Jurnal Elkolind Volume 12, Nomor 2, Juli 2025
DOI: http://dx.doi.org/10.33795/elkolind.v12i2/8730

Perbandingan Efisiensi Vision Transformer dan MobileNet untuk Optimasi Deteksi Objek di Edge Device

Mas Nurul Achmadiah[1], Novendra Setyawan[2], Anindya Dwi Risdhayanti[3],
e-mail: masnurul@polinema.ac.id, novendra@umm.ac.id, anindya@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 8 Juni 2025
Direvisi 24 Juli 2025
Diterbitkan 31 Juli 2025

Kata kunci:
Vision Transformer
MobileNet
Komputasi Tepi
Efisiensi Komputasi

ABSTRAK

Penelitian ini membandingkan kinerja Vision Transformer (ViT-Base) dan MobileNet untuk deteksi objek pada perangkat edge. Evaluasi dilakukan pada Jetson Orin Nano dengan menggunakan empat parameter: akurasi, latensi, konsumsi energi, dan efisiensi komputasi. Hasil pengujian menunjukkan MobileNet mencapai akurasi 100%, latensi 41,38 ms, konsumsi energi 0,3937 joule/frame, dan efisiensi 0,8332 %/msW. Sementara itu, ViT-Base memperoleh akurasi 93,72%, latensi 63,58 ms, konsumsi energi 0,5306 joule/frame, dan efisiensi 0,4466 %/msW. MobileNet lebih unggul dalam hal akurasi, efisiensi, kecepatan, dan penggunaan energi. Temuan ini membuktikan, MobileNet direkomendasikan untuk aplikasi edge real-time yang menuntut respon cepat dan hemat daya, serta sesuai untuk skenario yang membutuhkan akurasi tinggi.

Keywords:

Vision Transformer
MobileNet
Edge Computing
Computational Efficiency

ABSTRACT

This study compares the performance of the Vision Transformer (ViT-Base) and MobileNet for object detection on edge devices. The evaluation was conducted on the Jetson Orin Nano using four key parameters: accuracy, latency, energy consumption, and computational efficiency. Experimental results show that MobileNet achieved 100% accuracy, an average latency of 41.38 ms, energy consumption of 0.3937 joules/frame, and computational efficiency of 0.8332 %/msW. In contrast, ViT-Base obtained 93.72% accuracy, a latency of 63.58 ms, energy consumption of 0.5306 joules/frame, and computational efficiency of 0.4466 %/msW. MobileNet demonstrates superior performance in terms of accuracy, efficiency, speed, and energy utilization. These findings indicate that MobileNet is recommended for real-time edge applications that demand fast response and low power consumption, while still being suitable for scenarios requiring high accuracy.

Penulis Korespondensi:
Mas Nurul Achmadiah
Jurusan Teknik Elektro
Politeknik Negeri Malang,

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

Jalan Soekarno Hatta No.9 Malang, Indonesia
Email: masnurul@polinema.ac.id

1. PENDAHULUAN

Deteksi objek merupakan salah satu bidang penting dalam visi komputer yang bertujuan untuk mengenali dan melokalisasi objek dalam citra menggunakan bounding box [1] . Teknologi ini telah banyak diterapkan pada berbagai domain, seperti keamanan, kontrol lalu lintas, dan sistem pengawasan cerdas [2], [3] . Berkat kemajuan pembelajaran mendalam (deep learning), performa deteksi objek meningkat signifikan, namun tantangan tetap muncul ketika objek memiliki variasi posisi, orientasi, maupun atribut visual yang beragam [3].

Secara umum, algoritma deteksi objek berbasis deep learning dibagi menjadi dua kategori: pendekatan dua tahap (misalnya Faster R-CNN [4] dan Mask R-CNN [5]) dan pendekatan satu tahap (misalnya YOLO [6] dan SSD [7]). Pendekatan satu tahap umumnya lebih unggul dalam hal kecepatan karena proses deteksi yang lebih sederhana, namun sering kali mengorbankan akurasi [8] . Selain itu, model konvensional masih memiliki keterbatasan ketika digunakan pada kondisi kompleks, seperti peningkatan missed detection [9].

Untuk meningkatkan efisiensi pada perangkat dengan keterbatasan sumber daya, arsitektur ringan seperti MobileNet banyak digunakan karena dirancang untuk kecepatan dan konsumsi daya rendah. Di sisi lain, pendekatan Vision Transformer (ViT) memperkenalkan mekanisme self-attention yang mampu menangkap informasi global dari citra, sehingga lebih adaptif terhadap variasi objek yang kompleks, meskipun membutuhkan beban komputasi lebih tinggi.

Penelitian ini berfokus pada perbandingan efisiensi dua arsitektur populer, yaitu MobileNet dan Vision Transformer (ViT-Base), dengan implementasi pada perangkat edge. Evaluasi dilakukan menggunakan empat metrik utama: akurasi, latensi, konsumsi energi, dan efisiensi komputasi. Kontribusi utama penelitian ini meliputi:

(1) Analisis perbandingan kinerja MobileNet dan ViT-Base dalam deteksi objek;
(2) Implementasi kedua metode pada edge device Jetson Orin Nano untuk mengevaluasi kelayakan deteksi real-time yang hemat daya;
(3) Penyajian hasil eksperimen kuantitatif yang menekankan trade-off antara akurasi dan efisiensi, sebagai acuan dalam pemilihan model optimal untuk aplikasi deteksi objek berbasis edge computing.

Dengan demikian, penelitian ini difokuskan pada studi perbandingan efisiensi MobileNet dan Vision Transformer sebagai solusi deteksi objek yang cerdas, efisien, dan responsif untuk mendukung aplikasi komputasi tepi.

1. METODE PENELITIAN

2.1 Dataset

ImageNet adalah basis data citra berskala besar yang mendukung penelitian visi komputer, khususnya klasifikasi dan pengenalan objek [10] Dataset ini berisi lebih dari 14 juta citra beranotasi dalam lebih dari 20.000 kategori, jauh lebih luas dibanding dataset sebelumnya seperti MNIST atau CIFAR. Setiap kategori terdiri dari ratusan hingga ribuan citra yang diperoleh dari berbagai sumber daring dengan anotasi manual. Dalam penelitian ini, arsitektur ViT-Base dilatih awal (pre-trained) pada ImageNet untuk menghasilkan representasi fitur visual yang kuat sebelum digunakan pada tugas deteksi objek.

2.2 Metode Deteksi Objek Bergerak

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

Dalam penelitian ini, digunakan metode frame difference yang dikombinasikan dengan algoritma klasifikasi berbasis Transformer untuk melakukan deteksi objek secara efisien pada citra bergerak. Prinsip utamanya adalah hanya memproses frame yang mengandung perubahan signifikan, sehingga beban komputasi dapat ditekan. Alur metode ini dijelaskan dalam bentuk pseudocode pada Tabel 1, yang terdiri dari tiga tahap utama: deteksi gerakan (movement detection), pra-pemrosesan (pre-processing), dan klasifikasi (classification). Seluruh tahapan dieksekusi dalam struktur loop hingga seluruh frame video selesai dianalisis.

Pada tahap pertama, yaitu deteksi gerakan, dua frame video berturut-turut, $Frame_i$ dan $Frame_{\{i+1\}}$, diambil secara berkelanjutan. Kedua frame tersebut dibandingkan menggunakan metode frame difference untuk mengidentifikasi adanya perubahan antarframe. Hasil perbedaan kemudian diproses melalui serangkaian operasi morfologis, termasuk erosi, dilasi, blurring, dan thresholding. Proses ini bertujuan mempertegas perbedaan yang relevan sekaligus mengurangi noise. Selanjutnya dilakukan sampling untuk menilai apakah perubahan yang terdeteksi cukup besar. Jika ukuran sampel melampaui ambang yang telah ditetapkan, maka frame tersebut diteruskan ke tahap berikutnya.

Tahap kedua adalah pra-pemrosesan, di mana frame yang lolos seleksi akan dipangkas (cropping), diubah ukurannya (resizing), dan dinormalisasi melalui teknik pengurangan nilai rata-rata (mean subtraction). Proses ini memastikan bahwa data masukan berada dalam format yang sesuai untuk model klasifikasi. Tahap ketiga adalah klasifikasi, di mana citra hasil pra-pemrosesan dimasukkan ke dalam model Vision Transformer (ViT-Base) dan MobileNetsecara bergantian untuk membandingkan performa kedua metode tersebut. Model ini melakukan pengenalan objek dan menghasilkan keluaran berupa kotak pembatas (bounding box) serta label kelas yang sesuai. Jika objek berhasil dikenali, informasi tersebut ditampilkan pada frame. Proses ini berlangsung secara iteratif untuk setiap frame RGB, dan setelah satu frame selesai diproses, penghitung frame akan diperbarui. Sistem kemudian memeriksa apakah seluruh frame video telah dianalisis. Jika masih ada frame tersisa, maka siklus diulang hingga semua frame selesai.

TABEL 1 : PSEUDOCODE METODE FRAME DIFFERENCE DAN AI CLASSIFIER MENGGUNAKAN VISION TRANSFORMER (VIT-BASE) DAN MOBILENET

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)
```

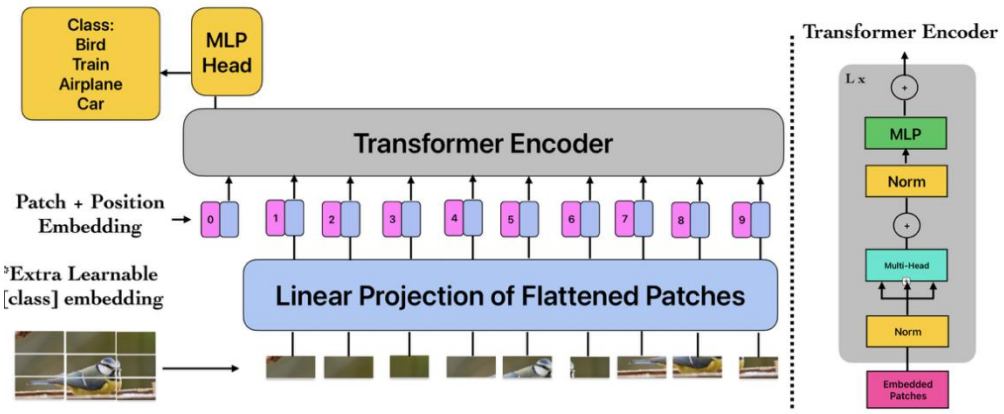
```
            if class_result in object_classes:
                bbox = get_bounding_box(class_result)
                write_to_frame(Frame_i, bbox)

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

Pendekatan ini tidak hanya meningkatkan efisiensi dengan mengurangi jumlah frame yang diproses, tetapi juga menjaga akurasi sistem dengan memastikan hanya frame yang relevan yang diklasifikasikan. Dengan demikian, integrasi frame difference dan Vision Transformer memberikan solusi efektif untuk deteksi objek real-time pada perangkat dengan keterbatasan sumber daya.

2.2 Arsitektur *Vision Transformer* (ViT) dan MobileNet

Vision Transformer (ViT), khususnya varian ViT-Base, merepresentasikan peralihan dari arsitektur jaringan konvolusional (convolutional neural networks) menuju struktur berbasis transformer untuk aplikasi pengolahan citra. 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 [11] . 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[8].

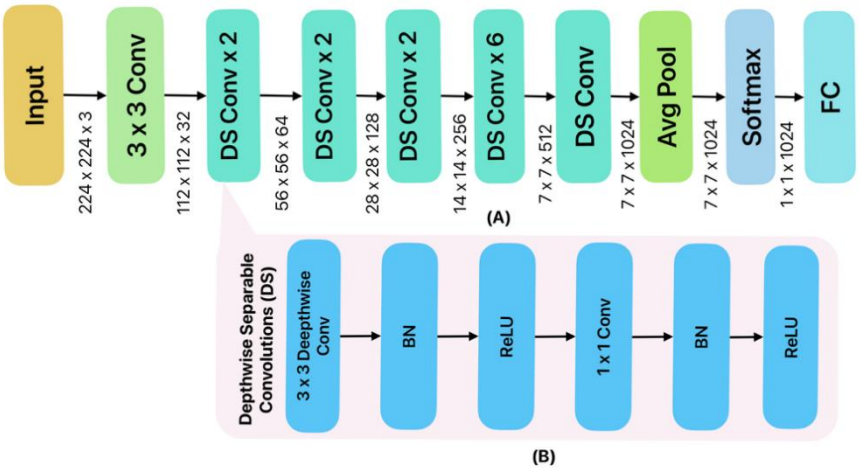


Gambar 1: Arsitektur Vision Transformer (ViT)

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).

Gambar 2 menyajikan arsitektur Mobilenet. MobileNet adalah arsitektur jaringan saraf konvolusional yang dikembangkan untuk menghadirkan efisiensi tinggi pada perangkat dengan keterbatasan komputasi dan daya, seperti ponsel pintar dan edge device. Keunggulan utama MobileNet terletak pada penggunaan Depthwise Separable Convolution, yaitu teknik yang membagi proses konvolusi menjadi dua tahap, yakni depthwise convolution untuk mengekstraksi fitur per kanal dan pointwise convolution (konvolusi 1×1) untuk menggabungkan hasil antar kanal. Pendekatan ini mampu memangkas jumlah parameter dan operasi komputasi secara signifikan dibandingkan konvolusi konvensional, sehingga mempercepat inferensi tanpa mengorbankan akurasi secara besar. Berdasarkan gambar arsitektur, proses MobileNet dimulai dari input layer berupa citra RGB berukuran 224×224×3 yang kemudian diproses melalui konvolusi awal 3×3 menghasilkan fitur berukuran 112×112×32. Setelah itu, citra melewati serangkaian blok Depthwise Separable Convolution seperti DS Conv×2, DS Conv×6, dan seterusnya.

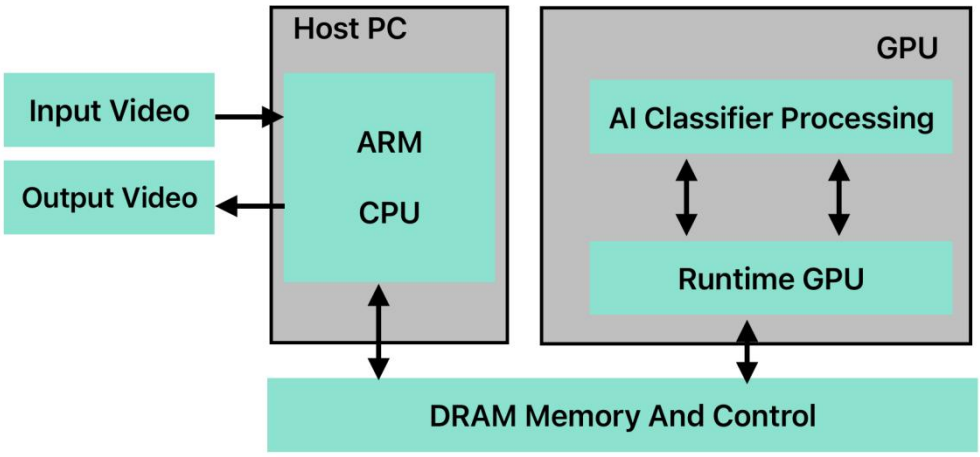


Gambar 2: Arsitektur MobileNet

Pada setiap blok, dilakukan operasi konvolusi, diikuti Batch Normalization serta fungsi aktivasi ReLU untuk menjaga stabilitas dan menambahkan non-linearitas. Selanjutnya, fitur dipadatkan melalui Average Pooling sehingga ukurannya lebih ringkas, lalu diteruskan ke fully connected layer dan softmax classifier guna menghasilkan prediksi kelas objek. Dengan rancangan yang ringkas, MobileNet mampu memberikan performa cepat, hemat energi, serta ideal untuk aplikasi visi komputer real-time pada perangkat edge.

2.3 Arsitektur Sistem dan Alur Pemrosesan Jetson Orin Nano

Sistem deteksi objek yang diusulkan dalam penelitian ini dibangun pada arsitektur pemrosesan video berbasis Jetson Orin Nano, sebagaimana ditunjukkan pada Gambar 3. Sistem terdiri atas dua komponen utama, yaitu host PC dan modul Jetson Orin Nano, yang bekerja sama dalam menjalankan pemrosesan visual dan inferensi model deep learning secara efisien. Pada tahap awal, video masukan diperoleh melalui kamera pengawas (CCTV) dan diterima oleh host PC. Video kemudian diproses menggunakan prosesor ARM dan CPU untuk melakukan praproses, termasuk frame extraction, resizing, dan normalisasi citra, sehingga data sesuai dengan format input model sebelum dikirimkan ke perangkat edge



Gambar 3. Arsitektur Sistem Jetson Orin Nano

Selanjutnya, data ditransmisikan ke Jetson Orin Nano melalui modul memori DRAM dan pengendali internal untuk memastikan aliran data berkecepatan tinggi selama inferensi. Jetson Orin Nano menjalankan inti deteksi objek dengan memanfaatkan GPU internal, yang mencakup dua fungsi utama: (1) eksekusi model AI, yaitu penerapan arsitektur MobileNet atau Vision Transformer (ViT-Base) untuk klasifikasi dan deteksi objek, serta (2) aktivitas runtime GPU, yang mendukung komputasi paralel dan distribusi beban kerja.

Baik MobileNet maupun ViT-Base yang digunakan telah melalui proses pre-training pada ImageNet, lalu di-fine-tuning menggunakan dataset yang sesuai dengan skenario deteksi objek. Untuk mengoptimalkan performa pada perangkat edge, model diakselerasi menggunakan TensorRT, sehingga mampu berjalan dengan latensi rendah dan konsumsi daya yang efisien. Dengan memanfaatkan kemampuan paralel GPU Jetson Orin Nano, sistem ini dapat mendukung deteksi objek real-time serta memungkinkan analisis perbandingan kinerja antara MobileNet dan ViT-Base secara menyeluruh.

2.3 Tahapan Pengujian Sistem

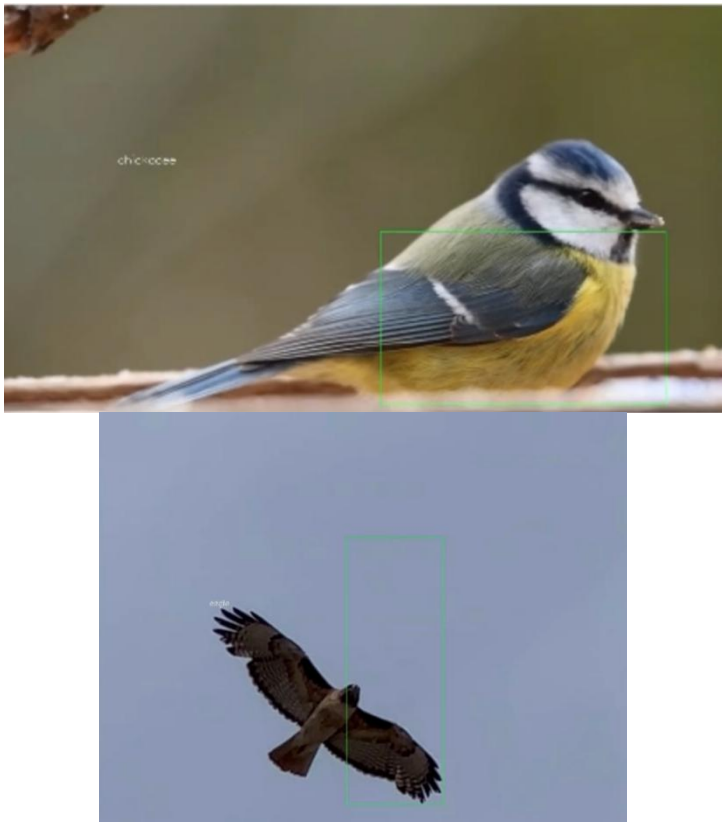
Pengujian sistem deteksi objek dalam penelitian ini dilakukan secara bertahap mengikuti rancangan pseudocode. Tahap pertama adalah deteksi gerakan (movement detection), di mana dua frame video berturut-turut diambil dan dibandingkan menggunakan metode frame difference untuk mendeteksi perubahan antar frame. Hasil perbedaan tersebut kemudian melalui serangkaian operasi morfologis seperti erosion, dilation, blurring, dan thresholding. Proses ini bertujuan untuk mempertegas area pergerakan yang signifikan sekaligus mengurangi noise. Frame yang memiliki perubahan di atas ambang batas (threshold) akan diteruskan ke tahap berikutnya.

Tahap kedua adalah pra-pemrosesan (pre-processing). Pada fase ini, frame yang lolos seleksi akan diproses melalui cropping, resizing, dan normalisasi menggunakan mean subtraction. Tahapan ini memastikan citra berada pada format standar sebelum dimasukkan ke dalam model klasifikasi. Selanjutnya, frame diproses menggunakan arsitektur MobileNet atau Vision Transformer (ViT-Base) untuk menghasilkan prediksi kelas. Jika objek terdeteksi, sistem akan menampilkan bounding box dan label kelas pada frame.

Tahap ketiga adalah klasifikasi akhir (classification). Pada tahap ini, hasil inferensi diperiksa dan disimpan hingga seluruh frame video selesai diproses. Evaluasi pengujian mencakup pengukuran akurasi deteksi, latensi rata-rata, konsumsi energi, serta efisiensi komputasi, guna membandingkan performa kedua model pada perangkat edge.

2. HASIL DAN PEMBAHASAN

Evaluasi kinerja sistem dilakukan untuk mengukur efektivitas arsitektur MobileNet dan Vision Transformer (ViT-Base) dalam mendeteksi objek pada perangkat edge. Pengujian dilakukan menggunakan data uji yang mewakili kondisi nyata dengan variasi pencahayaan dan latar belakang.



Gambar 4. Hasil visualisasi deteksi

Gambar 4 menunjukkan hasil visualisasi deteksi objek yang diperoleh dari eksperimen pada Jetson Orin Nano. Setiap gambar menampilkan hasil bounding box yang mengindikasikan lokasi serta kelas objek yang berhasil dikenali sistem. Hasil pengujian dirangkum pada Tabel 2. Dari hasil observasi, kedua model mampu mendeteksi objek secara konsisten, termasuk ketika terdapat perubahan kondisi visual yang dinamis. Hal ini memperlihatkan bahwa baik MobileNet maupun ViT-Base tidak hanya efektif pada kondisi statis, tetapi juga tetap stabil ketika intensitas cahaya dan latar belakang mengalami perubahan. Secara kuantitatif, MobileNet mencatat akurasi sebesar 100%, latensi rata-rata 41,38 ms, konsumsi energi 0,3937 joule/frame, serta efisiensi komputasi 0,8332 %/msW.

Tabel 2 : Hasil Pengujian

Metode	Akurasi (%)	Latensi (ms)	Energi (Joule)	Efisiensi (%/msW)
MobileNet	100.00	41.38	0.3937	0.8332
ViT-Base	93.72	63.58	0.5306	0.4466

Sementara itu, ViT-Base menghasilkan akurasi 93,72%, latensi 63,58 ms, konsumsi energi 0,5306 joule/frame, dan efisiensi 0,4466 %/msW. Metrik evaluasi yang digunakan mencakup akurasi, latensi, konsumsi energi, dan efisiensi komputasi. Konsumsi energi diukur dalam satuan joule menggunakan

perangkat lunak JTop pada platform Jetson Orin Nano, yang juga merekam pemanfaatan daya GPU dan CPU selama eksperimen berlangsung.

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

Efisiensi komputasi dihitung berdasarkan rasio antara performa deteksi dengan konsumsi sumber daya, sesuai dengan persamaan pada referensi [10] . Dengan kata lain, semakin tinggi nilai efisiensi, semakin optimal sebuah model dalam menghasilkan output akurat dengan latensi rendah dan konsumsi daya minimal.

3. KESIMPULAN

Penelitian ini berhasil mengimplementasikan arsitektur MobileNet dan Vision Transformer (ViT-Base) pada perangkat Jetson Orin Nano untuk melakukan deteksi objek secara real-time. Hasil pengujian menunjukkan bahwa MobileNet mampu mencapai akurasi 100%, dengan latensi rata-rata 41,38 ms, konsumsi energi 0,3937 joule/frame, dan efisiensi komputasi 0,8332 %/msW. Sebaliknya, ViT-Base mencatat akurasi 93,72%, latensi 63,58 ms, konsumsi energi 0,5306 joule/frame, serta efisiensi 0,4466 %/msW. Temuan ini menegaskan bahwa MobileNet lebih unggul dalam hal efisiensi energi, kecepatan, dan akurasi pada perangkat edge, sementara ViT-Base tetap menawarkan kinerja kompetitif terutama pada skenario dengan variasi objek yang lebih kompleks. Dengan demikian, penelitian ini membuktikan bahwa kedua arsitektur dapat dioperasikan secara efektif pada edge device dengan daya terbatas, meskipun terdapat trade-off antara akurasi dan efisiensi.

DAFTAR PUSTAKA

[1] M. Nurul Achmadiah, N. Setyawan, and A. D. Risdhayanti, "Deteksi Kepadatan Objek di Stasiun Kereta Api Berbasis ViT-Base pada Jetson Orin Nano," *Jurnal Elektronika dan Otomasi Industri*, vol. 12, no. 1, pp. 153–161, May 2025, doi: 10.33795/elkolind.v12i1.7495.

[2] 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.

[3] 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.

[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] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020.

[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. Pan *et al.*, "EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision Transformers," 2022, pp. 294–311. doi: 10.1007/978-3-031-20083-0_18.

[11] N. Setyawan, M. N. Achmadiah, C.-C. Sun, and W.-K. Kuo, "Multi-Stage Vision Transformer for Batik Classification," in *2024 International Electronics Symposium (IES)*, IEEE, Aug. 2024, pp. 449–453. doi: 10.1109/IES63037.2024.10665807.