Copyright © The Author(s) 2026
Jurnal Artificial, Vol 4 No 2, Oktober 2026

Implementasi Model Hibrida RT-DETR dan Vision Transformer untuk Deteksi dan Klasifikasi Produk Retail Berdasarkan Citra Kemasan

Putri Munawaroh [1*], Harma Oktafia Lingga Wijaya [2], Davit Irawan [3]

[1, 2, 3] Universitas Bina Insan, Indonesia

* putrimunawaroh903@gmail.com

Abstrak

Identifikasi produk retail pada kemasan merupakan salah satu penerapan *computer vision* yang dapat mendukung proses pengenalan produk secara otomatis. Penelitian ini bertujuan mengimplementasikan model hibrida Real-Time DEtection TRansformer (RT-DETR) dan Vision Transformer (ViT) untuk mendeteksi dan mengklasifikasikan produk retail berdasarkan citra kemasan. Penelitian menggunakan pendekatan eksperimental dengan dataset yang dikumpulkan secara langsung melalui dokumentasi citra kemasan produk retail. Dataset terdiri atas 1.202 citra dari 15 kelas produk, yang dibagi menjadi data pelatihan, validasi, dan pengujian dengan rasio 80:10:10. RT-DETR digunakan sebagai model deteksi untuk menentukan lokasi objek produk melalui *bounding box*, sedangkan ViT digunakan untuk mengklasifikasikan hasil *cropping* objek ke dalam kelas produk. Model RT-DETR-L dilatih selama 36 *epoch* dengan ukuran citra 640 × 640 piksel dan *batch size* 8, sedangkan ViT-Base Patch16-224 dilatih selama 6 *epoch* menggunakan citra berukuran 224 × 224 piksel. Evaluasi RT-DETR dilakukan menggunakan metrik *mean Average Precision* (mAP), *precision*, dan *recall*, sedangkan kinerja klasifikasi hibrida dievaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian menunjukkan bahwa RT-DETR mampu mendeteksi objek produk dengan nilai *precision* dan *recall* sekitar 0,90 serta mAP@50 sekitar 0,93–0,95. Pada tahap klasifikasi, ViT mencapai akurasi validasi sebesar 99,2% pada akhir pelatihan. Berdasarkan evaluasi rinci sistem hibrida, diperoleh nilai accuracy, precision, recall, dan F1-score masing-masing sebesar 100%. Hasil tersebut menunjukkan bahwa kombinasi RT-DETR sebagai pendeteksi objek dan ViT sebagai pengklasifikasi mampu menjalankan proses deteksi dan klasifikasi produk retail pada dataset penelitian dengan kinerja yang sangat baik.

Kata Kunci: *Citra Kemasan, Deteksi dan Klasifikasi Produk, Model Hibrida, RT-DETR, Vision Transformer*

Pendahuluan

Perkembangan *computer vision* dan *deep learning* telah mendorong pemanfaatan teknologi pengolahan citra untuk melakukan identifikasi objek secara otomatis pada berbagai bidang, termasuk sektor retail. Pengenalan produk berdasarkan citra menjadi salah satu permasalahan penting dalam *computer vision* karena dapat mendukung proses identifikasi produk secara lebih cepat dibandingkan pengenalan secara manual. Pada lingkungan retail, sistem pengenalan produk berbasis citra dapat digunakan untuk membantu proses pengelolaan produk, pemantauan rak, pengenalan barang, serta berbagai kebutuhan otomatisasi lainnya (Yanti et al., 2024). Namun, pengenalan produk retail memiliki tantangan tersendiri karena produk dapat memiliki kemiripan bentuk, warna, ukuran, desain kemasan, maupun karakteristik visual lainnya. Selain itu, perubahan posisi objek, kondisi pencahayaan, dan variasi tampilan produk dapat memengaruhi

https://doi.org/10.54065/artificial.4.2.2026.1982

kemampuan model dalam mengenali suatu produk secara tepat (Chen et al., 2022; Wei et al., 2020).

Penerapan *deep learning* pada pengenalan produk retail umumnya melibatkan dua permasalahan utama, yaitu deteksi objek dan klasifikasi citra. Deteksi objek bertujuan menentukan lokasi objek pada citra melalui *bounding box*, sedangkan klasifikasi bertujuan menentukan kelas dari objek berdasarkan karakteristik visual yang dipelajari oleh model (Zaidi et al., 2022; Gothai et al., 2022). Penelitian mengenai pengenalan produk kemasan menunjukkan bahwa pendekatan berbasis *deep learning* telah banyak digunakan untuk mendeteksi dan mengenali produk pada lingkungan retail. Studi terdahulu menunjukkan bahwa pendekatan *computer vision* dapat digunakan untuk melakukan deteksi dan klasifikasi produk kemasan pada lingkungan retail (Fuchs et al., 2019). Sementara itu, studi lainnya mengembangkan kerangka pengenalan produk grocery yang memisahkan proses deteksi produk dengan proses pengenalan informasi visual produk (Selvam & Koilraj, 2022). Pendekatan tersebut menunjukkan bahwa pemisahan proses lokalisasi dan pengenalan dapat digunakan untuk menangani permasalahan identifikasi produk secara bertahap.

Perkembangan metode deteksi objek selanjutnya mengalami perubahan dari pendekatan berbasis *Convolutional Neural Network* (CNN) menuju pendekatan berbasis *Transformer* (Shah & Tembhurne, 2023). Transformer memiliki kemampuan dalam memodelkan hubungan antarelemen visual melalui mekanisme *self-attention*, sehingga dapat digunakan untuk mempelajari informasi visual secara lebih luas. Perkembangan tersebut menghasilkan berbagai metode deteksi berbasis Transformer yang berusaha memperoleh keseimbangan antara akurasi dan efisiensi komputasi. Kajian mengenai perkembangan deteksi objek menunjukkan bahwa Transformer telah menjadi salah satu pendekatan penting dalam perkembangan metode deteksi objek modern, baik melalui pengem-bangan DETR maupun berbagai turunannya (Amjoud & Amrouch, 2023; Arkin et al., 2023).

Salah satu perkembangan penting dalam deteksi objek berbasis Transformer adalah *DEtection TRansformer* (DETR), yang memperkenalkan pendekatan *end-to-end object detection* melalui formulasi deteksi sebagai permasalahan *set prediction*. Perkembangan DETR kemudian menghasilkan berbagai arsitektur yang ditujukan untuk mengatasi keterbatasan model awal, khususnya dalam hal kecepatan konvergensi dan efisiensi deteksi. Salah satu pengembangan tersebut adalah Real-Time DEtection TRansformer (RT-DETR). Studi terdahulu mengembangkan RT-DETR sebagai detektor objek berbasis Transformer yang menggunakan *efficient hybrid encoder* dan mekanisme *query selection* untuk menghasilkan deteksi yang efisien (Zhao et al., 2024). Karakteristik tersebut menjadikan RT-DETR relevan untuk digunakan pada permasalahan deteksi objek yang membutuhkan kemampuan lokalisasi objek secara efektif.

Penggunaan RT-DETR pada objek produk retail juga telah dilakukan pada penelitian sebelumnya. Studi terdahulu menerapkan kombinasi YOLOv8 dan RT-DETR untuk mendeteksi produk retail dengan mempertimbangkan variasi karakteristik produk, seperti ukuran, warna, dan varian (Maulana et al., 2024). Penelitian tersebut menggunakan beberapa dataset produk retail dengan kondisi citra yang berbeda dan menunjukkan bahwa pendekatan berbasis RT-DETR dapat digunakan untuk mendukung proses deteksi produk retail (Melek et al., Pietrini et al., 2024). Penelitian lain mengenai pengenalan produk grocery juga menunjukkan bahwa sistem dua tahap yang

memisahkan deteksi produk dan pengenalan produk dapat digunakan untuk memperoleh informasi identitas produk dari citra (Gothai et al., 2022; Selvam & Koilraj, 2022). Hasil penelitian tersebut memperlihatkan bahwa karakteristik visual produk dan kondisi citra merupakan aspek penting dalam pengembangan sistem pengenalan produk retail.

Selain perkembangan pada deteksi objek, *Vision Transformer* (ViT) memberikan alternatif pendekatan untuk melakukan klasifikasi citra. Berbeda dengan CNN yang secara dominan menggunakan operasi konvolusi untuk memperoleh representasi fitur, ViT membagi citra menjadi sejumlah *patch* dan memprosesnya menggunakan mekanisme Transformer untuk mempelajari hubungan antarbagiannya. Pendekatan berbasis Vision Transformer telah digunakan pada berbagai permasalahan klasifikasi citra dan menunjukkan kemampuan dalam mempelajari representasi visual. Penelitian mengenai penerapan Vision Transformer pada klasifikasi citra juga menunjukkan bahwa arsitektur tersebut menjadi salah satu perkembangan penting dalam klasifikasi visual berbasis Transformer (Gothai et al., 2022). Dengan karakteristik tersebut, ViT dapat digunakan sebagai model klasifikasi terhadap area objek yang telah diperoleh dari proses deteksi.

Pada produk retail, informasi visual yang terdapat pada kemasan dapat menjadi karakteristik penting dalam membedakan suatu produk dengan produk lainnya. Logo, tulisan, warna, bentuk, dan elemen desain kemasan dapat menjadi bagian dari informasi visual yang dipelajari oleh model. Penelitian mengenai logo pada kemasan menunjukkan bahwa *deep learning* dapat digunakan untuk mendeteksi lokasi logo dan menganalisis karakteristik visual merek pada kemasan (Hosseini et al., 2025). Sementara itu, penelitian mengenai deteksi logo menunjukkan bahwa perkembangan *deep learning* telah menghasilkan berbagai pendekatan untuk mendeteksi informasi merek dalam citra (Hou et al., 2024). Namun, deteksi logo atau merek secara langsung merupakan permasalahan yang berbeda dengan klasifikasi kelas produk. Dalam penelitian ini, RT-DETR tidak digunakan untuk mendeteksi merek atau logo secara terpisah, melainkan digunakan untuk mendeteksi lokasi objek produk, sedangkan karakteristik visual pada area produk tersebut selanjutnya digunakan oleh ViT untuk menentukan kelas produk.

Karakteristik dataset juga menjadi faktor penting dalam pengembangan sistem pengenalan produk berbasis *deep learning*. Dataset produk retail dapat memiliki perbedaan dalam jumlah kelas, jumlah citra, kondisi pencahayaan, sudut pengambilan gambar, jarak objek, posisi produk, serta variasi kemasan. Penelitian mengenai pengenalan produk grocery menunjukkan bahwa keterbatasan jumlah data dan variasi sudut pengambilan gambar dapat memengaruhi kemampuan model dalam mengenali produk (Chen et al., 2022). Oleh karena itu, proses pengumpulan dan dokumentasi dataset perlu disesuaikan dengan karakteristik permasalahan yang diteliti. Dokumentasi dataset juga merupakan bagian penting dalam pengembangan sistem pembelajaran mesin karena karakteristik data, proses pengumpulan, dan tujuan penggunaannya dapat memengaruhi interpretasi terhadap hasil penelitian (Gebru et al., 2021; Kuznetsova et al., 2020).

Berdasarkan penelitian terdahulu, telah terdapat penelitian mengenai deteksi produk retail menggunakan *deep learning*, pengembangan RT-DETR untuk deteksi objek, serta penggunaan Vision Transformer untuk klasifikasi citra. Penelitian mengenai pengenalan produk retail juga telah menerapkan pendekatan bertahap yang memisahkan proses deteksi dan pengenalan produk (Fuchs et al., 2019; Selvam & Koilraj, 2022). Namun, berdasarkan telaah terhadap penelitian tersebut, penerapan RT-DETR sebagai pendeteksi

objek produk yang dilanjutkan dengan Vision Transformer sebagai pengklasifikasi terhadap area objek hasil deteksi pada dataset produk retail lokal masih memiliki ruang untuk dikaji lebih lanjut. Pendekatan tersebut memungkinkan proses deteksi dan klasifikasi dilakukan secara berurutan dengan fungsi model yang berbeda, yaitu RT-DETR bertugas menentukan lokasi objek, sedangkan ViT bertugas menentukan kelas produk berdasarkan karakteristik visual area objek.

Berdasarkan permasalahan tersebut, penelitian ini mengimplementasikan model hibrida RT-DETR–Vision Transformer untuk deteksi dan klasifikasi produk retail pada citra kemasan. Dataset penelitian diperoleh melalui dokumentasi langsung citra produk retail dan terdiri atas 1.202 citra yang terbagi ke dalam 15 kelas produk. Citra dikumpulkan dengan mempertimbangkan variasi pencahayaan, sudut kamera, jarak pengambilan, posisi, dan orientasi kemasan. RT-DETR digunakan pada tahap pertama untuk mendeteksi lokasi objek produk dan menghasilkan *bounding box*, kemudian area objek hasil deteksi dilakukan *cropping* dan digunakan sebagai masukan bagi Vision Transformer untuk menentukan kelas produk. Alur tersebut sesuai dengan implementasi sistem pada penelitian, yaitu RT-DETR menghasilkan lokasi objek dan nilai *confidence*, kemudian hasil *cropping* diproses oleh ViT untuk menghasilkan prediksi kelas produk.

Dengan demikian, penelitian ini berfokus pada penerapan dan evaluasi integrasi RT-DETR dan Vision Transformer dalam satu alur deteksi dan klasifikasi produk retail. Evaluasi dilakukan untuk mengetahui kemampuan RT-DETR dalam mendeteksi objek produk serta kemampuan ViT dan sistem hibrida dalam mengklasifikasikan produk berdasarkan kelas yang telah ditentukan. Pendekatan ini diharapkan dapat memberikan penerapan model berbasis Transformer untuk pengenalan produk retail pada dataset lokal yang dikembangkan secara khusus sesuai dengan objek penelitian.

Metode

Penelitian ini menggunakan pendekatan eksperimental untuk mengimplementasikan model *hibrida* RT-DETR–Vision Transformer (ViT) dalam proses deteksi dan klasifikasi produk retail berdasarkan citra kemasan. Sistem dirancang dalam dua tahap utama, yaitu tahap deteksi objek menggunakan RT-DETR dan tahap klasifikasi menggunakan ViT. RT-DETR digunakan untuk menentukan lokasi objek produk pada citra melalui *bounding box*, sedangkan ViT digunakan untuk mengklasifikasikan area objek hasil *cropping* ke dalam kelas produk. Tahapan penelitian meliputi pengumpulan dan seleksi citra, anotasi objek, pembagian dataset, pelatihan RT-DETR, pembentukan citra *cropping*, pelatihan ViT, integrasi kedua model, serta pengujian dan evaluasi sistem. Rangkaian tahapan tersebut disusun untuk menghasilkan alur pemrosesan citra yang konsisten dari tahap deteksi hingga klasifikasi.

Pengumpulan dan Karakteristik Dataset

Dataset penelitian diperoleh melalui pengambilan citra secara langsung terhadap kemasan produk retail. Citra diambil menggunakan kamera Samsung Galaxy A13 dengan resolusi kamera 64 MP. Pengambilan citra dilakukan dengan mempertimbangkan variasi kondisi pencahayaan, sudut pengambilan gambar, jarak kamera terhadap objek, serta posisi dan orientasi kemasan. Variasi tersebut digunakan untuk merepresentasikan kondisi citra produk yang berbeda selama proses pengembangan dan pengujian sistem.

Citra yang memiliki kualitas sangat rendah, buram, terpotong, atau tidak dapat digunakan untuk proses anotasi dan identifikasi objek dikeluarkan dari dataset. Dataset akhir terdiri atas 1.202 citra yang dikelompokkan ke dalam 15 kelas produk, yaitu Kacang Rosta, Entrasol, Cimory, Ultramilk, Baking Powder, Chitato, Silverqueen, Tango, Tongtji, Kapal Api, Kispray, Le Minerale, Soklin, Tolak Angin, dan Montiss.

Tabel 1. *Distribusi Dataset Produk*

No.	Kelas Produk	Jumlah
1	Kacang Rosta	72
2	Entrasol	90
3	Cimory	76
4	Ultramilk	83
5	Baking Powder	87
6	Chitato	71
7	Silverqueen	75
8	Tango	81
9	Tongtji	90
10	Kapal Api	68
11	Kispray	83
12	Le Minerale	81
13	Soklin	80
14	Tolak Angin	82
15	Montiss	83
	Total	1.202

Anotasi Citra

Proses anotasi dilakukan menggunakan LabelImg dengan format anotasi YOLO. Setiap citra diberi anotasi berupa *bounding box* yang menunjukkan lokasi objek produk serta identitas kelas objek. Hasil anotasi digunakan sebagai informasi acuan dalam proses pelatihan model RT-DETR. Anotasi dilakukan pada objek produk yang menjadi target penelitian, sehingga RT-DETR berfungsi untuk menentukan lokasi objek produk pada citra. Dengan demikian, label kelas pada anotasi merepresentasikan kelas produk, bukan anotasi terpisah untuk elemen logo atau merek pada kemasan. Format YOLO menyimpan informasi kelas dan koordinat *bounding box* yang selanjutnya digunakan oleh model deteksi.

Pembagian Dataset

Dataset dibagi menjadi tiga bagian, yaitu data pelatihan, validasi, dan pengujian dengan rasio 80:10:10. Dari total 1.202 citra, sebanyak 961 citra digunakan sebagai data pelatihan, 120 citra sebagai data validasi, dan 121 citra sebagai data pengujian. Data pelatihan digunakan untuk membentuk parameter model, sedangkan data validasi digunakan untuk memantau kinerja model selama proses pelatihan. Data pengujian digunakan pada tahap akhir untuk mengevaluasi kemampuan model terhadap citra yang tidak digunakan dalam proses pelatihan.

Tabel 2. *Pembagian Dataset*

Data	Rasio	Jumlah Citra
Pelatihan	80%	961
Validasi	10%	120
Pengujian	10%	121
Total	100%	1.202

Pelatihan Model RT-DETR

RT-DETR digunakan sebagai model deteksi objek untuk menentukan lokasi produk pada citra. RT-DETR merupakan pendekatan *end-to-end object detection* berbasis Transformer yang dirancang untuk melakukan deteksi objek tanpa memerlukan tahapan *non-maximum suppression* pada proses inferensi. Arsitektur RT-DETR digunakan pada penelitian ini sebagai tahap awal sebelum citra objek diproses oleh model klasifikasi (Zhao et al., 2024).

Model yang digunakan adalah RT-DETR-L dengan jumlah kelas sebanyak 15 kelas. Proses pelatihan dilakukan menggunakan framework Ultralytics dengan ukuran citra masukan 640 × 640 piksel. Model dilatih selama 36 *epoch* dengan *batch size* 8 dan menggunakan perangkat GPU.

Tabel 3. *Konfigurasi Pelatihan RT-DETR*

Parameter	Nilai
Model	RT-DETR-L
Jumlah kelas	15
Ukuran citra	640 × 640 piksel
Epoch	36
Batch size	8
Worker	2
Perangkat	GPU (Device 0)
Format Anotasi	YOLO
Framework	Ultralytics

Selama proses pelatihan, data anotasi digunakan sebagai acuan untuk mempelajari lokasi dan kelas objek pada citra. Model terbaik yang diperoleh dari proses pelatihan selanjutnya digunakan *sebagai* model deteksi pada tahap pembentukan *cropping* dan pengujian sistem *hibrida*.

Pembentukan Citra Cropping

Hasil anotasi *ground truth* digunakan untuk membentuk citra *cropping* sebagai data masukan pada tahap pelatihan ViT. Proses *cropping* dilakukan dengan mengambil area citra yang berada di dalam koordinat *bounding box* objek berdasarkan anotasi YOLO. Citra hasil *cropping* kemudian disimpan berdasarkan kelas produk masing-masing.

Tahap ini menghasilkan citra yang berfokus pada area objek produk sehingga dapat digunakan sebagai masukan untuk model klasifikasi. Perlu dibedakan antara proses *cropping* pada tahap pelatihan ViT dan tahap pengujian sistem *hibrida*. Pada pelatihan ViT, *cropping* dilakukan berdasarkan *ground-truth bounding box* dari anotasi dataset. Sementara itu, pada pengujian *hibrida*, area *cropping* diperoleh dari *bounding box* yang diprediksi oleh RT-DETR.

Pelatihan Vision Transformer

Vision Transformer digunakan sebagai model klasifikasi untuk menentukan kelas produk berdasarkan citra hasil *cropping*. ViT memproses citra sebagai rangkaian *patch* dan menggunakan mekanisme *self-attention* untuk mempelajari representasi visual. Pendekatan tersebut memungkinkan Transformer digunakan secara langsung untuk tugas klasifikasi citra (Dosovitskiy et al., 2020).

Arsitektur yang digunakan adalah ViT-Base Patch16-224. Citra hasil *cropping* diubah ukurannya menjadi 224 × 224 piksel dan dinormalisasi menggunakan parameter

normalisasi ImageNet sebelum diberikan sebagai masukan model. Proses pelatihan dilakukan selama 6 *epoch* dengan *batch size* 116, *learning rate* sebesar 0,0001, optimizer Adam, dan fungsi loss CrossEntropyLoss.

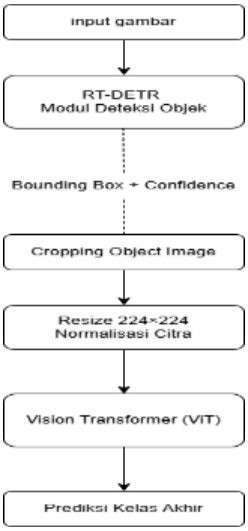
***Tabel 4**. Konfigurasi Pelatihan Vision Transformer*

Parameter	Konfigurasi
Model	ViT-Base Patch16-224
Jumlah kelas	15
Ukuran citra	224 × 224 piksel
Epoch	6
Batch size	16
Learning rate	0,0001
Optimizer	Adam
Loss function	CrossEntropyLoss
Normalisasi	ImageNet

Data *cropping* berdasarkan *ground-truth bounding box* digunakan pada tahap pelatihan dan validasi ViT. Model klasifikasi terbaik kemudian digunakan untuk mengklasifikasikan hasil *cropping* yang berasal dari prediksi RT-DETR pada tahap pengujian sistem *hibrida*.

Integrasi Model Hibrida RT-DETR–Vision Transformer

Dalam penelitian ini, istilah model hibrida merujuk pada integrasi dua model berbasis Transformer pada tingkat alur pemrosesan. Kedua model dilatih secara terpisah dan dioperasikan secara berurutan dalam satu sistem deteksi dan klasifikasi. Integrasi model dilakukan dengan menghubungkan keluaran RT-DETR dengan masukan ViT secara berurutan. Pada tahap pertama, citra masukan diberikan kepada RT-DETR untuk memperoleh lokasi objek berupa *bounding box* beserta kelas dan tingkat kepercayaan deteksi. Area objek berdasarkan *bounding box* tersebut kemudian dipotong menjadi citra *cropping*. Citra hasil *cropping* selanjutnya melalui proses *preprocessing*, yaitu perubahan ukuran menjadi 224 × 224 piksel dan normalisasi sesuai konfigurasi ViT. Citra tersebut kemudian diberikan kepada model ViT untuk menentukan kelas produk. Dengan mekanisme tersebut, RT-DETR berfungsi sebagai komponen deteksi, sedangkan ViT berfungsi sebagai komponen klasifikasi. Secara umum, alur sistem *hibrida* dapat digambarkan sebagai berikut:



***Gambar 1.** Alur Sistem Hibrida RT-DETR–Vision Transformer*

Pada tahap pelatihan, kedua model dilatih secara terpisah sesuai dengan fungsi masing-masing. RT-DETR dilatih menggunakan citra dan anotasi *bounding box*, sedangkan ViT dilatih menggunakan citra hasil *cropping* berdasarkan *ground-truth bounding box*. Pada tahap pengujian, kedua model digunakan secara berurutan sehingga *cropping* yang menjadi masukan ViT berasal dari hasil deteksi RT-DETR.

Pengujian Sistem

Pengujian dilakukan menggunakan data pengujian yang telah dipisahkan dari data pelatihan dan validasi. Setiap citra uji diberikan sebagai masukan ke dalam sistem *hibrida*. RT-DETR terlebih dahulu menghasilkan prediksi lokasi objek dan tingkat kepercayaan deteksi. Area objek berdasarkan hasil prediksi tersebut kemudian diproses melalui tahap *cropping* dan *preprocessing* sebelum diberikan kepada ViT. ViT selanjutnya menghasilkan prediksi kelas produk berdasarkan citra hasil *cropping*. Prediksi akhir sistem diperoleh dari hasil klasifikasi ViT dan dibandingkan dengan kelas sebenarnya pada data pengujian. Prosedur ini digunakan untuk mengevaluasi kemampuan sistem secara keseluruhan dalam melakukan deteksi objek dan klasifikasi produk secara berurutan.

Metode Evaluasi

Evaluasi dilakukan secara terpisah terhadap komponen deteksi, komponen klasifikasi, dan sistem *hibrida*. Evaluasi model RT-DETR dilakukan menggunakan *Precision*, *Recall*, *mean Average Precision* (mAP) pada IoU 0,50 (mAP@50), serta mAP pada rentang IoU 0,50–0,95 (mAP@50–95). Metrik tersebut digunakan untuk menilai kemampuan model dalam menemukan objek dan menghasilkan prediksi *bounding box* yang sesuai.

Evaluasi model ViT dan sistem *hibrida* dilakukan menggunakan *Accuracy*, *Precision*, *Recall*, dan *F1-score*. *Accuracy* digunakan untuk mengukur proporsi prediksi klasifikasi yang benar terhadap seluruh data pengujian. *Precision* menunjukkan ketepatan prediksi suatu kelas, sedangkan *Recall* menunjukkan kemampuan model dalam menemukan data yang termasuk ke dalam kelas tersebut. *F1-score* digunakan sebagai ukuran gabungan antara *Precision* dan *Recall*. Selain metrik numerik, hasil klasifikasi dianalisis menggunakan *classification report* dan *confusion matrix*. *Classification report* digunakan untuk menyajikan kinerja klasifikasi pada masing-masing kelas, sedangkan *confusion matrix* digunakan untuk melihat distribusi prediksi benar dan kesalahan klasifikasi antarkelas.

Dengan demikian, evaluasi dilakukan pada dua tingkat, yaitu evaluasi kemampuan RT-DETR dalam mendeteksi lokasi objek dan evaluasi kemampuan ViT serta sistem *hibrida* dalam menentukan kelas produk. Nilai evaluasi yang diperoleh pada tahap ini selanjutnya digunakan sebagai dasar penyajian hasil penelitian dan pembahasan.

Hasil

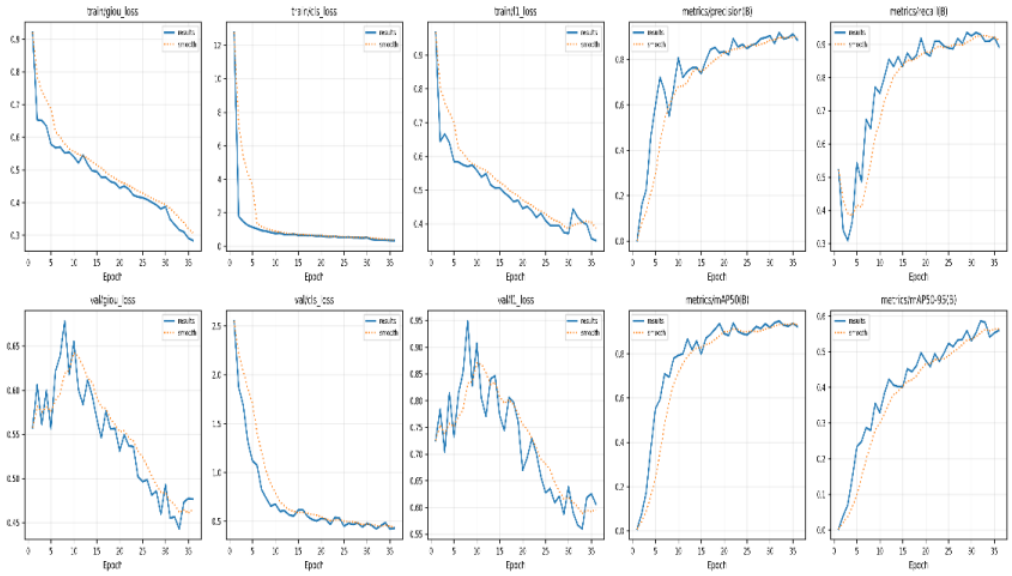
Hasil Pengembangan Model

Hasil implementasi menunjukkan bahwa pendekatan hibrida RT-DETR–Vision Transformer dapat digunakan untuk melakukan identifikasi produk retail melalui dua tahapan pemrosesan yang saling berkaitan. Hasil deteksi RT-DETR memberikan lokasi objek dalam citra, sedangkan Vision Transformer digunakan untuk menentukan kelas produk berdasarkan area objek yang telah diperoleh. Dengan pembagian fungsi tersebut, proses identifikasi tidak hanya mempertimbangkan keberadaan objek, tetapi juga karakteristik visual yang digunakan untuk membedakan kelas produk.

Dataset penelitian mencakup 15 kelas produk retail dengan total 1.202 citra. Distribusi jumlah citra antarkelas tidak sepenuhnya sama, dengan jumlah terbanyak terdapat pada kelas Entrasol dan Tongtji sebanyak 90 citra, sedangkan jumlah paling sedikit terdapat pada Kapal Api sebanyak 68 citra. Kondisi tersebut menunjukkan bahwa dataset memiliki ketidakseimbangan jumlah data antarkelas dalam tingkat yang relatif terbatas. Hasil tersebut menjadi dasar untuk menilai kemampuan masing-masing model dan kinerja sistem hibrida pada data pengujian. Pembahasan selanjutnya difokuskan pada hasil pelatihan RT-DETR, hasil pelatihan Vision Transformer, serta performa akhir sistem hibrida.

Hasil dan Analisis Model RT-DETR

Hasil pelatihan RT-DETR menunjukkan adanya kecenderungan penurunan nilai loss selama proses pembelajaran. Penurunan tersebut terlihat pada komponen training loss, classification loss, GIoU loss, dan L1 loss. Pada data validasi, nilai loss juga menunjukkan kecenderungan menurun meskipun terdapat fluktuasi pada beberapa epoch. Pola tersebut menunjukkan bahwa model secara bertahap memperoleh representasi yang lebih baik terhadap karakteristik objek yang terdapat dalam dataset.



Gambar 2. *Grafik Hasil Pelatihan RT-DETR*

Dari sisi kinerja deteksi, nilai Precision dan Recall RT-DETR berada pada kisaran 0,90. Sementara itu, nilai mAP@50 berada pada kisaran 0,93–0,95**,** sedangkan mAP@50–95 berada pada kisaran 0,58–0,60.

Tabel 5. *Hasil Evaluasi RT-DETR*

Metrik	Hasil
Precision	≈ 0,90
Recall	≈ 0,90
mAP@50	≈ 0,93–0,95
mAP@50–95	≈ 0,58–0,60

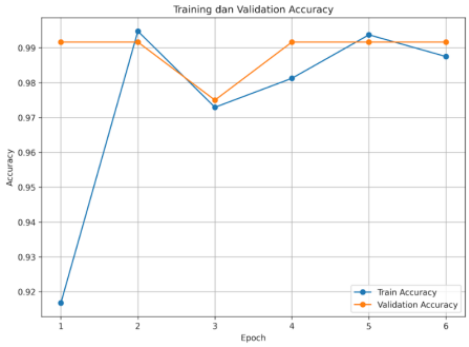
Nilai mAP@50 yang relatif tinggi menunjukkan bahwa RT-DETR mampu menghasilkan deteksi yang sesuai dengan ground truth pada ambang IoU 0,50. Namun, nilai mAP@50–95 yang lebih rendah menunjukkan bahwa ketika kriteria IoU dibuat lebih ketat, ketepatan posisi bounding box mengalami penurunan. Perbedaan kedua nilai

tersebut penting karena menunjukkan bahwa kemampuan model dalam menemukan objek sudah baik, tetapi presisi lokasi bounding box pada tingkat overlap yang lebih ketat masih memiliki ruang untuk ditingkatkan.

Temuan tersebut juga memperlihatkan bahwa keberhasilan tahap deteksi menjadi bagian penting dalam keseluruhan sistem hibrida. Bounding box yang dihasilkan RT-DETR menjadi dasar bagi tahap klasifikasi berikutnya. Dengan demikian, kualitas deteksi tidak hanya menentukan kemampuan sistem dalam menemukan produk, tetapi juga menentukan kualitas area visual yang diteruskan kepada Vision Transformer. Hasil ini sejalan dengan penelitian terdahulu yang menerapkan RT-DETR pada deteksi produk retail dan menunjukkan bahwa RT-DETR memiliki kemampuan deteksi yang tinggi pada objek produk (Maulana et al., 2024). Namun, penelitian tersebut berfokus pada deteksi produk, sedangkan penelitian ini melanjutkannya dengan tahap klasifikasi menggunakan Vision Transformer sehingga proses identifikasi tidak berhenti pada lokasi objek.

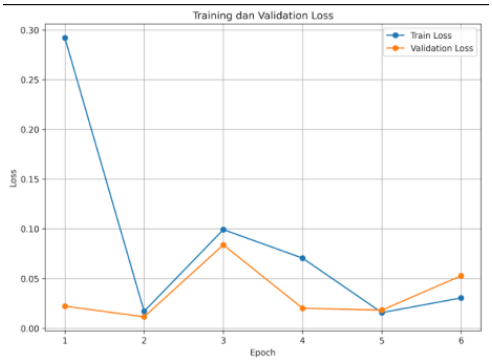
Hasil dan Analisis Vision Transformer

Performa Vision Transformer menunjukkan perkembangan yang cukup konsisten selama proses pembelajaran. Pada awal pelatihan, Training Accuracy berada pada sekitar 91,7%, kemudian meningkat menjadi sekitar 99,5% pada epoch kedua. Pada epoch ketiga terjadi penurunan menjadi sekitar 97,3%, kemudian meningkat kembali pada epoch berikutnya. Pada epoch keenam, Training Accuracy berada pada sekitar 98,7%, sedangkan Validation Accuracy mencapai sekitar 99,2%.



Gambar 3. *Grafik Training Accuracy dan Validation Accuracy*

Perkembangan loss memberikan pola yang mendukung hasil tersebut. Training Loss yang pada epoch pertama berada di sekitar 0,29 mengalami penurunan menjadi sekitar 0,02 pada epoch kedua. Validation Loss secara umum berada pada nilai rendah dan relatif stabil, walaupun terdapat fluktuasi pada beberapa epoch. Pada epoch keenam, Validation Loss berada pada sekitar 0,05.



Gambar 4. *Grafik Training Loss dan Validation Loss Vision Transformer*

Tabel 6. *Ringkasan Performa Vision Transformer*

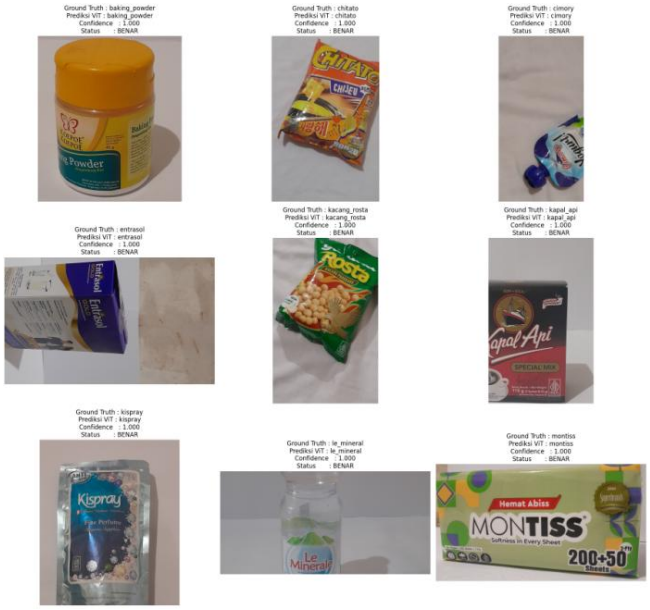
Indikator	Hasil
Training Accuracy awal	≈ 91,7%
Training Accuracy tertinggi yang tercatat	≈ 99,5%
Training Accuracy epoch 6	≈ 98,7%
Validation Accuracy epoch 6	≈ 99,2%
Training Loss epoch 1	≈ 0,29
Training Loss epoch 2	≈ 0,02
Validation Loss epoch 6	≈ 0,05

Validation Accuracy sebesar 99,2% menunjukkan bahwa Vision Transformer mampu membedakan kelas produk dengan tingkat ketepatan yang tinggi pada data validasi. Hasil ini menunjukkan bahwa informasi visual yang terdapat pada area produk cukup representatif untuk membedakan 15 kelas yang digunakan dalam penelitian.

Penggunaan area objek sebagai masukan klasifikasi juga memberikan keuntungan dari sisi representasi visual. Model dapat berfokus pada karakteristik produk seperti bentuk kemasan, warna, tulisan, logo, dan pola visual tanpa harus memproses keseluruhan latar belakang citra. Dalam konteks ini, hasil klasifikasi menunjukkan bahwa tahap klasifikasi mampu memanfaatkan informasi visual yang tersedia pada objek hasil deteksi. Hasil tersebut sejalan dengan karakteristik Vision Transformer yang memanfaatkan mekanisme self-attention untuk menangkap hubungan antarbagiannya secara lebih menyeluruh. Dosovitskiy et al. menunjukkan kemampuan ViT dalam melakukan klasifikasi citra dengan memanfaatkan representasi berbasis patch dan hubungan global antarbagian citra.

Hasil Integrasi RT-DETR–Vision Transformer

Hasil integrasi menunjukkan bahwa kedua model dapat menjalankan fungsi yang berbeda dalam satu sistem. RT-DETR berperan pada tahap penentuan lokasi objek, sedangkan Vision Transformer berperan pada tahap penentuan kelas produk. Hasil visualisasi memperlihatkan adanya keluaran berupa bounding box, kelas produk, dan confidence dari proses prediksi.



Gambar 5. *Contoh Hasil Prediksi Sistem Hibrida*

Integrasi tersebut menghasilkan alur pemrosesan yang saling bergantung. Hasil deteksi RT-DETR menentukan bagian citra yang diteruskan ke Vision Transformer. Dengan demikian, kinerja sistem hibrida tidak hanya ditentukan oleh kemampuan klasifikasi ViT, tetapi juga oleh kualitas area objek yang diperoleh dari tahap deteksi.

Hal penting dari hasil penelitian ini adalah adanya perbedaan antara data yang digunakan untuk melatih ViT dan data yang diterima ViT ketika sistem hibrida diuji. Pada proses pelatihan ViT, citra objek diperoleh berdasarkan ground-truth bounding box, sedangkan pada pengujian hibrida citra diperoleh dari bounding box hasil prediksi RT-DETR. Perbedaan tersebut membuat evaluasi sistem hibrida lebih merepresentasikan kondisi penggunaan kedua model secara berurutan. Dengan demikian, keberhasilan sistem hibrida menunjukkan bahwa model klasifikasi tetap mampu memberikan prediksi ketika menerima area objek yang berasal dari hasil deteksi model lain, bukan hanya area yang diperoleh langsung dari anotasi ground truth.

Hasil Pengujian Sistem Hibrida

Berdasarkan evaluasi rinci pada data pengujian, sistem hibrida memperoleh hasil sebagai berikut.

Tabel 7. *Hasil Evaluasi Sistem Hibrida RT-DETR–Vision Transformer*

Metrik	Nilai
Accuracy	100,00%
Precision	100,00%
Recall	100,00%
F1-score	100,00%

Hasil evaluasi menunjukkan bahwa pada data pengujian yang berhasil diproses oleh sistem, seluruh prediksi kelas produk sesuai dengan label sebenarnya. Precision sebesar 100,00% menunjukkan bahwa prediksi kelas yang dihasilkan sistem sesuai dengan kelas aktual, sedangkan Recall sebesar 100,00% menunjukkan bahwa data kelas yang diuji berhasil dikenali. F1-score sebesar 100,00% merupakan konsekuensi dari nilai Precision dan Recall yang sama-sama mencapai 100,00%.

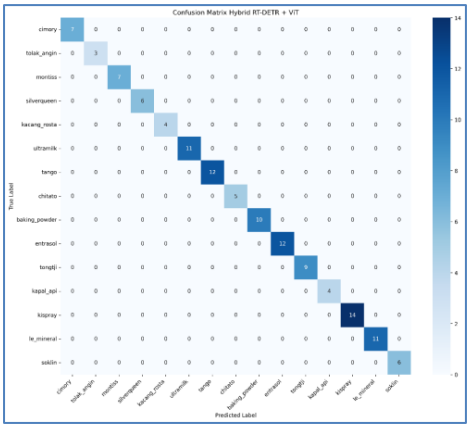
Hasil ini menunjukkan performa klasifikasi yang sangat tinggi pada dataset pengujian penelitian. Dengan demikian, kombinasi tahap deteksi dan klasifikasi dapat bekerja secara efektif pada 15 kelas produk yang menjadi objek penelitian.

Namun, nilai 100,00% perlu diinterpretasikan secara proporsional. Nilai tersebut menggambarkan kinerja sistem pada dataset dan skenario pengujian penelitian, sehingga belum dapat digunakan untuk menyimpulkan bahwa sistem akan memperoleh kinerja yang sama pada seluruh jenis produk retail atau kondisi lingkungan nyata. Perbedaan desain kemasan, kondisi pencahayaan, sudut pengambilan, ukuran objek, latar belakang, maupun produk yang belum terdapat dalam dataset masih dapat memengaruhi performa ketika sistem diterapkan di luar data penelitian.

Analisis Classification Report dan Confusion Matrix

Classification Report digunakan untuk melihat performa sistem pada masing-masing kelas, sedangkan Confusion Matrix digunakan untuk melihat hubungan antara kelas aktual dan kelas hasil prediksi. Berdasarkan hasil evaluasi yang dilaporkan dalam penelitian, nilai keseluruhan Precision, Recall, dan F1-score mencapai 100,00%.

	precision	recall	f1-score	support
cimory	1.00	1.00	1.00	7
tolak_angin	1.00	1.00	1.00	3
montiss	1.00	1.00	1.00	7
silverqueen	1.00	1.00	1.00	6
kacang_rosta	1.00	1.00	1.00	4
ultramilk	1.00	1.00	1.00	11
tango	1.00	1.00	1.00	12
chitato	1.00	1.00	1.00	5
baking_powder	1.00	1.00	1.00	10
entrasol	1.00	1.00	1.00	12
tongtji	1.00	1.00	1.00	9
kapal_api	1.00	1.00	1.00	4
kispray	1.00	1.00	1.00	14
le_mineral	1.00	1.00	1.00	11
soklin	1.00	1.00	1.00	6
accuracy			1.00	121
macro avg	1.00	1.00	1.00	121
weighted avg	1.00	1.00	1.00	121



***Gambar 6**. Classification Report dan Confusion Matrix*

Hasil tersebut menunjukkan bahwa evaluasi klasifikasi pada data pengujian tidak menunjukkan adanya perbedaan antara kelas aktual dan kelas prediksi pada hasil yang berhasil diproses oleh sistem. Dengan kata lain, setiap kelas yang dievaluasi memperoleh hasil prediksi yang sesuai dengan label sebenarnya. Temuan tersebut memperkuat hasil Accuracy sebesar 100,00%. Accuracy saja sebenarnya hanya memberikan gambaran umum, sedangkan Classification Report dan Confusion Matrix memberikan informasi yang lebih terperinci mengenai distribusi prediksi. Oleh karena itu, kesesuaian antara hasil metrik keseluruhan dengan analisis kelas memberikan dukungan tambahan terhadap hasil pengujian sistem hibrida.

Pembahasan

Hasil penelitian menunjukkan bahwa integrasi RT-DETR dan Vision Transformer mampu memberikan kinerja yang sangat baik pada identifikasi produk retail. RT-DETR memperoleh nilai Precision dan Recall sekitar 0,90 dengan mAP@50 pada kisaran 0,93–0,95, sedangkan Vision Transformer mencapai Validation Accuracy sebesar 99,2%. Ketika kedua model diintegrasikan dalam sistem hibrida, hasil evaluasi pada data pengujian menunjukkan nilai Accuracy, Precision, Recall, dan F1-score masing-masing sebesar 100,00%. Hasil tersebut menunjukkan bahwa RT-DETR mampu menyediakan hasil deteksi yang memadai sebagai masukan bagi tahap klasifikasi, sementara Vision Transformer mampu mengenali kelas produk berdasarkan karakteristik visual objek yang diberikan.

Kinerja tersebut menunjukkan bahwa pembagian fungsi antara deteksi dan klasifikasi menjadi salah satu aspek penting dalam pendekatan hibrida. RT-DETR berfokus pada penentuan lokasi objek, kemudian informasi visual dari area objek digunakan oleh Vision Transformer untuk menentukan kelas produk. Hasil ini sejalan dengan penelitian terdahulu yang menunjukkan kemampuan RT-DETR dalam deteksi produk retail (Maulana et al., 2024). Serta penelitian lainnya yang menunjukkan kemampuan Vision Transformer dalam tugas klasifikasi citra (Dosovitskiy et al., 2020). Perbedaan penelitian ini terletak pada penggabungan kedua pendekatan tersebut dalam satu alur untuk mengenali produk retail berdasarkan citra kemasan.

Meskipun memperoleh nilai 100,00% pada evaluasi sistem hibrida, hasil tersebut perlu dipahami dalam batasan dataset penelitian. Dataset hanya mencakup 15 kelas produk dengan jumlah citra yang tidak sepenuhnya seimbang, sehingga hasil tersebut belum dapat digunakan untuk menyatakan bahwa sistem memiliki generalisasi yang sama pada seluruh produk dan kondisi lingkungan retail. Variasi yang lebih luas, seperti penambahan

kelas produk, kondisi pencahayaan, sudut pengambilan, serta objek yang saling bertumpuk, masih diperlukan untuk menguji ketahanan sistem pada kondisi yang lebih kompleks. Dengan demikian, hasil penelitian menunjukkan bahwa pendekatan hibrida RT-DETR–Vision Transformer efektif pada dataset dan skenario pengujian yang digunakan dalam penelitian ini.

Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, model hibrida RT-DETR–Vision Transformer berhasil diimplementasikan untuk melakukan deteksi dan klasifikasi produk retail berdasarkan citra kemasan. RT-DETR berperan dalam mendeteksi lokasi objek produk, sedangkan Vision Transformer digunakan untuk menentukan kelas produk berdasarkan area objek yang terdeteksi. Sistem dikembangkan menggunakan dataset sebanyak 1.202 citra dari 15 kelas produk dan menunjukkan bahwa pendekatan deteksi dan klasifikasi secara berurutan dapat diterapkan pada skenario penelitian yang digunakan. Hasil evaluasi menunjukkan bahwa model RT-DETR memperoleh nilai Precision dan Recall sekitar 0,90, dengan mAP@50 sebesar 0,93–0,95 dan mAP@50–95 sebesar 0,58–0,60. Pada tahap klasifikasi, Vision Transformer mencapai Validation Accuracy sebesar 99,2% pada akhir pelatihan. Berdasarkan evaluasi rinci terhadap sistem hibrida pada data pengujian, diperoleh nilai Accuracy, Precision, Recall, dan F1-score masing-masing sebesar 100,00%. Hasil tersebut menunjukkan bahwa sistem mampu melakukan deteksi dan klasifikasi produk dengan kinerja yang sangat baik pada dataset dan skenario pengujian penelitian.

Meskipun demikian, hasil tersebut masih terbatas pada 15 kelas produk dan karakteristik dataset yang digunakan, sehingga belum dapat secara langsung digeneralisasikan pada seluruh kondisi lingkungan retail. Pengembangan selanjutnya dapat dilakukan dengan memperluas jumlah kelas dan dataset serta meningkatkan keragaman kondisi citra, seperti pencahayaan, sudut pengambilan, latar belakang, dan objek yang saling bertumpuk. Dengan demikian, penelitian ini menunjukkan bahwa kombinasi RT-DETR sebagai model deteksi dan Vision Transformer sebagai model klasifikasi memiliki potensi untuk mendukung pengenalan produk retail secara otomatis berdasarkan citra kemasan.

Acknowledgment

-

Daftar Pustaka

Amjoud, A. B., & Amrouch, M. (2023). Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review. *IEEE Access*, 11, 35479–35516. https://doi.org/10.1109/ACCESS.2023.3266093

Arkin, E., Yadikar, N., Xu, X., Aysa, A., & Ubul, K. (2023). A survey: Object detection methods from CNN to transformer. *Multimedia Tools and Applications*, 82(14), 21353–21383. https://doi.org/10.1007/s11042-022-13801-3

Chen, S., Liu, D., Pu, Y., & Zhong, Y. (2022). Advances in deep learning-based image recognition of product packaging. *Image and Vision Computing*, 128, 104571. https://doi.org/10.1016/j.imavis.2022.104571

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929

Fuchs, K., Grundmann, T., & Fleisch, E. (2019). Towards Identification of Packaged Products via Computer Vision: Convolutional Neural Networks for Object Detection and Image Classification in Retail Environments. *Proceedings of the 9th International Conference on the Internet of Things*, 1–8. https://doi.org/10.1145/3365871.3365899

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. https://doi.org/10.1145/3458723

Gothai, E., Bhatia, S., Alabdali, A. M., Sharma, D. K., Kondamudi, B. R., & Dadheech, P. (2022). Design features of grocery product recognition using deep learning. *Intelligent Automation & Soft Computing*, 34(2), 1231–1246. https://doi.org/10.32604/iasc.2022.026264

Gothai, E., Bhatia, S., M. Alabdali, A., Kumar Sharma, D., Raj Kondamudi, B., & Dadheech, P. (2022). Design Features of Grocery Product Recognition Using Deep Learning. *Intelligent Automation & Soft Computing*, 34(2), 1231–1246. https://doi.org/10.32604/iasc.2022.026264

Hosseini, A., Hooshanfar, K., Omrani, P., Toosi, R., Toosi, R., Ebrahimian, Z., & Akhaee, M. A. (2025). Brand visibility in packaging: A deep learning approach for logo detection, saliency-map prediction, and logo placement analysis. *Discover Applied Sciences*, 7(6), 537. https://doi.org/10.1007/s42452-025-07043-9

Hou, S., Li, J., Min, W., Hou, Q., Zhao, Y., Zheng, Y., & Jiang, S. (2024). Deep Learning for Logo Detection: A Survey. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(3), 1–23. https://doi.org/10.1145/3611309

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Maulana, A. W., Wibowo, S. A., Sunarya, U., Rahmania, R., & Insani, A. (2024). Performance Improvement of YOLOv8-RTDETR Method Based Retail Product Detection. *Journal of Image and Graphics*, 12(3), 292–301. https://doi.org/10.18178/joig.12.3.292-301

Melek, C. G., Battini Sönmez, E., & Varlı, S. (2024). Datasets and methods of product recognition on grocery shelf images using computer vision and machine learning approaches: An exhaustive literature review. *Engineering Applications of Artificial Intelligence*, 133, Article 108452. https://doi.org/10.1016/j.engappai.2024.108452

Pietrini, R., Paolanti, M., Mancini, A., Frontoni, E., & Zingaretti, P. (2024). Shelf management: A deep learning-based system for shelf visual monitoring. *Expert*

Systems with Applications, 255(Part B), Article 124635. https://doi.org/10.1016/j.eswa.2024.124635

Selvam, P., & Koilraj, J. A. S. (2022). A Deep Learning Framework for Grocery Product Detection and Recognition. *Food Analytical Methods*, 15(12), 3498–3522. https://doi.org/10.1007/s12161-022-02384-2

Shah, S., & Tembhurne, J. (2023). Object detection using convolutional neural networks and transformer-based models: A review. *Journal of Electrical Systems and Information Technology*, 10, Article 54. https://doi.org/10.1186/s43067-023-00123-z

Wei, Y., Tran, S., Xu, S., Kang, B., & Springer, M. (2020). Deep Learning for Retail Product Recognition: Challenges and Techniques. *Computational Intelligence and Neuroscience*, 2020, 1–23. https://doi.org/10.1155/2020/8875910

Yanti, R., Susilawati, F. E., & Akramunnisa, A. (2024). Rancang Bangun Aplikasi Penjualan Berbasis Website pada Toko MD Komputer Palopo. *Journal Artificial: Informatika Dan Sistem Informasi*, 2(2), 147–160. https://doi.org/10.54065/artificial.552

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, 126, Article 103514. https://doi.org/10.1016/j.dsp.2022.103514

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 16965–16974). https://doi.org/10.1109/CVPR52733.2024.01605