Jurnal Elkolind Volume 12, Nomor 1, Mei 2025
DOI: http://dx.doi.org/10.33795/elkolind.v12i1/7495

Deteksi Kepadatan Penumpang di Stasiun Kereta Api Menggunakan Vision Transformer pada Jetson Orin Nano

Mas Nurul Achmadiah[1], Novendra Setyawan[2], Anindya Dwi Risdhayanti[3],
e-mail: masnurul@polinema.ac.id, novendra@umm.ac.id, risdhayanti@polinema.ac.id
[1,3]Jurusan Teknik Elektro, Politeknik Negeri Malang, Jalan Soekarno Hatta No.9 Malang, Indonesia
[2]Jurusan Teknik Elektro, Universitas Muhammadiyah Malang, Jl. Raya Tlogomas No.246 Malang, Indonesia
[1,2]Departement Electro-Optical, National Formosa University, Huwei Township, Yunlin County, Taiwan

Informasi Artikel

Riwayat Artikel

Diterima 2 Mei 2025
Direvisi 20 Mei 2025
Diterbitkan 31 Mei 2025

Kata kunci:

Deteksi Objek
Vision Transformer
Jetson Orin Nano
Komputasi Tepi
Frame difference

ABSTRAK

Penelitian ini mengusulkan implementasi sistem deteksi kepadatan penumpang dengan menggabungkan metode *frame difference* dan model Vision Transformer (ViT-Base). Metode *frame difference* digunakan sebagai tahap awal untuk mendeteksi ada atau tidaknya perbedaan antar frame, sebelum adanya proses lebih lanjut, sehingga komputasi menjadi lebih efisien karna tidak dijalankan sepanjang waktu. Model ViT-Base yang digunakan telah dilatih awal dengan dataset ImageNet-21K dan diimplementasikan pada perangkat Jetson Orin Nano. Evaluasi dilakukan berdasarkan empat parameter utama: akurasi, latensi, konsumsi energi, dan efisiensi komputasi. Hasil pengujian menunjukkan sistem mencapai akurasi 91,17%, latensi 46,59 ms, konsumsi energi $0{,}1332\ joule/frame$, dan efisiensi $0{,}171\ \%/msW$. Hasil penelitian ini menunjukkan bahwa integrasi *frame difference* dengan ViT-Base mampu menghasilkan akurasi yang tinggi, latensi yang rendah, konsumsi energi yang rendah juga efisiensi komputasi yang relative rendah. Pendekatan ini layak diterapkan dalam sistem pemantauan kepadatan penumpang secara real-time, khususnya pada lingkungan transportasi publik yang menuntut akurasi tinggi, waktu respons cepat, dan konsumsi daya rendah.

Keywords:

Object Detection
Vision Transformer
Edge Computing
Frame Difference

ABSTRACT

This study proposes the implementation of a passenger density detection system by combining the frame difference method with the Vision Transformer (ViT-Base) model. The frame difference method is employed as an initial stage to detect changes between frames, allowing the system to skip further processing when no significant difference is present. This improves computational efficiency, as the detection model does not run continuously. The ViT-Base model used in this research was pre-trained on the ImageNet-21K dataset and deployed on the Jetson Orin Nano edge device. The system was evaluated using four key performance metrics: accuracy, latency, energy consumption, and computational efficiency. Experimental results demonstrate that the system achieved 91.17% *accuracy, an average latency of* $46.59\ ms$*, energy consumption of* $0.1332\ joules/frame$*, and computational efficiency of* $0.171\ \%/msW$*. These findings indicate that integrating the frame difference method with ViT-Base enables the system to achieve high accuracy, low latency, and low power consumption with relatively efficient computation. This approach is suitable for real-time passenger density monitoring, especially in public transportation environments that demand high accuracy, fast response times, and energy-efficient performance..*

Penulis Korespondensi:
Mas Nurul Achmadiah,

p-ISSN: 2356-0533; e-ISSN: 2355-9195

Jurusan Teknik Elektro,
Politeknik Negeri Malang,
Jalan Soekarno Hatta No.9 Malang, Indonesia
Email: masnurul@polinema.ac.id
Nomor HP/WA aktif: +62 81217807060

1. PENDAHULUAN

Pada stasiun kereta api, kepadatan penumpang yang tinggi sering kali menyebabkan risiko keselamatan yang serius, terutama ketika penumpang melintasi garis kuning pembatas peron saat kereta akan tiba. Di negara berkembang, kasus kecelakaan di stasiun semakin meningkat dari waktu ke waktu [1]. Sistem keamanan manual terbukti belum mampu secara efektif mengurangi angka kecelakaan tersebut. Oleh karena itu, dibutuhkan sistem deteksi otomatis yang mampu memberikan peringatan dini ketika penumpang berada di area berbahaya, terutama di dekat jalur kereta. Salah satu pendekatan yang berkembang pesat untuk mendukung sistem monitoring keamanan adalah teknologi deteksi objek berbasis visi komputer. Deteksi objek bertujuan untuk mengenali dan melokalisasi objek dalam citra menggunakan bounding box. Berkat kemajuan pembelajaran mendalam (deep learning), teknologi ini telah banyak diterapkan pada bidang keamanan, kontrol lalu lintas, dan sistem pengawasan cerdas [2] . Meskipun deteksi objek berbasis Convolutional Neural Network (CNN) telah berhasil diterapkan untuk berbagai jenis objek, tantangan besar masih ditemui dalam mendeteksi manusia di lingkungan yang padat. Deteksi orang menjadi lebih kompleks karena variasi posisi tubuh, perilaku, serta atribut visual seperti pakaian dan orientasi tubuh [3].

Secara umum, algoritma deteksi objek berbasis deep learning dibagi menjadi dua kategori: pendekatan dua tahap (seperti Faster R-CNN [4] dan Mask R-CNN [5]) dan pendekatan satu tahap (seperti YOLO [6] dan SSD [7]). Pendekatan satu tahap umumnya lebih unggul dalam hal kecepatan karena menyederhanakan proses deteksi, namun sering kali mengorbankan akurasi [8] . Selain itu, model-model konvensional tersebut cenderung mengalami keterbatasan saat digunakan untuk mendeteksi objek dalam area yang padat, karena jumlah missed detection yang tinggi[9].

Untuk meningkatkan efisiensi proses deteksi, penelitian ini menggunakan metode *frame difference* digunakan sebagai tahap praproses guna menyaring frame yang mengandung pergerakan signifikan sebelum dikirim ke model deteksi utama. Strategi ini efektif dalam mengurangi beban komputasi, terutama pada perangkat dengan keterbatasan sumber daya seperti sistem berbasis edge. Seiring berkembangnya teknologi, beberapa pendekatan baru mulai dikembangkan untuk mengatasi keterbatasan tersebut, salah satunya adalah Vision Transformer (ViT). Berbeda dengan pendekatan konvolusional seperti YOLO atau SSD yang mengandalkan operasi lokal (convolutional filters), ViT memanfaatkan mekanisme self-attention untuk menangkap informasi global dari seluruh citra, sehingga lebih adaptif terhadap variasi objek yang kompleks dan kondisi kepadatan tinggi.

Menanggapi tantangan dalam efisiensi komputasi dan akurasi deteksi pada perangkat edge, penelitian ini mengusulkan implementasi integrasi metode *frame difference* dengan klasifikasi AI menggunakan model Vision Transformer (ViT-Base) yang telah melalui tahap pre-training menggunakan dataset ImageNet-21K, untuk diterapkan dalam deteksi dan estimasi kepadatan penumpang di stasiun kereta api. Kontribusi utama dari penelitian ini meliputi:

(1) Penggabungan metode *frame difference* sebagai pra-prosessing dengan model ViT-Base untuk meningkatkan akurasi dan efisiensi deteksi objek;
(2) Implementasi metode pada Jetson Orin Nano untuk evaluasi kelayakan ViT-Base sebagai solusi deteksi objek real-time yang hemat daya dan akurat
(3) Evaluasi eksperimental kuantitatif yang menunjukkan bahwa sistem ini mampu mencapai akurasi 91,17%, latensi rata-rata 46,59 ms, konsumsi energi 0,1332 joule/frame, dan efisiensi 0,171 %/msW, menjadikannya solusi andal untuk implementasi pada sistem pemantauan penumpang secara real-time.

Oleh karena itu, penelitian ini difokuskan pada pengembangan sistem deteksi dan estimasi kepadatan penumpang berbasis Vision Transformer yang terintegrasi dengan metode *frame difference* dan diimplementasikan pada perangkat edge Jetson Orin Nano, sebagai solusi cerdas, efisien, dan responsif untuk meningkatkan keselamatan dan kenyamanan di lingkungan transportasi publik.

2. METODE PENELITIAN

2.1 Dataset

ImageNet merupakan sebuah basis data citra berskala besar yang dikembangkan untuk mendukung penelitian dalam bidang visi komputer, khususnya pada tugas klasifikasi dan pengenalan objek. Dataset ini pertama kali diperkenalkan oleh [10] dan dibangun berdasarkan struktur hierarkis WordNet, dengan mencakup lebih dari 14 juta citra beranotasi yang tersebar dalam lebih dari 20.000 kategori (synsets). Berbeda dengan dataset citra sebelumnya yang bersifat terbatas dari segi jumlah kelas dan volume data (seperti MNIST atau CIFAR), ImageNet menawarkan cakupan kelas yang lebih luas, kedalaman hierarki yang lebih kompleks, serta jumlah sampel yang jauh lebih besar. Setiap kategori dalam ImageNet diwakili oleh ratusan hingga ribuan citra, yang diperoleh dari berbagai sumber daring dan melalui proses anotasi manual untuk memastikan akurasi label. Dalam konteks penelitian ini, arsitektur ViT-Base dilatih awal (pre-trained) menggunakan ImageNet untuk memperoleh representasi fitur visual yang kuat sebelum diterapkan pada tugas deteksi kepadatan penumpang.

2.2 Metode Deteksi Objek Bergerak

Metode *frame difference* yang dikombinasikan dengan algoritma klasifikasi berbasis Transformer digunakan dalam penelitian ini untuk mendeteksi objek secara efisien pada citra bergerak. Tabel 1 merupakan pseudocode pada metode ini. Algoritma dirancang menjadi tiga tahapan utama, yaitu Deteksi Gerakan (Movement Detection) dan Pra-pemrosesan serta Klasifikasi (Pre-processing & Classification), yang dijalankan dalam struktur loop hingga seluruh frame video selesai diproses.

Pada fase pertama (Deteksi Gerakan), dua frame video berturut-turut, yaitu $Frame_I$ dan $Frame_{\{i+1\}}$, diambil secara berkelanjutan. Kedua frame ini dibandingkan menggunakan metode *frame difference* untuk mendeteksi perubahan antarframe. Hasil perbedaan ini kemudian melalui serangkaian proses morfologis seperti **erosi, dilasi, perataan (blur),** dan penyaringan menggunakan ambang batas (**thresholding)**. Sampling dilakukan terhadap hasil threshold untuk menentukan apakah perubahan yang terdeteksi cukup signifikan. Jika ukuran sampel melebihi ambang yang telah ditentukan, maka sistem akan masuk ke fase selanjutnya.

Pada fase kedua (Pra-pemrosesan), setiap frame yang telah diseleksi akan dipangkas (cropping), diubah ukurannya (resizing), dan dinormalisasi melalui pengurangan nilai rata-rata (mean subtraction). Setelah itu, citra hasil normalisasi akan dilanjutkan pada fase ketiga (Klasifikasi), citra diklasifikasikan menggunakan model Vision Transformer (ViT-Base). Jika objek berhasil dikenali, sistem akan membuat kotak pembatas (bounding box) dan menampilkan kelas objek pada citra. Setiap iterasi dari proses ini merepresentasikan satu unit pemrosesan dari input video berwarna tiga kanal (RGB). Setelah satu frame selesai diproses, penghitung frame akan diperbarui dan sistem memeriksa apakah semua frame dalam video telah diproses. Jika belum, proses akan berulang untuk frame berikutnya hingga seluruh video selesai dianalisis.

2.2 Arsitektur *Vision Transformer* (ViT)

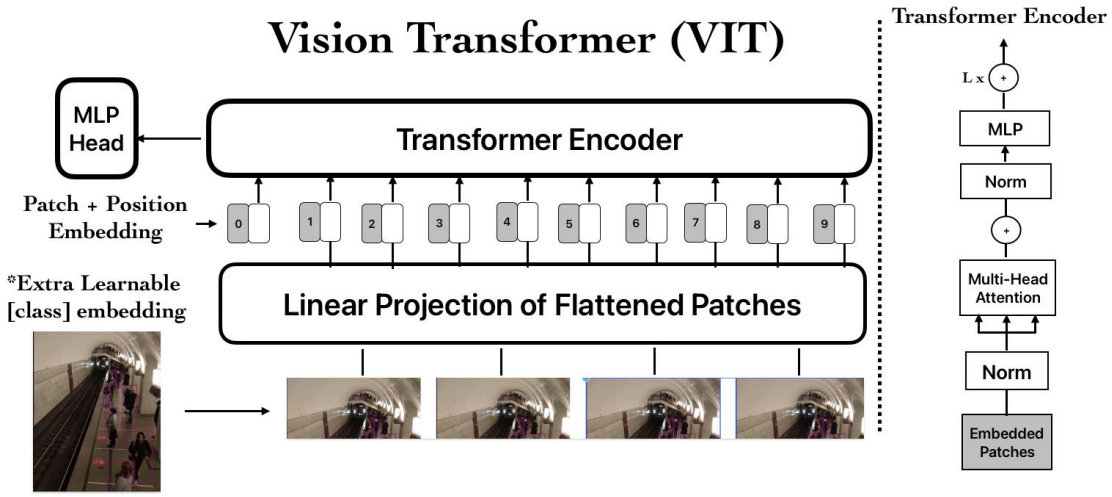
Vision Transformer (ViT), khususnya varian ViT-Base, merepresentasikan peralihan dari arsitektur jaringan konvolusional (convolutional neural networks) menuju struktur berbasis transformer untuk aplikasi pengolahan citra.

TABEL 1 : PSEUDOCODE METODE FRAME DIFFERENCE MENGGUNAKAN VISION TRANSFORMER SEBAGAI AI CLASSIFIER

Pseudocode

```
Start
while True:
    # PHASE 1: MOVEMENT DETECTION
    Frame_i = capture_frame()
    Frame_i_plus1 = capture_next_frame()

    diff_frame = frame_difference(Frame_i, Frame_i_plus1)
    eroded_frame = apply_erosion(diff_frame)
    dilated_frame = apply_dilation(eroded_frame)
    blurred_frame = apply_blur(dilated_frame)
    threshold_frame = apply_threshold(blurred_frame)
    sampled_frame = sample(threshold_frame)
    if sample_size(sampled_frame) > frame_threshold:
        goto PreProcessing

# PHASE 2: PRE-PROCESSING
PreProcessing:
for i in range(sample_size):
    Frame_i = get_frame(i)

    # Pre-processing
    cropped = crop(Frame_i)
    resized = resize(cropped)
    normalized = mean_subtraction(resized)

    # Transformer classification
    class_result = classifier(normalized)

    if class_result in object_classes:
        bbox = get_bounding_box(class_result)
        write_to_frame(Frame_i, bbox)

# PHASE 3: CLASSIFICATION
    # DPU Classification check
    if i == frame_number:
        break
End
```

Model ini menggunakan mekanisme self-attention untuk mengumpulkan dan mengintegrasikan ketergantungan global dalam citra masukan secara efektif. ViT-Base dikenal mampu memberikan presisi yang lebih tinggi, meskipun memiliki kelemahan berupa waktu pemrosesan yang lebih lama dan kebutuhan komputasi yang lebih besar. Gambar 1 menyajikan gambaran umum arsitektur Vision Transformer (ViT), yang memperlihatkan bagaimana citra diproses sebagai rangkaian potongan (patches) alih-alih sebagai struktur dua dimensi utuh[11].

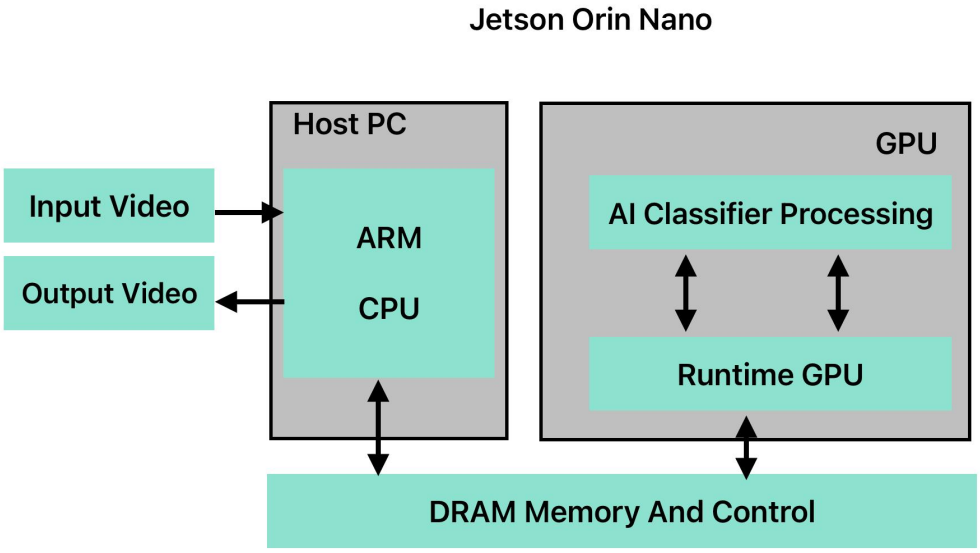


Gambar 1: Arsitektur Vision Transformer (ViT)

Proses dimulai dengan membagi citra masukan menjadi potongan-potongan berukuran tetap, kemudian dilakukan flattening dan embedding linier pada masing-masing potongan untuk memproyeksikannya ke dalam ruang vektor berdimensi seragam. Embedding posisi (positional embeddings) ditambahkan ke setiap patch untuk menyandikan informasi spasial, sehingga model tetap memiliki pemahaman terhadap posisi relatif setiap patch dalam citra asli. Selanjutnya, representasi tersebut dimasukkan ke dalam encoder standar milik arsitektur Transformer, yang terdiri atas beberapa lapisan multi-head self-attention dan jaringan feed-forward, dengan penerapan residual connection dan layer normalization di setiap tahap. Arsitektur ini memanfaatkan skalabilitas dan efektivitas Transformer yang awalnya dikembangkan untuk pemrosesan bahasa alami (Natural Language Processing), namun kini berhasil diadaptasi untuk tugas pengenalan citra (image recognition).

2.3 Arsitektur Sistem dan Alur Pemrosesan Jetson Orin Nano

Sistem deteksi kepadatan penumpang yang diusulkan dalam penelitian ini dibangun di atas arsitektur pemrosesan video berbasis Jetson Orin Nano, sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Arsitektur Sistem Jetson Orin Nano

p-ISSN: 2356-0533; e-ISSN: 2355-9195

Sistem ini terdiri atas dua komponen utama, yaitu **host PC** dan **modul Jetson Orin Nano**, yang berperan penting dalam menjalankan proses pemrosesan visual dan inferensi kecerdasan buatan secara efisien. Pada tahap awal, **video masukan** diperoleh melalui kamera pengawas (CCTV) dan diterima oleh **host PC**. Video tersebut kemudian diproses secara awal oleh prosesor **ARM** dan **CPU** yang terdapat pada host. **Prosesor ARM** digunakan untuk menangani tugas-tugas dasar pemrosesan awal dengan efisiensi daya yang tinggi, sementara **CPU** bertugas mengelola komputasi praproses yang lebih kompleks, seperti *frame extraction*, *resizing*, dan normalisasi citra. Proses ini bertujuan untuk menyiapkan data agar sesuai dengan format input model ViT-Base sebelum dilakukan inferensi [12].

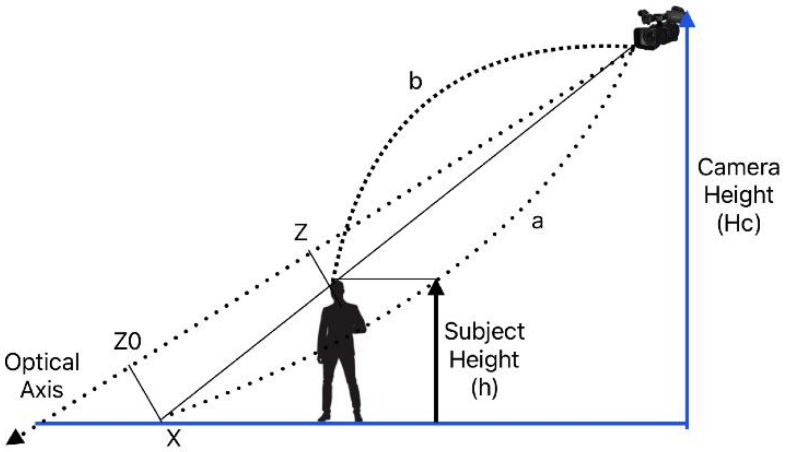
Setelah proses praproses selesai, data ditransmisikan ke Jetson Orin Nano melalui **memori DRAM** dan modul kontrol. Peran **memori dan pengendali DRAM** sangat krusial dalam menyediakan jalur data berkecepatan tinggi dan manajemen memori yang optimal selama proses inferensi berlangsung. Jetson Orin Nano menjalankan inti proses deteksi objek dengan memanfaatkan **GPU internal**, yang terdiri dari dua fungsi utama, yaitu:

1. **Pemrosesan Klasifikasi AI**, yaitu eksekusi model ViT-Base untuk mendeteksi dan mengestimasi jumlah penumpang secara visual, serta
2. **Aktivitas Runtime GPU**, yang mendukung pemrosesan paralel dan pengelolaan beban kerja *deep learning*.

Model **Vision Transformer (ViT-Base)** yang digunakan sebelumnya telah melalui proses *pre-training* menggunakan ImageNet dan kemudian difine-tuning pada dataset kepadatan penumpang. Model ini dioptimalkan menggunakan **TensorRT** untuk meningkatkan efisiensi inferensi pada perangkat edge Jetson Orin Nano. Pemanfaatan **kemampuan paralel GPU** pada Jetson Orin Nano memungkinkan sistem untuk menjalankan inferensi secara real-time dengan latensi rendah dan konsumsi daya yang efisien. Proses akhir mencakup penghitungan jumlah objek (penumpang) dan estimasi tingkat kepadatan berdasarkan output dari bounding box yang dihasilkan oleh ViT-Base.

2.3 Tahapan Pengujian Sistem

Gambar 3 menunjukkan konsep perhitungan tinggi suatu objek berdasarkan sudut pandang kamera dan prinsip proyeksi geometri. Titik x merupakan titik pada permukaan tanah yang diperoleh dari perpanjangan garis yang menghubungkan kamera dengan kepala subjek.



Gambar 3 Proyeksi Geometri Objek

Dalam hal ini, jarak a adalah jarak dari kamera ke titik X dipermukaan tanah, sedangkan b adalah jarak dari kamera ke kepala subjek secara langsung. Rasio antara kedua jarak ini, yaitu b/a, dianggap setara dengan

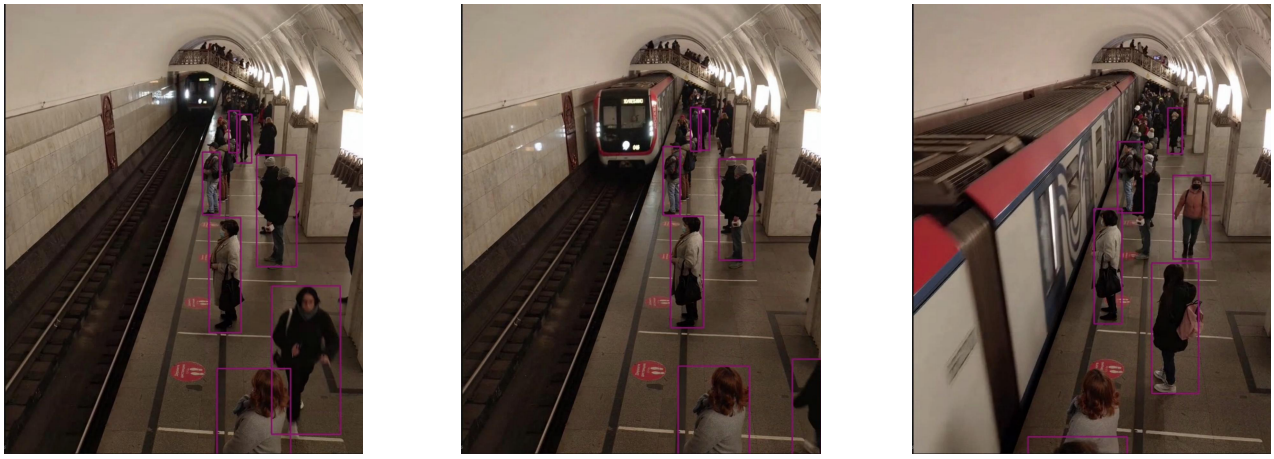
rasio jarak z/z_0, di mana z adalah proyeksi horizontal dari kamera ke kepala subjek pada sumbu optik, dan z_0 merupakan total jarak horizontal dari kamera ke titik proyeksi sejajar di permukaan tanah.

$$h = H_c \times \left(1 - \frac{b}{a}\right) = H_c \times \left(1 - \frac{z}{z_0}\right) \quad (1)$$

Dengan menggunakan kesetaraan rasio tersebut, tinggi objek atau subjek dalam citra dapat dihitung menggunakan pendekatan geometris secara tidak langsung. Hubungan antara tinggi kamera (Hc) dan tinggi subjek (h) dapat ditentukan menggunakan persamaan (1).

3. HASIL DAN PEMBAHASAN

Evaluasi kinerja sistem dilakukan untuk mengukur efektivitas metode *frame difference* yang diintegrasikan dengan model ViT-Base dalam mendeteksi kepadatan penumpang pada stasiun kereta api. Pengujian dilakukan menggunakan data uji yang representatif dari lingkungan nyata, dengan variasi kepadatan penumpang dan kondisi pencahayaan. Gambar 4 menunjukkan hasil visualisasi deteksi kepadatan penumpang yang diperoleh dari eksperimen yang telah diimplementasikan pada perangkat edge Jetson Orin Nano. Pengujian dilakukan pada lingkungan nyata, yaitu stasiun kereta api yang terletak di Stasiun Kereta Bawah Tanah Leninskiy Prospekt, dengan kondisi pencahayaan dan kepadatan penumpang yang bervariasi. Masing-masing gambar merepresentasikan urutan kedatangan kereta sebelum kereta datang, saat kereta mendekat, dan ketika kereta berhenti di peron. Hasil deteksi ditampilkan melalui bounding box berwarna ungu yang mengindikasikan lokasi dan jumlah penumpang yang berhasil dikenali oleh sistem.



Gambar 4. Hasil visualisasi deteksi kepadatan penumpang

Hasil pengujian disajikan pada Tabel 2. Dari hasil pengamatan, metode ini mampu mendeteksi objek manusia secara konsisten pada seluruh tahap, termasuk ketika kereta bergerak masuk yang menyebabkan gangguan visual dinamis. Hal ini menunjukkan bahwa model tidak hanya efektif dalam mendeteksi individu pada kondisi statis, tetapi juga mampu mempertahankan performa ketika terjadi perubahan intensitas cahaya dan latar belakang bergerak. Secara kuantitatif, sistem mencatat akurasi deteksi sebesar 91,17%, latensi rata-rata 46,59 ms, konsumsi energi 0,1332 joule/frame, dan efisiensi komputasi 0,171 %/msW. Temuan ini mengindikasikan bahwa ViT-Base yang dipadukan dengan metode *frame difference* dapat beroperasi secara real-time dengan efisiensi tinggi, menjadikannya cocok untuk sistem monitoring berbasis edge di lingkungan transportasi publik.

TABEL 2 : HASIL PENGUJIAN

Parameter Evaluasi	Hasil	Keterangan
Akurasi	91,17%	Tingkat akurasi model.
Latensi	46,59 ms	Waktu inferensi per frame
Konsumsi Energi	0,1332 joule/frame	Energi yang dibutuhkan untuk setiap inferensi
Efisiensi Komputasi	0,171 %/msW	Perbandingan antara akurasi, waktu, dan daya

Dalam penelitian ini, metrik evaluasi yang digunakan mencakup akurasi, latensi, konsumsi energi, dan efisiensi komputasi. Pengukuran konsumsi energi dilakukan untuk mengetahui seberapa besar energi (dalam satuan joule) yang digunakan oleh sistem saat menjalankan proses inferensi deteksi objek. Pada platform Jetson Orin Nano, proses pengukuran dilakukan menggunakan perangkat lunak JTop, yaitu alat pemantau yang secara khusus dirancang untuk perangkat NVIDIA Jetson. JTop dapat merekam data mengenai konsumsi daya, latensi proses, serta pemanfaatan sumber daya GPU dan CPU selama eksperimen berlangsung Efisiensi komputasi bertujuan untuk mengukur sejauh mana sistem dapat menjalankan tugas deteksi objek dengan akurat, cepat, dan hemat daya. Dalam penelitian ini, perhitungan efisiensi mengacu pada rumus yang diadopsi dari referensi [13], yang digunakan untuk membandingkan rasio performa terhadap konsumsi sumber daya. Ditunjukkan dalam persamaan 2.

$$Efisiensi = \frac{Akurasi\ (\%}{Latensi\ (ms)\ x\ Daya\ (Watt)} \quad (2)$$

Dengan kata lain, semakin tinggi nilai efisiensi, maka model semakin optimal dalam menghasilkan output yang akurat dengan waktu proses dan konsumsi daya yang lebih rendah.

4. KESIMPULAN

Penelitian ini berhasil mengintegrasikan metode *frame difference* dengan arsitektur Vision Transformer (ViT-Base) pada perangkat Jetson Orin Nano untuk mendeteksi dan mengestimasi kepadatan penumpang di stasiun kereta api secara real-time. Metode ini menggunajan model ViT-Base yang telah dilatih awal (pre-trained) menggunakan dataset ImageNet-21K menunjukkan performa yang kompetitif dalam skenario deteksi objek dengan tingkat kepadatan tinggi. Berdasarkan hasil evaluasi, sistem yang dikembangkan mampu mencapai akurasi sebesar 91,17%, dengan latensi rata-rata 46,59 ms, konsumsi energi 0,1332 joule/frame, dan efisiensi komputasi sebesar 0,171 %/msW. Pencapaian ini menunjukkan bahwa metode ini secara efisien dapat diterapkan pada perangkat edge dengan konsumsi daya rendah, tanpa mengorbankan akurasi dan kecepatan. Secara keseluruhan, penelitian ini layak diterapkan untuk sistem monitoring kepadatan penumpang berbasis edge computing, khususnya pada lingkungan transportasi publik yang menuntut efisiensi energi dan respons waktu nyata. Ke depan, penelitian ini dapat dikembangkan lebih lanjut dengan menambahkan kemampuan prediksi pelanggaran zona bahaya (danger zone prediction) dan integrasi dengan sistem peringatan dini visual atau suara untuk mendukung keselamatan penumpang.

DAFTAR PUSTAKA

[1] S. Terabe, T. Kato, H. Yaginuma, N. Kang, and K. Tanaka, "Risk Assessment Model for Railway Passengers on a Crowded Platform," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 1, pp. 524–531, Jan. 2019, doi: 10.1177/0361198118821925.

[2] L. Jiao *et al.*, "A Survey of Deep Learning-Based Object Detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/ACCESS.2019.2939201.

[3] M. Ahmad, I. Ahmed, and A. Adnan, "Overhead View Person Detection Using YOLO," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, Oct. 2019, pp. 0627–0633. doi: 10.1109/UEMCON47517.2019.8992980.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[5] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 2961–2969.
[6] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," Jul. 2021.
[7] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," Dec. 2015, doi: 10.1007/978-3-319-46448-0_2.
[8] M. Nurul Achmadiah, A. Ahamad, C.-C. Sun, and W.-K. Kuo, "Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems," *IEEE Internet Things J*, vol. 12, no. 11, pp. 16681–16694, Jun. 2025, doi: 10.1109/JIOT.2025.3536526.
[9] M. N. Achmadiah, N. Setyawan, A. A. Bryantono, C.-C. Sun, and W.-K. Kuo, "Fast Person Detection Using YOLOX With AI Accelerator For Train Station Safety," in *2024 International Electronics Symposium (IES)*, IEEE, Aug. 2024, pp. 504–509. doi: 10.1109/IES63037.2024.10665874.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[11] J. Pan *et al.*, "EdgeViTs: Competing Light-weight CNNs on Mobile Devices with Vision Transformers," May 2022.
[12] Youvan, "Developing and deploying ai applications on nvidia jetson orin nx: A comprehensive guide," 2024.

p-ISSN: 2356-0533; e-ISSN: 2355-9195