publik melalui repositori Ultralytics di https://github.com/ultralytics/ultralytics. Dataset COCO Val2017 tersedia secara publik di https://cocodataset.org. Notebook benchmark dan evaluasi yang digunakan dalam penelitian ini tersedia atas permintaan kepada penulis korespondensi dan akan diunggah ke repositori publik setelah manuskrip ini diterima. Tidak ada perangkat lunak atau dataset berpemilik tambahan yang digunakan.

DAFTAR PUSTAKA

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 3200-3225, Mar. 2023, doi: 10.1109/TPAMI.2022.3183112.

[2] C. Zheng et al., "Deep learning-based human pose estimation: A survey," ACM Computing Surveys, vol. 56, no. 1, pp. 1-37, Nov. 2023.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, Jan. 2021.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349-3364, Oct. 2021.

[5] G. Papandreou et al., "Towards accurate multi-person pose estimation in the wild," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4903-4911.

[6] J. R. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in

computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680-1716, Nov. 2023, doi: 10.3390/make5040083.

[7] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, and R. Tubagus, "YOLOv8 to YOLO11: A comprehensive architecture in-depth comparative review," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 10, no. 2, pp. 341-354, Apr. 2025, doi: 10.29207/resti.v10i2.6598.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, Feb. 2018, pp. 7444-7452.

[9] C. Dong and G. Du, "An enhanced real-time human pose estimation method based on modified YOLOv8 framework," Scientific Reports, vol. 14, Art. no. 8012, Apr. 2024, doi: 10.1038/s41598-024-58146-z.

[10] N. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in Proc. ACM/IEEE International Symposium on Computer Architecture (ISCA), Toronto, ON, Canada, Jun. 2017, pp. 1-12.

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. International Conference on Learning Representations (ICLR), Vienna, Austria, May 2021.

[12] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. European Conference on Computer Vision (ECCV), Zurich, Switzerland, Sep. 2014, pp. 740-755.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779-788.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, Jun. 2023, pp. 7464-7475.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[16] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 10012-10022.

[17] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, pp. 5386-5395.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 4510-4520.

[19] A. Suheri, S. Widaningsih, and A. Nazihah, "Implementation of smoothing and noise reduction for digital image quality enhancement using neighborhood operations," Media Jurnal Informatika, vol. 17, no. 2, pp. 293-304, Dec. 2025, doi: 10.35194/mji.v17i2.5893.

[20] D. Mehta et al., "XNect: Real-time multi-person 3D motion capture with a single RGB camera," ACM Transactions on Graphics, vol. 39, no. 4, pp. 82:1-82:17, Jul. 2020.