ISSN (Online) 2747-0563
Volume 5 No. 2 Tahun 2025

Arsitektur U-Net MobileNetV2 yang Efisien dan Akurat untuk Segmentasi Buah Multi-Kelas

Muhammad Adeva

Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
22081010077@student.upnjatim.ac.id

***Abstrak*— Sistem pertanian presisi modern, seperti robot pemanen, sangat bergantung pada kemampuan *computer vision* untuk segmentasi buah secara *real-time*. Identifikasi objek sederhana tidaklah cukup; sistem ini memerlukan pemahaman spasial tingkat piksel untuk lokalisasi yang akurat, yang hanya dapat disediakan oleh segmentasi semantik. Namun, arsitektur segmentasi semantik yang umum seperti U-Net seringkali terlalu berat secara komputasi untuk aplikasi di perangkat *edge*. Penelitian ini mengusulkan dan mengevaluasi arsitektur U-Net yang dioptimalkan, menggunakan *backbone* MobileNetV2 *pre-trained* [1] untuk mencapai keseimbangan antara akurasi dan efisiensi. Model ini dilatih pada dataset FruitSeg30, yang terdiri dari 30 kelas buah. Sebuah langkah pra-pemrosesan krusial dilakukan untuk mengonversi *mask* biner per-kelas dari dataset asli menjadi *mask* multi-kelas tunggal (0-30). Model menerima input gambar berukuran 256x256 piksel. Untuk mengatasi *class imbalance*, kami menggunakan *hybrid loss function* yang menggabungkan *Categorical Cross-Entropy* dan *Dice Loss* [2], serta dilatih menggunakan optimizer Adam. Hasil eksperimen menunjukkan kinerja yang luar biasa: model mencapai akurasi Dice Coefficient 97,84% dengan jumlah parameter yang ringan, yaitu 11,3 Juta. Lebih penting lagi, model ini mencapai waktu inferensi rata-rata 19,80 ms per gambar (~50,5 FPS). Kinerja ini membuktikan kelayakannya untuk aplikasi segmentasi *real-time* di lapangan dan menawarkan solusi praktis yang menjembatani kesenjangan antara akurasi tinggi dan kebutuhan komputasi yang rendah pada perangkat *edge*.**

Kata Kunci— **Segmentasi Semantik, U-Net, MobileNetV2, Segmentasi Buah, *Real-Time*.**

I. PENDAHULUAN

Industri agrikultur modern semakin bergerak menuju otomatisasi dan *Pertanian Presisi (Precision Agriculture)* untuk meningkatkan efisiensi, mengurangi biaya tenaga kerja, dan mengoptimalkan hasil panen. Dalam ranah ini, sistem robotik canggih, seperti lengan pemanen otomatis dan *drone* pemantau, memegang peranan krusial. Keberhasilan sistem-sistem ini sangat bergantung pada kemampuan mereka untuk "melihat" dan menginterpretasi lingkungan secara akurat, di mana tugas identifikasi dan lokalisasi buah secara *real-time* menjadi tantangan fundamental. [1]

Untuk mengatasi tugas ini, teknologi *Computer Vision*, khususnya *Deep Learning*, telah menunjukkan kinerja yang luar biasa. Berbeda dari klasifikasi citra sederhana yang hanya memberi label pada seluruh gambar (misal, 'ada apel'), sistem robotik memerlukan pemahaman spasial yang mendetail. Di sinilah teknik **Segmentasi Semantik** menjadi sangat relevan . Segmentasi semantik memberikan klasifikasi pada tingkat piksel, yang memungkinkan sistem untuk secara presisi mengetahui *di mana* letak setiap buah dan membedakannya dari latar belakang (daun, cabang, atau buah lainnya) [2][3].

Meskipun arsitektur canggih seperti U-Net telah menjadi standar untuk segmentasi presisi , implementasinya seringkali menghadapi tantangan komputasi. Arsitektur U-Net standar, atau yang menggunakan *encoder* berat (seperti VGG16 atau ResNet), memiliki puluhan juta parameter. Model yang besar ini memerlukan sumber daya komputasi tinggi, membuatnya lambat dan tidak praktis untuk diimplementasikan pada perangkat *on-device* atau *edge computing* yang umumnya digunakan oleh robot di lapangan [4][5].

Berdasarkan tantangan tersebut, terdapat kebutuhan mendesak akan sebuah model segmentasi yang tidak hanya akurat, tetapi juga efisien secara komputasi (ringan dan cepat). Oleh karena itu, penelitian ini merancang, mengimplementasikan, dan mengevaluasi kinerja arsitektur U-Net yang dioptimalkan untuk tugas segmentasi buah multi-kelas. Kami mengusulkan penggunaan *backbone* MobileNetV2 *pre-trained* untuk menggantikan *encoder* U-Net standar. Penelitian ini bertujuan untuk membuktikan secara kuantitatif bahwa arsitektur yang diusulkan mampu mencapai keseimbangan optimal antara akurasi segmentasi yang tinggi dan efisiensi komputasi (jumlah parameter rendah dan waktu inferensi cepat) untuk aplikasi *real-time* [6][7].

II. METODOLOGI

Bab ini menguraikan langkah-langkah metodologis yang digunakan dalam penelitian, mulai dari akuisisi dan pra-pemrosesan data, perancangan arsitektur model, hingga skenario pelatihan dan evaluasi.

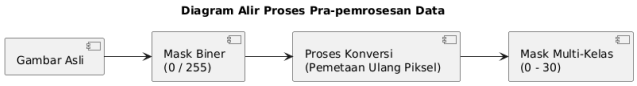
A. Dataset dan Pra-pemrosesan Data

Penelitian ini menggunakan dataset publik "FruitSeg30_Segmentation Dataset & Mask Annotations". Dataset mentah ini terdiri dari 30 kelas buah, di mana setiap kelas memiliki gambar dan *mask* segmentasi biner (nilai piksel 0 untuk latar belakang dan 255 untuk objek buah).

Karena tujuan penelitian adalah segmentasi multi-kelas, proses pra-pemrosesan krusial dilakukan. Kami mengonversi 30 set data biner ini menjadi satu set data tunggal dengan masker segmentasi multi-kelas. Nilai piksel pada *mask* baru dipetakan ulang di mana 0 merepresentasikan 'Latar Belakang', dan nilai 1 hingga 30 merepresentasikan masing-

ISSN (Online) 2747-0563
Volume 5 No. 2 Tahun 2025

masing kelas buah (misal, 'Apple_Gala' = 1, 'Avocado' = 2, dst.). Proses konversi data ini diilustrasikan pada Gambar 1.

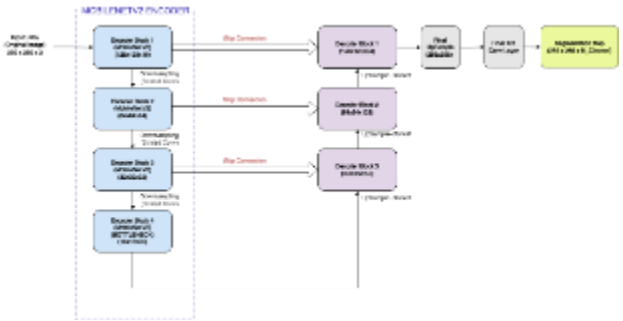


Gambar 1. Konversi Data

Seluruh gambar dan mask kemudian distandarisasi ukurannya (resized) menjadi resolusi input 256x256 piksel. Nilai piksel pada gambar RGB juga dinormalisasi dari rentang [0, 255] ke rentang [0, 1].

B. Arsitektur Model

Arsitektur yang diusulkan didasarkan pada model U-Net, yang terdiri dari jalur *encoder* (kontraksi) dan *decoder* (ekspansi) simetris . Untuk mencapai efisiensi komputasi, penelitian ini mengadopsi pendekatan *transfer learning* dengan mengganti jalur *encoder* U-Net standar menggunakan arsitektur **MobileNetV2** *pre-trained.*



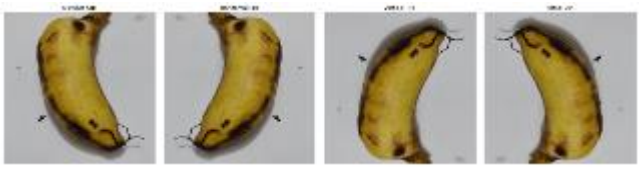
Gambar 2. Arsitektur U-Net dengan *Encoder* MobileNetV2

Bobot *encoder* MobileNetV2 diinisialisasi menggunakan bobot dari ImageNet untuk mempercepat konvergensi. *Skip connections* (Gambar 2, garis penghubung) diambil dari keluaran blok-blok relevan di MobileNetV2 dan digabungkan (concatenated) dengan jalur *decoder* untuk mempertahankan informasi spasial beresolusi tinggi, yang krusial untuk lokalisasi batas objek.

Lapisan keluaran akhir model menggunakan konvolusi 1x1 dengan jumlah filter diatur ke **31** (sesuai dengan 30 kelas buah + 1 kelas latar belakang). Fungsi aktivasi **Softmax** diterapkan pada lapisan ini untuk menghasilkan prediksi probabilitas per-piksel untuk setiap kelas.

C. Skenario Pelatihan dan Implementasi

Dataset yang telah diproses (total 1969 gambar) dibagi menjadi 80% data latih (1575 gambar) dan 20% data validasi (394 gambar). Augmentasi data *on-the-fly* diterapkan hanya pada data latih untuk meningkatkan variasi dan generalisasi model. Transformasi augmentasi geometris yang digunakan mencakup **Horizontal Flip** (p=0.5) , **Vertical Flip** (p=0.5), dan **Random Rotate 90°** (p=0.5). Contoh hasil augmentasi visual ditunjukkan pada Gambar 3.



Gambar 3. Hasil Augmentasi

Untuk menangani potensi ketidakseimbangan kelas (*class imbalance*), penelitian ini mengadopsi Hybrid Loss (Combo Loss). Fungsi *loss* ini merupakan penjumlahan berbobot 50:50 dari *Categorical Cross-Entropy (CCE)* dan *Multi-class Dice Loss* .

Model dilatih menggunakan optimizer Adam dengan *learning rate* (LR) awal $1x10^{-4}$. Ukuran *batch* ditetapkan pada 16. Dua *callback* digunakan: *ReduceLROnPlateau* dan *EarlyStopping* untuk menghentikan pelatihan dan menyimpan bobot model terbaik.

III. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari skenario pelatihan dan evaluasi yang telah diuraikan pada bab metodologi. Evaluasi dibagi menjadi tiga bagian: metrik performa kuantitatif (akurasi), metrik efisiensi (komputasi), dan analisis kualitatif (visual).

A. Hasil Kinerja Model (Kuantitatif)

Model dilatih hingga mencapai konvergensi, yang secara otomatis dihentikan oleh *callback* EarlyStopping untuk mencegah *overfitting* dan memastikan bobot model terbaik yang disimpan .

Evaluasi akhir dilakukan pada 394 gambar di *validation set* yang belum pernah dilihat oleh model selama pelatihan. Model yang diusulkan (U-Net + MobileNetV2) menunjukkan kinerja yang sangat baik. Metrik evaluasi utama, yaitu Dice Coefficient (yang mengukur tumpang tindih antara prediksi dan *ground truth*), mencapai nilai 97,84%.

Hasil ini mengindikasikan bahwa model memiliki kemampuan yang sangat tinggi dalam melokalisasi dan mempartisi piksel buah dari latar belakang secara akurat.

B. Hasil Efisiensi Komputasi

Tujuan utama dari penelitian ini adalah untuk memvalidasi arsitektur yang efisien. Dua metrik efisiensi utama diukur: jumlah parameter model dan waktu inferensi.

ISSN (Online) 2747-0563
Volume 5 No. 2 Tahun 2025

Seluruh skenario pelatihan dan evaluasi model dilakukan pada platform Google Colaboratory yang dilengkapi dengan GPU NVIDIA T4 (VRAM 16GB). Ukuran *batch* 16 dipilih untuk memaksimalkan penggunaan memori GPU tanpa menyebabkan *error out-of-memory*. Pelatihan diatur untuk 100 *epoch* namun dihentikan secara otomatis oleh *callback* EarlyStopping pada epoch ke-59, karena metrik *validation loss* tidak lagi menunjukkan perbaikan. Total waktu pelatihan yang tercatat adalah 4107 detik (sekitar 68,5 menit). Bobot model terbaik yang digunakan untuk evaluasi akhir adalah bobot yang disimpan dari epoch ke-49.

TABEL I
HASIL METRIK KERJA DAN EFISIENSI MODEL

Kategori	Metrik	Nilai
Akurasi	Validation Dice Coefficient	97,84%
Akurasi	Validation Loss	0,0367
Efisiensi	Jumlah Parameter	11.346.687
Efisiensi	Waktu Inferensi (per Gambar)	19,80 ms
Efisiensi	Throughput (FPS)	~50,5

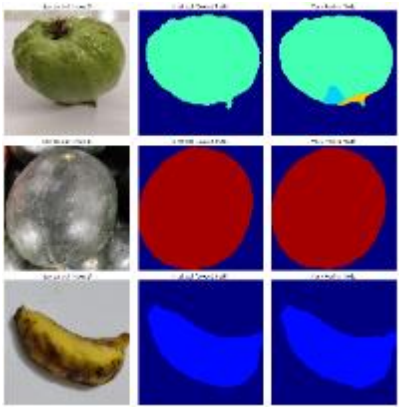
Hasil pengukuran disajikan pada Tabel 1. Model yang diusulkan hanya memiliki 11,3 Juta parameter. Angka ini secara signifikan lebih ringan jika dibandingkan dengan arsitektur U-Net standar yang menggunakan *encoder* berat seperti VGG16.

Lebih penting lagi, waktu inferensi rata-rata per gambar (resolusi 256x256) adalah 19,80 ms. Ini membuktikan bahwa model mampu memproses sekitar 50,5 *Frames Per Second* (FPS) (1000 ms / 19,80 ms), yang sepenuhnya memenuhi syarat untuk aplikasi *real-time*.

C. *Analisis Kualitatif (Visual)*

Selain metrik kuantitatif, analisis visual dilakukan dengan membandingkan hasil prediksi model dengan *mask ground truth* seperti yang ditunjukkan di gambar 4.

Secara visual, hasil prediksi model menunjukkan kualitas segmentasi yang sangat tinggi. Batas-batas (tepi) dari setiap buah dapat diprediksi dengan presisi dan hampir identik dengan *mask ground truth*. Model juga terbukti *robust* dalam memisahkan beberapa buah yang saling berdekatan atau tumpang tindih, serta mengabaikan area latar belakang yang kompleks. Hasil kualitatif ini mengkonfirmasi keakuratan kuantitatif yang ditunjukkan oleh skor *Dice Coefficient*.



Gambar 4. Komparasi Gambar Asli, *Ground Truth,* dan Hasil Prediksi

D. *Pembahasan*

Hasil penelitian ini secara komprehensif membuktikan hipotesis awal. Arsitektur U-Net dengan *encoder* MobileNetV2 berhasil mencapai dua tujuan sekaligus:

1. **Akurasi Tinggi:** Skor Dice 97,84% menunjukkan bahwa penggantian *encoder* dengan arsitektur ringan tidak mengorbankan kemampuan model untuk belajar representasi fitur yang kompleks untuk segmentasi.
2. **Efisiensi Tinggi:** Dengan 11,3 Juta parameter dan kecepatan inferensi 50,5 FPS, model ini terbukti efisien dan praktis untuk implementasi di perangkat dengan sumber daya terbatas (seperti perangkat *edge* atau robot pemanen), yang merupakan tantangan utama yang diuraikan dalam pendahuluan.

IV. KESIMPULAN

Penelitian ini telah berhasil merancang, mengimplementasikan, dan mengevaluasi sebuah arsitektur U-Net yang dioptimalkan untuk tugas segmentasi buah multi-kelas. Dengan mengadopsi pendekatan *transfer learning* dan mengganti *encoder* U-Net standar dengan *backbone* MobileNetV2 , penelitian ini berhasil menjawab tantangan utama yaitu menciptakan model yang akurat sekaligus efisien. Hasil evaluasi kuantitatif menunjukkan bahwa arsitektur yang diusulkan mampu mencapai tingkat akurasi segmentasi yang sangat tinggi, dibuktikan dengan perolehan skor Dice Coefficient 97,84% pada *validation set*. Selain itu, model ini terbukti sangat efisien secara komputasi, dengan total hanya 11,3 Juta parameter. Efisiensi ini dikonfirmasi lebih lanjut oleh waktu inferensi 19,80 ms per gambar, yang setara dengan throughput ~50,5 FPS, membuktikan kemampuannya untuk beroperasi dalam skenario *real-time*. Temuan ini menunjukkan bahwa model yang diusulkan adalah kandidat yang kuat dan praktis untuk implementasi di dunia nyata, khususnya untuk sistem pertanian presisi, robot pemanen, atau

ISSN (Online) 2747-0563
Volume 5 No. 2 Tahun 2025

aplikasi *edge computing* lain yang memiliki keterbatasan sumber daya komputasi.

UCAPAN TERIMA KASIH

Penelitian ini dapat dilaksanakan dengan baik berkat bantuan dari berbagai pihak yang turut serta membantu dalam penyelesaian jurnal ini, Dan terimakasih juga atas rekan team pembuatan jurnal ini sehingga jurnal ini dapat terselesaikan secara tepat waktu.

REFERENSI

[1] P. Maheswari, P. Raja, O. E. Apolo-Apolo, and M. Pérez-Ruiz, "Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—A review," *Frontiers in Plant Science*, vol. 12, p. 684328, 2021.

[2] Z. Xie, Z. Ke, K. Chen, Y. Wang, Y. Tang, and W. Wang, "A lightweight deep learning semantic segmentation model for optical-image-based post-harvest fruit ripeness analysis of sugar apples (Annona squamosa)," *Agriculture*, vol. 14, no. 4, p. 591, 2024.

[3] I. Ulku, "ContexNestedU-Net: Efficient context-aware semantic segmentation architecture for precision agriculture applications based on multispectral remote sensing imagery," *Technical Sciences*, vol. 41, no. 5, 2024.

[4] H. Song, Y. Shang, and D. He, "Review on deep learning technology for fruit target recognition," *Transactions of the Chinese Society of Agricultural Machinery*, vol. 54, no. 1, 2023.

[5] B. C. Bag, H. K. Maity, and C. Koley, "UNet MobileNetV2: Medical image segmentation using deep neural network (DNN)," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 18, no. 1, pp. 21–29, Jan. 2023.

[6] Y. Xiao, H. Wang, Y. Xu, and R. Zhang, "Fruit detection and recognition based on deep learning for automatic harvesting: An overview and review," *Agronomy*, vol. 13, no. 6, p. 1625, 2023.

[7] "Evaluation of the effectiveness of the UNet model with different backbones in the semantic segmentation of tomato leaves and fruits," *Horticulturae*, 2023.