Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Comparison of Yolo26 and RT-DETR-L for Manga Visual Element Detection on The Manga109 Dataset

Daniel Sande Bona[1*], Tindia Febriyati[2]

[1,2]Desain Komunikasi Visual, Institut Seni Budaya Indonesia Tanah Papua, Jayapura, Indonesia
Email: [1*]daniel.sande.bona@gmail.com, [2]tindia.febriyati2@gmail.com
(*: *corresponding author*)

Abstract - *Automatic comic element detection is a critical prerequisite for manga digital analysis applications such as indexing, translation, and accessibility enhancement. Manga109 is a widely used benchmark for this task, yet the rapid progress of real-time object detectors has not been matched by a systematic comparison among state-of-the-art models in this domain, particularly between YOLO-family detectors employing Small-Target-Aware Label assignment (STAL) and transformer-based detectors such as RT-DETR. This study benchmarks YOLO26 (nano, small, and medium variants) against RT-DETR-L for detecting four comic element classes (panel, character, text, and face) on Manga109. All models were trained for 100 epochs at 1024×1024 pixels and evaluated using mAP and per-size AP following COCO conventions. YOLO26m achieves the best performance (mAP@0.5:0.95 of 0.7471), while RT-DETR-L obtains the lowest (0.7001) despite the largest parameter count and slowest inference. For small text detection, YOLO26m outperforms RT-DETR-L by 42.7% relatively, and the AP gap across object sizes narrows monotonically as model size grows, supporting the STAL design claim. RT-DETR-L also exhibits significant training instability. The main contribution is a systematic, multi-dimensional benchmark of YOLO26 against RT-DETR-L on Manga109 that evidence STAL effectiveness for small-text detection and offers practical guidance for selecting real-time detectors in manga image analysis.*

Keywords: *Manga Image Analysis, Manga109, Object Detection, RT-DETR, YOLO26*

Abstrak - Deteksi elemen komik secara otomatis merupakan tahap awal yang menentukan keberhasilan berbagai aplikasi analisis manga digital, seperti pengindeksan, terjemahan, dan peningkatan aksesibilitas. Manga109 merupakan dataset standar untuk tugas ini, tetapi pesatnya perkembangan detektor objek *real-time* belum diimbangi perbandingan sistematis di antara model-model terkini pada domain tersebut, khususnya antara keluarga YOLO yang mengadopsi *Small-Target-Aware Label assignment* (STAL) dan detektor berbasis *transformer* seperti RT-DETR. Penelitian ini melakukan *benchmark* YOLO26 (varian *nano*, *small*, dan *medium*) terhadap RT-DETR-L untuk mendeteksi empat kelas elemen komik, yaitu *panel*, *character*, *text*, dan *face*, pada Manga109. Seluruh model dilatih selama 100 *epoch* dengan ukuran citra 1024×1024 piksel dan dievaluasi menggunakan metrik mAP serta AP berbasis ukuran objek sesuai konvensi COCO. YOLO26m memperoleh kinerja terbaik dengan mAP@0.5:0.95 sebesar 0,7471, sedangkan RT-DETR-L terendah (0,7001) meskipun memiliki parameter terbanyak dan inferensi paling lambat. Pada deteksi teks berukuran kecil, YOLO26m mengungguli RT-DETR-L sebesar 42,7% relatif, dan selisih AP antarukuran objek menyempit secara monoton seiring bertambahnya ukuran model, mendukung klaim desain STAL. RT-DETR-L juga menunjukkan ketidakstabilan pelatihan yang signifikan. Kontribusi penelitian ini adalah *benchmark* sistematis dan multidimensi YOLO26 terhadap RT-DETR-L pada Manga109 yang membuktikan efektivitas STAL untuk deteksi teks kecil sekaligus menjadi acuan praktis pemilihan detektor *real-time* dalam analisis citra manga.

Kata Kunci: Analisis Citra Manga, Deteksi Objek, Manga109, RT-DETR, YOLO26

1. PENDAHULUAN

Komik, khususnya manga, merupakan salah satu medium visual-tekstual yang populer sekaligus bagian penting dari warisan budaya. Seiring digitalisasi koleksi manga dalam jumlah besar, kebutuhan untuk memahami elemen-elemen penyusun halaman komik secara otomatis menjadi semakin penting bagi berbagai aplikasi seperti pengindeksan, temu kembali (*retrieval*), terjemahan, dan peningkatan aksesibilitas konten [1]. Halaman manga umumnya tersusun atas beberapa elemen visual yang khas, yaitu panel (*frame*), wajah (*face*), tubuh (*body*) karakter, dan teks (*text*). Kemampuan mendeteksi elemen-elemen tersebut secara akurat menjadi prasyarat bagi sebagian besar tugas analisis komik tingkat lanjut [2], [3]. *Benchmark* terbaru seperti CoMix [4], [5] juga menunjukkan bahwa deteksi objek menjadi fondasi bagi tugas pemahaman komik yang lebih luas, termasuk identifikasi pembicara, re-identifikasi karakter, urutan baca, dan penalaran multimodal. Jumlah koleksi manga digital yang terus bertambah membuat anotasi manual tidak lagi memungkinkan, sehingga deteksi elemen komik yang akurat sekaligus efisien menjadi kebutuhan yang mendesak. Kebutuhan ini semakin nyata karena banyak aplikasi analisis komik dituntut berjalan secara *real-time* pada perangkat dengan sumber daya terbatas, sehingga pemilihan detektor yang tepat, yaitu yang menyeimbangkan akurasi, kecepatan, dan efisiensi, menjadi persoalan praktis yang belum banyak dikaji pada domain ini.

Tugas deteksi elemen komik memiliki karakteristik yang berbeda dengan deteksi objek pada citra natural. Halaman manga umumnya bersifat monokrom, memiliki kepadatan objek yang tinggi, variasi gaya gambar antarjudul, serta keberadaan elemen teks berukuran sangat kecil yang tersebar di dalam panel [6]. Penelitian awal pada domain ini banyak memanfaatkan arsitektur deteksi konvensional. Ogawa dkk. [3] menerapkan pendekatan

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

berbasis *Single Shot MultiBox Detector* (SSD) pada anotasi Manga109 untuk mendeteksi kelas *frame*, *face*, *body*, dan *text*, sehingga menjadi salah satu rujukan utama deteksi objek pada komik. Studi komprehensif terkini mengenai pemahaman komik [4] menunjukkan bahwa sebagian besar metode masih mengadaptasi arsitektur deteksi dari domain citra natural, sementara evaluasi yang berfokus pada elemen berukuran kecil seperti teks masih terbatas.

Di sisi lain, bidang deteksi objek *real-time* berkembang pesat. Model berbasis *You Only Look Once* (YOLO) telah menjadi salah satu kerangka kerja deteksi yang paling banyak digunakan karena keseimbangannya antara akurasi dan kecepatan inferensi [7]. Sebagian besar detektor YOLO masih bergantung pada *Non-Maximum Suppression* (NMS) sebagai tahap pascaproses. NMS menambah variasi latensi dan menggunakan *hyperparameter* ambang (*threshold*) yang dapat memengaruhi stabilitas akurasi. Sebagai alternatif, muncul pendekatan deteksi *end-to-end* berbasis *transformer*. *Detection Transformer* (DETR) [8] menghilangkan kebutuhan NMS dengan merumuskan deteksi sebagai prediksi himpunan, namun biaya komputasinya yang tinggi membuatnya kurang praktis untuk aplikasi *real-time*. *Real-Time Detection Transformer* (RT-DETR) [9] mengatasi keterbatasan tersebut dan menjadi detektor *transformer end-to-end* pertama yang mampu beroperasi secara *real-time*, dengan klaim mengungguli detektor YOLO baik pada kecepatan maupun akurasi.

Perkembangan terbaru keluarga YOLO hadir melalui YOLO26 [10]. YOLO26 menghapus *Distribution Focal Loss* (DFL) untuk menyederhanakan kepala deteksi, mengadopsi inferensi *NMS-free* secara native, serta memperkenalkan *Progressive Loss* dan *Small-Target-Aware Label assignment* (STAL). STAL dirancang secara khusus untuk menjaga cakupan label positif pada objek berukuran kecil, sehingga secara teoretis relevan untuk deteksi elemen teks pada manga. Namun, klaim desain ini belum diuji secara empiris pada domain komik, dan sejauh ini belum tersedia perbandingan sistematis antara YOLO26 dan RT-DETR pada dataset Manga109.

Perbandingan sistematis antara detektor YOLO dan RT-DETR telah mulai dilakukan pada berbagai domain citra. González Hernández dkk. [11] membandingkan beberapa varian YOLO dengan RT-DETR untuk deteksi multi-objek pada citra umum, sementara Le dkk. [7] melakukan *benchmarking* YOLO dan RT-DETR dari sisi kecepatan, akurasi, dan efisiensi. Namun, perbandingan tersebut belum menyentuh domain komik atau manga yang memiliki karakteristik objek kecil dan padat, sehingga evaluasi yang komprehensif dan tidak bertumpu pada satu metrik agregat pada domain ini masih diperlukan. Sintesis literatur terdahulu beserta celah penelitian yang menjadi dasar studi ini dirangkum pada Tabel 1.

Tabel 1. Sintesis literatur terdahulu dan celah penelitian

Ref.	Peneliti	Domain / Objek Studi	Metode	Temuan Utama dan Celah (Gap)
[3]	Ogawa dkk.	Manga109 (komik)	*Single Shot MultiBox Detector* (SSD)	Menjadi rujukan deteksi elemen komik pada Manga109; berfokus pada satu arsitektur dan belum menekankan evaluasi objek berukuran kecil.
[11]	González Hernández dkk.	Citra umum (multi-objek)	Beberapa varian YOLO vs RT-DETR	Membandingkan YOLO dan RT-DETR pada citra umum; belum mencakup domain komik atau manga.
[7]	Le dkk.	Citra umum	*Benchmarking* YOLO vs RT-DETR	Mengevaluasi kecepatan, akurasi, dan efisiensi; tidak menyoroti objek kecil dan padat khas komik.
[4]	CoMix	Komik (multi-tugas)	*Benchmark* pemahaman komik	Menempatkan deteksi sebagai fondasi; masih mengadaptasi detektor citra natural dengan evaluasi per-ukuran objek yang terbatas.
—	Penelitian ini (2026)	Manga109 (komik)	YOLO26 (n/s/m) vs RT-DETR-L	*Benchmark* multidimensi antar-paradigma dengan analisis AP per-ukuran objek dan validasi klaim STAL untuk teks kecil.

Berdasarkan uraian di atas, terdapat beberapa kesenjangan yang menjadi dasar penelitian ini. Pertama, belum tersedia *benchmark* yang membandingkan secara sistematis detektor *real-time* terkini, yaitu YOLO26 dan RT-DETR, pada dataset Manga109. Kedua, deteksi elemen teks berukuran kecil masih menjadi tantangan utama sekaligus pembeda desain antararsitektur, sehingga memerlukan analisis kinerja yang dipisahkan berdasarkan ukuran objek. Ketiga, klaim desain STAL pada YOLO26 untuk objek kecil belum divalidasi secara empiris pada domain komik. Keempat, stabilitas pelatihan detektor berbasis *transformer* pada citra non-natural seperti manga masih jarang dilaporkan. Penelitian ini bertujuan untuk (1) Membandingkan kinerja deteksi YOLO26 (varian *nano*, *small*, dan *medium*) dengan RT-DETR-L pada dataset Manga109; (2) Menganalisis kinerja deteksi berdasarkan ukuran objek (*small*, *medium*, *large*), khususnya pada kelas teks; (3) Menguji secara empiris klaim desain STAL pada deteksi objek berukuran kecil; (4) Melaporkan stabilitas pelatihan masing-masing model sebagai pertimbangan praktis dalam pemilihan detektor.

Kontribusi penelitian ini ada tiga. Pertama, penelitian ini menyediakan *benchmark* sistematis pertama yang membandingkan YOLO26 (varian *nano*, *small*, dan *medium*) dengan RT-DETR-L pada dataset Manga109 dalam

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

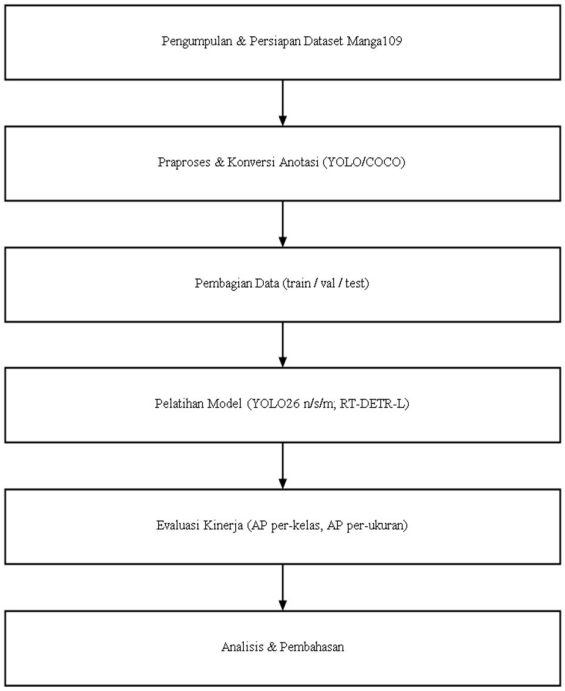
kondisi eksperimen yang identik. Kedua, penelitian ini memperkenalkan kerangka evaluasi multidimensi yang mencakup akurasi keseluruhan, kinerja berdasarkan ukuran objek, perilaku konvergensi, efisiensi pelatihan, kecepatan inferensi, serta analisis kesalahan kualitatif. Ketiga, penelitian ini memberikan bukti empiris yang mendukung efektivitas mekanisme *Small-Target-Aware Label assignment* (STAL) untuk deteksi teks berukuran kecil pada citra manga.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian eksperimental kuantitatif yang membandingkan kinerja dua kelompok detektor objek *real-time*, yaitu YOLO26 dan RT-DETR, pada tugas deteksi elemen komik menggunakan dataset Manga109. Tahapan penelitian secara keseluruhan ditunjukkan pada Gambar 1.

2.1. Tahapan Penelitian

Tahapan penelitian terdiri atas enam tahap, yaitu pengumpulan dan persiapan dataset Manga109, praproses dan konversi anotasi, pembagian data menjadi data latih, validasi, dan uji, pelatihan model deteksi, evaluasi kinerja, serta analisis dan pembahasan hasil. Urutan logis tahapan tersebut ditunjukkan pada Gambar 1.

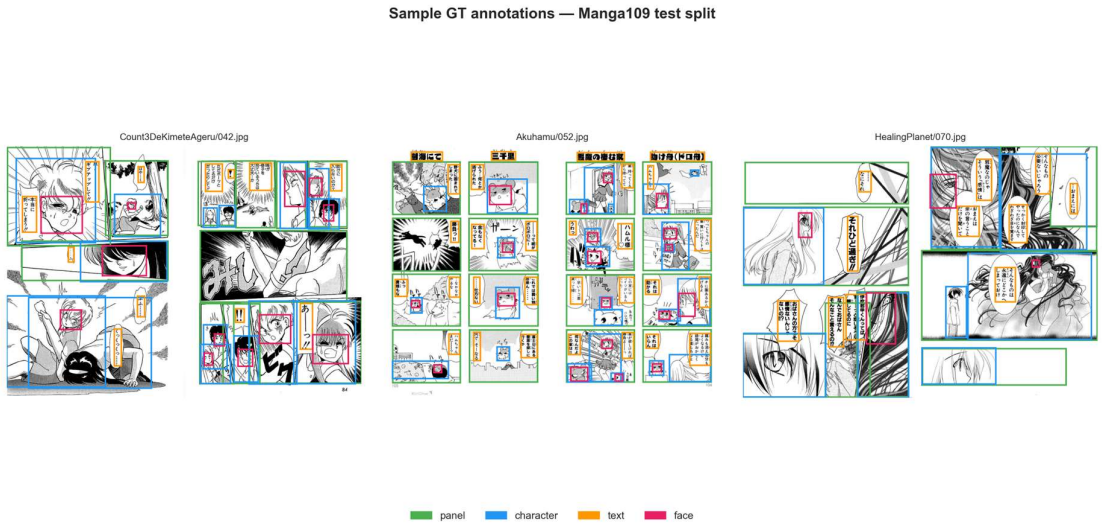


Gambar 1. Tahapan penelitian

2.2. Dataset Manga109

Dataset yang digunakan pada penelitian ini adalah Manga109 v2021.12.30 [1], [2], yaitu dataset citra manga beranotasi yang banyak digunakan untuk penelitian analisis komik. Dataset ini diperoleh dari repositori HuggingFace milik Aizawa-Yamakata-Matsui Laboratory, University of Tokyo (huggingface.co/datasets/hal-utokyo/Manga109) dan terdiri atas 109 judul manga karya sekitar 93 pengarang dengan total sekitar 10.600 halaman. Anotasi yang digunakan mencakup empat kelas elemen visual, yaitu *panel*, *character*, *text*, dan *face*, mengikuti definisi kelas pada penelitian deteksi objek komik sebelumnya [3]. Anotasi asli Manga109 tersimpan dalam berkas XML per judul dengan tag *frame*, *body*, *text*, dan *face*. Tag *frame* dan *body* dipetakan berturut-turut menjadi kelas *panel* dan *character*, sedangkan *text* dan *face* dipertahankan. Atribut identitas karakter dan isi transkripsi teks pada anotasi asli diabaikan karena penelitian ini hanya memanfaatkan kotak pembatas dan label kelas. Seluruh anotasi selanjutnya dikonversi ke format pelatihan detektor berupa koordinat kotak pembatas ternormalisasi beserta indeks kelas. Contoh anotasi *ground truth* pada beberapa halaman uji ditunjukkan pada Gambar 2.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

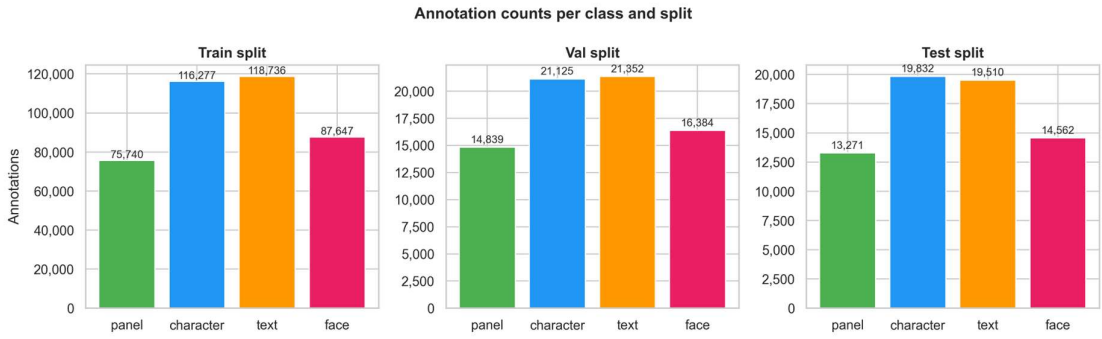


Gambar 2. Contoh anotasi ground truth Manga109 pada split uji (panel, character, text, face)

Data dibagi berdasarkan judul (*split by-title*) dengan *seed* 42 dan rasio sekitar 73/14/13 untuk data latih, validasi, dan uji. Set uji berisi 1.356 gambar dengan 67.175 anotasi, yaitu jumlah seluruh anotasi keempat kelas pada kolom Uji Tabel 2 (baris Total). Jumlah anotasi per kelas dan per split disajikan pada Tabel 2 dan divisualisasikan pada Gambar 3.

Tabel 2. Jumlah anotasi per kelas dan split

Kelas	Latih	Validasi	Uji
panel	75.740	14.839	13.271
character	116.277	21.125	19.832
text	118.736	21.352	19.510
face	87.647	16.384	14.562
Total	398.400	73.700	67.175



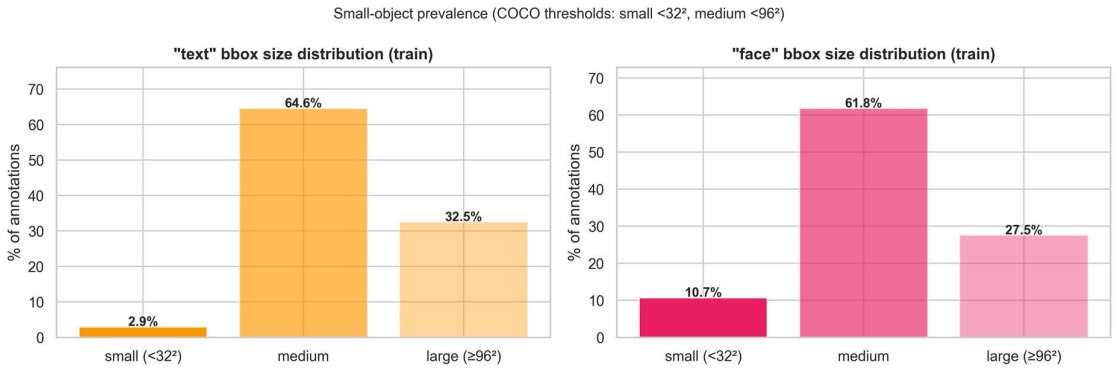
Gambar 3. Jumlah anotasi per kelas dan split

Untuk keperluan analisis kinerja berdasarkan ukuran objek, definisi kategori ukuran objek mengikuti konvensi COCO [12], yaitu objek berukuran kecil (*small*) dengan luas area kurang dari 32×32 piksel (1.024 px2), sedang (*medium*) antara 32×32 dan 96×96 piksel, dan besar (*large*) lebih dari 96×96 piksel. Distribusi ukuran objek pada split uji ditunjukkan pada Tabel 3 dan Gambar 4. Distribusi ini penting karena kelas *text* dan *face* memiliki proporsi objek kecil yang jauh lebih tinggi dibandingkan kelas *panel*, sehingga menjadi penentu utama pada analisis deteksi objek kecil di Bagian 3.

Tabel 3. Distribusi ukuran objek per kelas pada split uji

Kelas	Small	Medium	Large	Total
panel	0	9	13.262	13.271
character	75	3.239	16.518	19.832
text	351	11.914	7.245	19.510
face	1.177	8.775	4.610	14.562

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



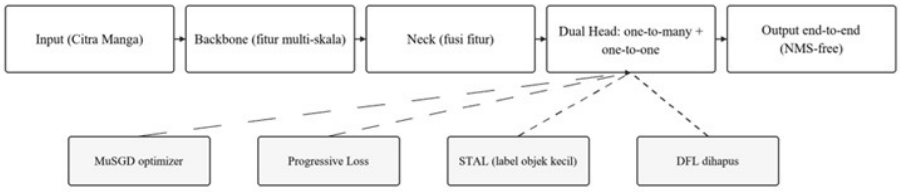
Gambar 4. Distribusi ukuran bounding box kelas text dan face pada data latih

2.3. Arsitektur Model Deteksi

Kedua model yang dibandingkan mewakili dua paradigma dominan dalam deteksi objek *real-time*. YOLO26 merupakan detektor satu tahap (*one-stage*) berbasis *Convolutional Neural Network* yang bersifat *anchor-free*, sedangkan RT-DETR-L merupakan detektor *end-to-end* berbasis *transformer* dengan arsitektur *encoder-decoder*. Perbedaan mendasar keduanya terletak pada cara menghasilkan prediksi: YOLO26 menghasilkan prediksi secara padat (*dense prediction*) melalui konvolusi pada seluruh posisi peta fitur di berbagai skala, sementara RT-DETR-L menggunakan sejumlah *object query* dan mekanisme *self-attention* untuk memodelkan relasi antar-objek secara global. Kedua metode dipilih karena sama-sama mendukung inferensi tanpa NMS dan menargetkan kinerja *real-time*, sehingga dapat dibandingkan secara adil dari sisi akurasi maupun kecepatan pada perangkat GPU kelas konsumen. Uraian rinci masing-masing arsitektur diberikan pada Subbagian a dan b.

a. YOLO26

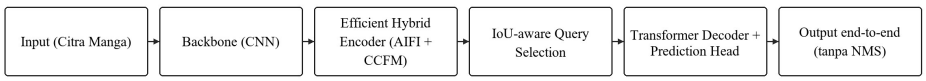
YOLO26 [10] merupakan anggota terbaru keluarga YOLO yang dirancang dengan penekanan pada kesederhanaan dan efisiensi penerapan. Berbeda dengan detektor YOLO sebelumnya, YOLO26 menghilangkan *Distribution Focal Loss* (DFL) untuk menyederhanakan kepala deteksi (*detection head*) dan menerapkan inferensi *NMS-free* secara native melalui desain kepala ganda (*dual head*). Pada tahap pelatihan, YOLO26 memanfaatkan tiga komponen utama, yaitu *Progressive Loss* yang menggeser penekanan supervisi ke arah kepala inferensi, *Small-Target-Aware Label assignment* (STAL) yang menjaga cakupan label positif untuk objek berukuran kecil, serta *optimizer* MuSGD. Pada penelitian ini digunakan tiga varian YOLO26, yaitu *nano* (YOLO26n), *small* (YOLO26s), dan *medium* (YOLO26m). Diagram blok arsitektur YOLO26 ditunjukkan pada Gambar 5.



Gambar 5. Diagram blok arsitektur YOLO26

b. RT-DETR

RT-DETR [9] merupakan detektor objek *real-time* berbasis *transformer* yang dikembangkan dari kerangka *Detection Transformer* (DETR) [8]. RT-DETR mempertahankan sifat *end-to-end* tanpa NMS, namun menekan biaya komputasi melalui *efficient hybrid encoder* yang memisahkan interaksi fitur intra-skala dan fusi fitur lintas-skala, serta pemilihan *query* yang memperhatikan kualitas (*IoU-aware query selection*). Pada penelitian ini digunakan varian RT-DETR-L. Diagram blok arsitektur RT-DETR ditunjukkan pada Gambar 6.



Gambar 6. Diagram blok arsitektur RT-DETR

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

2.4. Setup Pelatihan

Seluruh proses pelatihan dilakukan pada perangkat ASUS ROG Strix G634JZ dengan GPU NVIDIA RTX 4080 Mobile (12 GB), sistem operasi Ubuntu 22.04, CUDA 12.6, dan PyTorch 2.7.0. Kerangka kerja yang digunakan adalah Ultralytics v8.4.67 dengan ukuran citra masukan 1024×1024 piksel dan pelatihan selama 100 *epoch*. Ukuran *batch* disesuaikan dengan kapasitas model, yaitu 16 untuk YOLO26n, 8 untuk YOLO26s, 4 untuk YOLO26m, dan 4 untuk RT-DETR-L.

Seluruh model diinisialisasi dari bobot *pretrained* dataset COCO, lalu di-*fine-tune* secara penuh pada data latih Manga109 tanpa membekukan satu lapisan pun sehingga seluruh parameter jaringan ikut diperbarui selama pelatihan. Proses pelatihan mengaktifkan presisi campuran (*Automatic Mixed Precision*, AMP) secara default untuk menekan penggunaan memori GPU, dengan jumlah kelas keluaran diatur menjadi empat (nc = 4) sesuai elemen komik yang dideteksi. Di luar ukuran citra, jumlah *epoch*, dan ukuran *batch* yang telah disebutkan, seluruh *hyperparameter* lainnya, seperti *learning rate*, *optimizer*, dan skema augmentasi, dibiarkan pada nilai default masing-masing model di Ultralytics v8.4.67 sehingga perbandingan antar-model tetap adil dan mudah direplikasi.

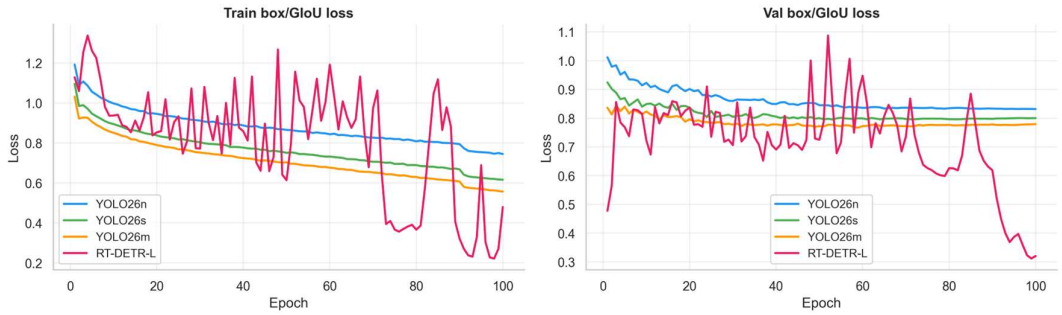
Pada penelitian ini hanya varian RT-DETR-L original [9] yang dilatih sebagai pembanding berbasis *transformer*, bukan RT-DETRv2 [13]. Konfigurasi ini dipilih agar perbandingan berfokus pada keluarga YOLO26 dan satu *baseline* RT-DETR yang representatif pada perangkat GPU kelas konsumen 12 GB. Waktu pelatihan dan kecepatan konvergensi setiap model dicatat sebagai bagian dari analisis efisiensi pada Bagian 3.

3. HASIL DAN PEMBAHASAN

3.1. Tahap Pelatihan dan Stabilitas Konvergensi

Seluruh model dilatih selama 100 *epoch* dengan ukuran citra 1024×1024 piksel. Kurva *mean Average Precision* (mAP) validasi selama pelatihan ditunjukkan pada Gambar 7, sedangkan kurva *loss* pelatihan dan validasi ditunjukkan pada Gambar 8. Ketiga varian YOLO26 menunjukkan pola konvergensi yang halus dan stabil: nilai mAP meningkat tajam pada *epoch* awal kemudian mendatar, dan nilai *loss* menurun secara monoton. Sebaliknya, RT-DETR-L menunjukkan ketidakstabilan pelatihan yang jelas, ditandai oleh osilasi besar pada kurva mAP validasi dan lonjakan *loss* yang berulang sepanjang pelatihan. RT-DETR-L baru mencapai tingkat kinerja yang sebanding dengan YOLO26 pada *epoch* akhir.

Gambar 7. Kurva mAP@0.5 dan mAP@0.5:0.95 validasi selama pelatihan (100 epoch)



Gambar 8. Kurva box/GIoU loss pelatihan dan validasi

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Ketidakstabilan ini dikuantifikasi melalui kecepatan konvergensi, yaitu jumlah *epoch* yang dibutuhkan untuk mencapai persentase tertentu dari mAP@0.5 final, serta waktu pelatihan total seperti disajikan pada Tabel 4. YOLO26n selesai dilatih sekitar 5,0 jam, YOLO26s sekitar 6,3 jam, dan YOLO26m sekitar 10,6 jam. Sebaliknya, RT-DETR-L membutuhkan sekitar 37,6 jam, atau sekitar empat kali lebih lama daripada YOLO26m. Dari sisi konvergensi, ketiga varian YOLO26 mencapai 95% dari mAP final dalam paling lama 4 *epoch*, sedangkan RT-DETR-L membutuhkan 92 *epoch* untuk mencapai ambang yang sama.

Tabel 4. Efisiensi pelatihan dan kecepatan konvergensi

Model	Batch	Waktu latih	Epoch ke 95% mAP	Catatan
YOLO26n	16	~5,0 jam	4	Paling cepat
YOLO26s	8	~6,3 jam	1	Trade-off kecepatan-akurasi
YOLO26m	4	~10,6 jam	1	Akurasi terbaik
RT-DETR-L	4	~37,6 jam	92	Paling lambat dan ukuran model paling besar

Untuk mengkaji stabilitas pelatihan secara lebih mendalam, dianalisis pula kestabilan kurva setelah masing-masing model mencapai titik konvergennya. Setelah *epoch* ke-92, kurva mAP@0.5 RT-DETR-L relatif stabil (standar deviasi 0,011 dan tanpa penurunan tajam lebih dari 0,02 antar-*epoch*), tetapi metrik mAP@0.5:0.95 tetap jauh lebih fluktuatif (standar deviasi 0,048) dibandingkan YOLO26m (0,008), dan standar deviasi loss validasinya sekitar 2,6 kali lebih besar. Hal ini mengindikasikan bahwa kualitas lokalisasi presisi tinggi RT-DETR-L belum sepenuhnya stabil hingga *epoch* ke-100. Fluktuasi ini diduga bersumber dari mekanisme *Hungarian bipartite matching* pada arsitektur *transformer*: pada citra manga yang padat dan saling tumpang-tindih (panel di dalam panel, wajah di dalam tubuh karakter), penugasan slot prediksi ke objek acuan dapat berubah antar-*epoch* sehingga menghasilkan sinyal gradien yang kurang stabil. Selain itu, penggunaan *Automatic Mixed Precision* (AMP) diduga memperbesar variansi pada fase pertengahan pelatihan ketika magnitudo loss masih besar. Temuan ini menyiratkan bahwa RT-DETR-L pada domain manga kemungkinan memerlukan anggaran pelatihan yang lebih panjang, sekitar 150-200 *epoch*, untuk mencapai kestabilan penuh.

3.2. Kinerja Deteksi Keseluruhan

Kinerja deteksi keseluruhan beserta kompleksitas dan kecepatan inferensi tiap model disajikan pada Tabel 5. YOLO26m memperoleh kinerja tertinggi dengan mAP@0.5 sebesar 0,9366 dan mAP@0.5:0.95 sebesar 0,7471. Perlu dicatat bahwa RT-DETR-L justru memperoleh mAP@0.5:0.95 terendah (0,7001) meskipun memiliki jumlah parameter terbesar (32,81 juta) dan kecepatan inferensi paling rendah (42,76 FPS). Bahkan YOLO26n yang hanya memiliki 2,51 juta parameter memperoleh mAP@0.5 (0,9249) sedikit lebih tinggi daripada RT-DETR-L (0,9218), dengan kecepatan inferensi sekitar 6,5 kali lebih cepat. Temuan ini mengindikasikan bahwa pada deteksi elemen manga, keunggulan jumlah parameter dan kapasitas RT-DETR-L tidak tercermin pada akurasi.

Tabel 5. Kinerja deteksi keseluruhan, kompleksitas, dan kecepatan inferensi (batch=1, imgsz=1024)

Model	Params (M)	FLOPs (G)	FPS	VRAM (MB)	mAP@0.5	mAP@0.5:0.95
YOLO26n	2,51	14,79	276,7	111,1	0,9249	0,7223
YOLO26s	9,95	57,63	156,1	235,7	0,9321	0,7399
YOLO26m	21,78	191,34	69,69	458,3	0,9366	0,7471
RT-DETR-L	32,81	271,14	42,76	480,0	0,9218	0,7001

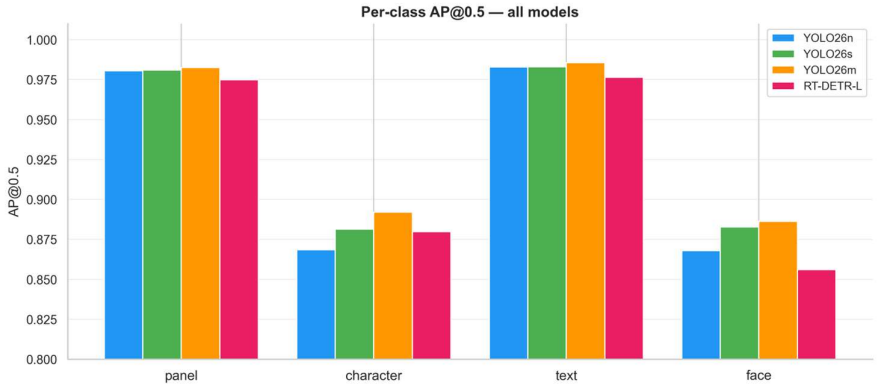
3.3. Kinerja Per-Kelas

Nilai AP@0.5 dan AP@0.5:0.95 untuk tiap kelas disajikan pada Tabel 6 dan divisualisasikan pada Gambar 9. Pada metrik AP@0.5, kelas *panel* dan *text* relatif mudah dideteksi oleh seluruh model (AP@0.5 di atas 0,97), sedangkan kelas *character* dan *face* lebih sulit (AP@0.5 sekitar 0,86-0,89). Perbedaan yang lebih tajam muncul pada metrik AP@0.5:0.95 yang lebih ketat terhadap kualitas lokalisasi. Pada kelas *text*, YOLO26m memperoleh AP@0.5:0.95 sebesar 0,8878, sedangkan RT-DETR-L hanya 0,8115, yaitu selisih sekitar 9,4% relatif. Hal ini mengindikasikan bahwa meskipun RT-DETR-L mampu mendeteksi keberadaan teks (AP@0.5 tinggi), kualitas lokalisasi kotak pembatasnya pada teks lebih rendah dibandingkan YOLO26.

Tabel 6. AP@0.5 dan AP@0.5:0.95 per kelas

Model	panel	character	text	face
YOLO26n	0,9804 / 0,9347	0,8685 / 0,6292	0,9828 / 0,8722	0,8680 / 0,4531
YOLO26s	0,9810 / 0,9391	0,8813 / 0,6575	0,9831 / 0,8816	0,8828 / 0,4814
YOLO26m	0,9825 / 0,9445	0,8920 / 0,6781	0,9855 / 0,8878	0,8862 / 0,4782
RT-DETR-L	0,9749 / 0,9204	0,8798 / 0,6309	0,9764 / 0,8115	0,8561 / 0,4377

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 9. AP@0.5 per kelas untuk seluruh model

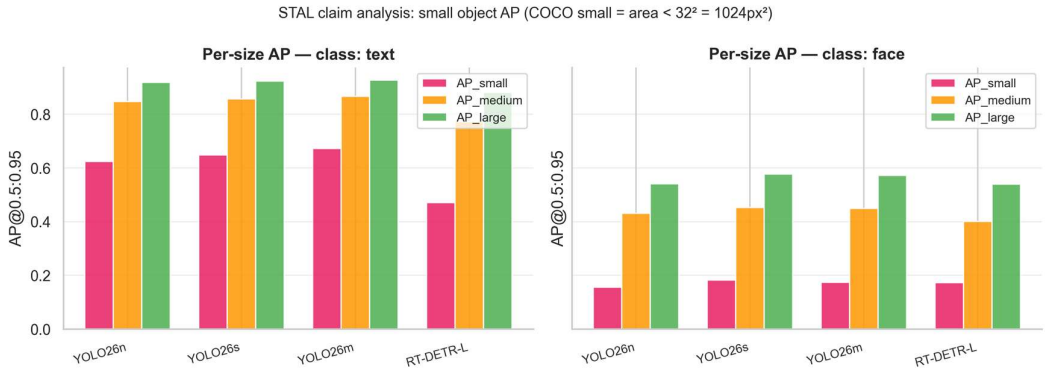
3.4. Analisis AP Berdasarkan Ukuran Objek dan Validasi Klaim STAL

Bagian ini merupakan analisis utama penelitian. Nilai AP berdasarkan ukuran objek untuk kelas *text* dan *face* disajikan pada Tabel 7 dan Gambar 10 (kelas *panel* tidak memiliki objek berukuran *small* sehingga AP-nya tidak terdefinisi). Pada kelas *text*, AP objek kecil (*AP-small*) meningkat secara monoton seiring bertambahnya ukuran model YOLO26, yaitu 0,6242 (YOLO26n), 0,6479 (YOLO26s), dan 0,6722 (YOLO26m). Ketiganya jauh lebih tinggi daripada RT-DETR-L yang hanya memperoleh 0,4709. Dengan demikian, YOLO26m mengungguli RT-DETR-L pada deteksi teks berukuran kecil sebesar 42,7% relatif.

Lebih lanjut, selisih antara AP teks objek besar dan objek kecil (*AP-large* dikurangi *AP-small*) pada YOLO26 menyempit secara monoton seiring bertambahnya ukuran model, yaitu 0,294 (YOLO26n), 0,276 (YOLO26s), dan 0,255 (YOLO26m), sedangkan pada RT-DETR-L selisih tersebut jauh lebih besar, yaitu 0,411. Pola ini konsisten dengan klaim desain *Small-Target-Aware Label assignment* (STAL) pada YOLO26 [10], yaitu menjaga cakupan label positif untuk objek berukuran kecil. Hasil ini mengindikasikan bahwa STAL memberikan keuntungan empiris yang nyata pada deteksi teks kecil di domain manga, dan keuntungan tersebut bertambah seiring meningkatnya kapasitas model.

Tabel 7. AP berdasarkan ukuran objek untuk kelas text dan face

Model	text small	text medium	text large	face small	face medium
YOLO26n	0,6242	0,8470	0,9186	0,1560	0,4306
YOLO26s	0,6479	0,8567	0,9234	0,1823	0,4530
YOLO26m	0,6722	0,8671	0,9275	0,1742	0,4488
RT-DETR-L	0,4709	0,7720	0,8816	0,1725	0,4003



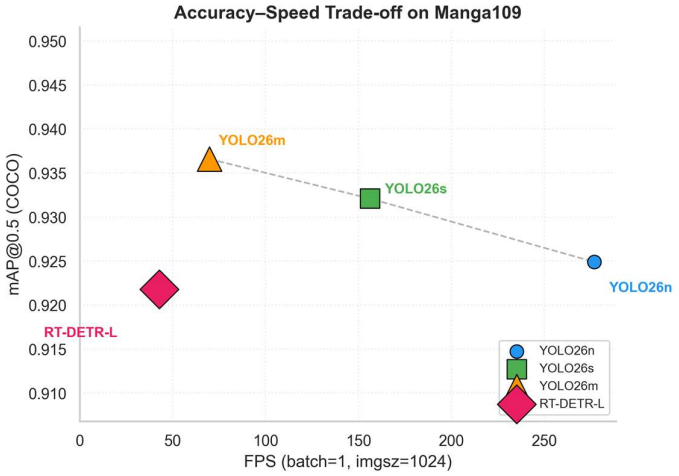
Gambar 10. AP berdasarkan ukuran objek (small/medium/large) untuk kelas text dan face

Berbeda dengan kelas *text*, pada kelas *face* nilai *AP-small* seluruh model tetap rendah (berkisar 0,156-0,182) dan keunggulan YOLO26 atas RT-DETR-L tidak sebesar pada teks. Hal ini kemungkinan disebabkan oleh karakteristik wajah pada manga yang sering berukuran sangat kecil, bergaya gambar bervariasi, dan memiliki kontras rendah, sehingga deteksi wajah kecil tetap menjadi tantangan terlepas dari strategi penetapan label. Temuan ini menunjukkan bahwa keuntungan STAL lebih menonjol pada kelas teks dibandingkan kelas wajah.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3.5. Trade-off Akurasi, Kecepatan, dan Efisiensi

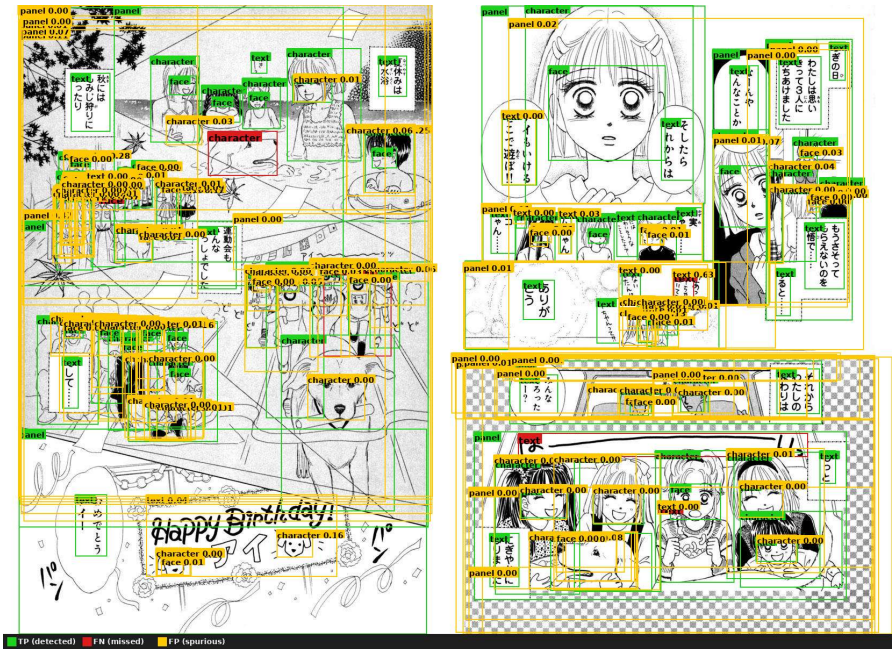
Hubungan antara akurasi, kecepatan, dan ukuran model ditunjukkan pada Gambar 11. Ketiga varian YOLO26 berada pada *frontier* yang lebih unggul dibandingkan RT-DETR-L: untuk setiap tingkat akurasi, YOLO26 menawarkan kecepatan inferensi yang lebih tinggi, jumlah parameter yang lebih kecil, dan waktu pelatihan yang lebih pendek. RT-DETR-L berada di bawah *frontier* tersebut, yaitu memiliki akurasi lebih rendah sekaligus lebih lambat, lebih besar, dan jauh lebih mahal untuk dilatih. Dengan demikian, pada tugas deteksi elemen Manga109, RT-DETR-L tidak menawarkan keunggulan *trade-off* dibandingkan YOLO26.



Gambar 11. Trade-off akurasi (mAP@0.5) terhadap kecepatan (FPS)

3.6. Analisis Kesalahan

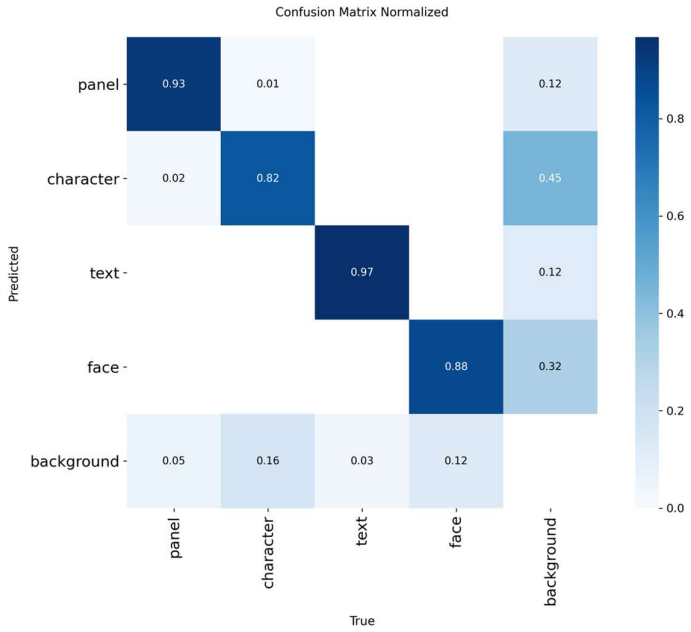
Pemeriksaan kualitatif dilakukan terhadap kasus dengan kinerja deteksi rendah untuk memahami pola kesalahan model. Gambar 12 menampilkan contoh kesalahan pada teks berukuran kecil. Pola kesalahan yang teramati antara lain teks kecil yang tidak terdeteksi (*false negative*), deteksi ganda atau kotak pembatas yang tidak presisi (*false positive*) pada halaman dengan kepadatan objek tinggi, serta kesulitan pada panel dengan latar bertekstur atau *screen tone*. Pola ini konsisten dengan analisis kuantitatif pada Bagian 3.4, yaitu objek berukuran kecil, khususnya teks dan wajah, merupakan sumber kesalahan utama.



Gambar 12. Contoh kesalahan deteksi pada teks berukuran kecil (hijau: TP, merah: FN, oranye: FP)

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Untuk melengkapi analisis kualitatif tersebut, dihitung pula *confusion matrix* empat kelas ditambah kelas *background* pada split uji guna mengukur pola kesalahan secara kuantitatif. Hasilnya menunjukkan bahwa moda kegagalan dominan pada seluruh model adalah kegagalan deteksi (objek yang seharusnya terdeteksi justru diprediksi sebagai *background*), bukan kesalahan klasifikasi antar-kelas; kebingungan antar-kelas maksimum hanya 0,02, yang mengindikasikan keempat kelas terdefinisi dengan baik secara visual. Kelas character memiliki laju kegagalan tertinggi (0,45 pada YOLO26m dan 0,42 pada RT-DETR-L), konsisten dengan tubuh karakter yang sering terpotong tepi panel atau saling tumpang-tindih. Sebaliknya, kelas text paling andal dideteksi (laju kegagalan 0,07-0,13) karena kontras tinggi karakter cetak terhadap latar putih. Untuk kelas face yang relevan dengan analisis objek kecil, RT-DETR-L justru memiliki laju kegagalan lebih rendah (0,24) dibandingkan YOLO26m (0,32); namun keunggulan ini tidak tercermin pada AP-small karena perbedaan *trade-off precision-recall*, yaitu YOLO26m menghasilkan lebih banyak *false positive* pada wajah yang tertekan ketika ambang keyakinan ditinggikan. Gambar 13 menampilkan *confusion matrix* ternormalisasi untuk YOLO26m sebagai model dengan kinerja terbaik.



Gambar 13. Confusion Matrix Ternormalisasi YOLO26m pada Split Uji (Baris: Kelas Acuan, Kolom: Kelas Prediksi)

3.7. Pembahasan

Secara keseluruhan, hasil penelitian menunjukkan tiga temuan utama. Pertama, YOLO26 mengungguli RT-DETR-L pada deteksi elemen Manga109 baik dari sisi akurasi, kecepatan, efisiensi parameter, maupun biaya pelatihan. Temuan ini menarik karena RT-DETR pada awalnya diajukan sebagai detektor yang mengungguli YOLO pada citra natural [9], namun keunggulan tersebut tidak terlihat pada domain manga dalam penelitian ini. Kedua, keunggulan YOLO26 paling menonjol pada deteksi objek kecil, khususnya teks, yang konsisten dengan klaim desain STAL [10]. Ketiga, RT-DETR-L mengalami ketidakstabilan pelatihan yang signifikan pada domain ini, sehingga membutuhkan jumlah *epoch* yang jauh lebih banyak untuk konvergen.

Jika dibandingkan dengan penelitian terdahulu, hasil ini memperlihatkan pola yang menarik. Pada citra natural, RT-DETR diklaim mampu menyaingi bahkan mengungguli detektor YOLO [9], dan studi perbandingan pada domain citra umum [7], [11] menempatkan RT-DETR sebagai pesaing yang kompetitif. Namun, keunggulan tersebut tidak terbukti pada domain manga: seluruh varian YOLO26, termasuk YOLO26n yang jauh lebih ringan, secara konsisten mengungguli RT-DETR-L. Perbedaan arah temuan ini kemungkinan dipengaruhi oleh karakteristik domain manga yang berbeda dari citra natural, yaitu dominasi objek berukuran sangat kecil dan padat seperti teks, citra monokrom, serta variasi gaya gambar antarjudul. Karakteristik tersebut cenderung menguntungkan mekanisme STAL dan prediksi padat berbasis konvolusi pada resolusi tinggi yang diterapkan YOLO26, sementara pendekatan *set prediction* berbasis atensi global pada RT-DETR-L tampak lebih sulit dioptimalkan pada pola citra yang repetitif dan padat, sebagaimana tercermin dari kebutuhan *epoch* pelatihan yang jauh lebih banyak. Dibandingkan penelitian deteksi elemen komik terdahulu yang umumnya berfokus pada satu

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 314-325
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

arsitektur seperti SSD [3], penelitian ini menyediakan perbandingan multidimensi antar-paradigma detektor *real-time* pada dataset yang sama. Ringkasan perbandingan penelitian ini dengan studi terdahulu disajikan pada Tabel 8.

Tabel 8. Perbandingan penelitian ini dengan studi detektor terdahulu
(hasil pada domain, dataset, dan protokol berbeda sehingga tidak dapat dibandingkan secara langsung)

Ref.	Studi	Domain / Dataset	Fokus Perbandingan	Temuan Relatif terhadap Penelitian Ini
[11]	González Hernández dkk.	Citra umum (dataset HORD)	YOLOv8-v11 vs RT-DETR	YOLOv8l unggul (mAP50 0,918; mAP50-95 0,703) di atas RT-DETR-L (mAP50 0,783; mAP50-95 0,550); arah serupa dengan temuan pada manga.
[7]	Le dkk.	Pascal VOC	YOLO vs RT-DETR	YOLOv8x (mAP50-95 0,480; 133 FPS) sedikit di atas RT-DETRv1 (mAP50-95 0,475; 100 FPS); domain berbeda dari komik.
[3]	Ogawa dkk.	Manga109	SSD300-fork (arsitektur tunggal)	mAP 84,2% untuk empat elemen komik; baseline satu arsitektur dengan protokol dan metrik yang tidak setara langsung.
—	Penelitian ini	Manga109	YOLO26m	mAP@0.5 0,9366; mAP@0.5:0.95 0,7471 (terbaik keseluruhan).
—	Penelitian ini	Manga109	RT-DETR-L	mAP@0.5 0,9218; mAP@0.5:0.95 0,7001 (terendah) meski parameter terbesar.

Hasil ini perlu ditafsirkan secara hati-hati. Penelitian ini hanya menggunakan satu dataset (Manga109) dan satu varian RT-DETR (RT-DETR-L), sehingga generalisasi terhadap dataset komik lain maupun varian RT-DETR lain masih memerlukan pengujian lebih lanjut. Selain itu, evaluasi ini menggunakan satu skema pembagian data *by-title* (seed=42) tanpa validasi silang *k-fold*; validasi silang *by-title* (k=5) idealnya diperlukan untuk melaporkan nilai rata-rata dan standar deviasi (mean ± std) mAP yang lebih *robust*, tetapi biaya komputasinya yang besar (estimasi sekitar 298 jam GPU untuk keempat model pada satu GPU) menjadikannya tidak *feasible* dalam cakupan penelitian ini, sehingga ditinggalkan sebagai agenda penelitian lanjutan. Selain itu, fluktuasi validasi RT-DETR-L kemungkinan berkaitan dengan instabilitas konfigurasi pelatihan pada domain manga dan masih perlu diverifikasi ulang, misalnya melalui penyetelan *hyperparameter* yang lebih ekstensif atau pengujian ulang dengan konfigurasi AMP yang berbeda.

4. KESIMPULAN

Penelitian ini membandingkan kinerja YOLO26 (varian *nano*, *small*, dan *medium*) dengan RT-DETR-L untuk deteksi empat elemen komik, yaitu *panel*, *character*, *text*, dan *face*, pada dataset Manga109. YOLO26m memperoleh kinerja keseluruhan tertinggi dengan mAP@0.5 sebesar 0,9366 dan mAP@0.5:0.95 sebesar 0,7471, sedangkan RT-DETR-L memperoleh mAP@0.5:0.95 terendah (0,7001) meskipun memiliki jumlah parameter terbesar, kecepatan inferensi paling rendah, dan waktu pelatihan paling panjang. Bahkan YOLO26n dengan 2,51 juta parameter memperoleh mAP@0.5 yang sedikit lebih tinggi daripada RT-DETR-L pada kecepatan sekitar 6,5 kali lebih cepat.

Temuan utama penelitian terdapat pada deteksi objek berukuran kecil. Pada kelas teks, YOLO26m mengungguli RT-DETR-L dalam *AP-small* sebesar 42,7% relatif (0,6722 berbanding 0,4709), dan selisih AP antara teks objek besar dan kecil menyempit secara monoton seiring bertambahnya ukuran model YOLO26. Pola ini memberikan bukti empiris yang mendukung klaim desain *Small-Target-Aware Label assignment* (STAL) pada YOLO26 untuk domain manga. Selain itu, RT-DETR-L menunjukkan ketidakstabilan pelatihan, yaitu memerlukan 92 *epoch* untuk mencapai 95% dari mAP finalnya dan waktu latih sekitar 37,6 jam, sedangkan seluruh varian YOLO26 mencapai 95% dalam paling lama 4 *epoch* dengan waktu latih sekitar 5,0-10,6 jam.

Kontribusi utama penelitian ini adalah menyediakan *benchmark* multidimensi YOLO26 versus RT-DETR-L pada Manga109 beserta bukti empiris efektivitas STAL pada deteksi teks berukuran kecil, yang dapat menjadi acuan praktis dalam pemilihan detektor *real-time* untuk analisis citra komik.

Penelitian ini memiliki beberapa keterbatasan. Evaluasi hanya dilakukan pada satu dataset (Manga109) dan satu varian RT-DETR (RT-DETR-L), sehingga generalisasi terhadap dataset komik lain maupun varian RT-DETR lain belum dapat diklaim. Ketidakstabilan pelatihan RT-DETR-L kemungkinan masih dapat dikurangi melalui penyetelan *hyperparameter* yang lebih ekstensif. Selain itu, deteksi wajah berukuran kecil tetap menjadi tantangan bagi seluruh model.

Penelitian selanjutnya disarankan untuk menguji varian RT-DETR yang lebih baru seperti RT-DETRv2 [13], memperluas evaluasi ke beberapa dataset komik agar kemampuan generalisasi lintas-gaya dapat diukur, serta mengeksplorasi penambahan modul atensi seperti CBAM [14] pada kepala deteksi untuk meningkatkan akurasi pada objek berukuran kecil. Evaluasi lanjutan juga dapat mempertimbangkan skema *multi-task learning* untuk

analisis komik seperti Comic MTL [15], validasi silang *k-fold by-title* untuk estimasi generalisasi yang lebih *robust*, serta pembaruan anotasi Manga109-v2026 [16] untuk menguji apakah temuan ini bertahan pada anotasi yang lebih baru.

DAFTAR PUSTAKA

[1] K. Aizawa et al., "Building a Manga Dataset 'Manga109' with Annotations for Multimedia Applications," *IEEE MultiMedia*, vol. 27, no. 2, pp. 8-18, 2020, doi: 10.1109/MMUL.2020.2987895.

[2] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based Manga Retrieval using Manga109 Dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21811-21838, 2017, doi: 10.1007/s11042-016-4020-z.

[3] T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki, and K. Aizawa, "Object Detection for Comics using Manga109 Annotations," arXiv:1803.08670, 2018.

[4] E. Vivoli, M. Bertini, and D. Karatzas, "CoMix: A Comprehensive Benchmark for Multi-Task Comic Understanding," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2024.

[5] S. Paval, P. Meißner, and I. P. Yamshchikov, "ComicScene154: A Scene Dataset for Comic Analysis," in*Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2025, pp. 31574-31580, doi: 10.18653/v1/2025.emnlp-main.1608.

[6] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey,"*Mathematical Biosciences and Engineering*, vol. 20, no. 4, pp. 6551-6590, 2023, doi: 10.3934/mbe.2023282.

[7] N.-T. Le, N. T. Thai, and C. V. Bui, "Benchmarking Real-Time Object Detection: Evaluating YOLO and RT-DETR on Speed, Accuracy, and Efficiency," in*Information and Communication Technology (SOICT 2024), Communications in Computer and Information Science*. Springer, 2025, pp. 224-234, doi: 10.1007/978-981-96-4285-4_19.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proc. ECCV*, 2020.

[9] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs Beat YOLOs on Real-time Object Detection," in *Proc. IEEE/CVF CVPR*, 2024.

[10] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection," arXiv:2509.25164, 2026.

[11] A. J. González Hernández, J. P. Sánchez Hernández, D. L. Hernández Rabadán, J. Frausto Solis, and J. González Barbosa, "An analysis of YOLO models versus RT-DETR applied to multi-object detection in images,"*International Journal of Combinatorial Optimization Problems and Informatics*, vol. 16, no. 3, pp. 360-377, 2025, doi: 10.61467/2007.1558.2025.v16i3.778.

[12] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in *Proc. ECCV*, 2014.

[13] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu, "RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer," arXiv:2407.17140, 2024.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. ECCV*, 2018.

[15] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, "Comic MTL: Optimized Multi-task Learning for Comic Book Image Analysis," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, pp. 265-284, 2019.

[16] J. Baek, A. Miyai, S. Onohara, H. Ikuta, and K. Aizawa, "Manga109-v2026: Revisiting Manga109 Annotations for Modern Manga Understanding," in *Proc. Culture × AI Workshop at ICML*, 2026.