BAB 2
LANDASAN TEORI

2.1 Computer Vision dalam Konteks Layanan Restoran

Computer Vision merupakan cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk memahami dan menafsirkan informasi visual dari dunia nyata dalam bentuk citra atau video [2]. Berbeda dengan pendekatan sensor fisik konvensional yang berbiaya tinggi dan tidak fleksibel, sistem berbasis *computer vision* memanfaatkan kamera sebagai sensor utama sehingga lebih mudah dipasang, dikonfigurasi, dan dikembangkan sesuai kebutuhan [1]. Dalam konteks industri restoran, teknologi ini membuka peluang untuk mengotomasi pemantauan kondisi meja makan secara objektif, konsisten, dan *real-time* tanpa memerlukan intervensi manusia yang terus-menerus [1]. Penerapan *computer vision* berbasis kamera pada lingkungan restoran terbukti mampu meningkatkan efisiensi operasional manajemen meja secara signifikan dibandingkan dengan sistem manual maupun sistem berbasis sensor fisik [1].

2.2 Deep Learning untuk Deteksi Objek

Deep Learning merupakan subbidang dari *machine learning* yang menggunakan jaringan saraf tiruan berlapis banyak (*deep neural networks*) untuk mempelajari representasi data secara hierarkis dan otomatis dari data mentah [2]. Arsitektur *Convolutional Neural Network* (CNN) menjadi fondasi utama dalam model *deep learning* untuk visi komputer, karena kemampuannya mengekstraksi fitur spasial dari citra melalui operasi konvolusi berlapis yang semakin abstrak pada setiap lapisan. CNN telah terbukti mampu mendeteksi kondisi objek secara otomatis pada berbagai domain aplikasi industri, mencakup inspeksi visual otomatis hingga deteksi cacat pada lini produksi [3]. Dalam konteks deteksi objek pada domain makanan dan peralatan makan, *deep learning* terbukti mampu mengenali perbedaan kondisi visual yang halus dan beragam, seperti perbedaan antara piring yang bersih, aktif digunakan, dan kotor dengan sisa makanan, bahkan dalam kondisi pencahayaan dan sudut pandang kamera yang bervariasi [11]. Penelitian yang dilakukan oleh Alahmari dan Salem membuktikan bahwa *deep learning* mampu mengenali kondisi makanan (*food state recognition*) dengan akurasi yang tinggi menggunakan dataset yang relatif terbatas, yang mendukung kelayakan penerapan

pendekatan serupa pada domain peralatan makan restoran [4].

2.3 Object Detection

Object detection adalah teknik dalam *computer vision* yang bertujuan untuk mendeteksi keberadaan satu atau lebih objek dalam suatu citra sekaligus menentukan lokasi objek tersebut dalam bentuk *bounding box* [2]. Metode ini menggabungkan dua tugas utama secara simultan, yaitu klasifikasi objek (apa yang terdeteksi) dan lokalisasi objek (di mana posisinya dalam gambar). Terdapat dua pendekatan utama dalam *object detection*: *two-stage detector* seperti Faster R-CNN yang memisahkan tahap proposal wilayah dan klasifikasi sehingga lebih akurat namun lebih lambat, serta *single-stage detector* seperti YOLO yang melakukan kedua tugas dalam satu tahap pemrosesan sehingga jauh lebih cepat dan lebih sesuai untuk kebutuhan inferensi *real-time* [2, 5]. Untuk aplikasi pemantauan meja makan restoran yang membutuhkan respons segera, pendekatan *single-stage detector* merupakan pilihan yang lebih tepat karena mampu memproses *frame* kamera secara langsung tanpa penundaan yang berarti [10].

2.4 Evolusi Algoritma YOLO

You Only Look Once (YOLO) adalah keluarga algoritma *object detection* berbasis *deep learning* yang pertama kali diperkenalkan pada tahun 2016 [2]. YOLO memformulasikan deteksi objek sebagai masalah regresi tunggal: citra dibagi menjadi *grid* sel dan model secara langsung memprediksi *bounding box* serta probabilitas kelas untuk setiap sel dalam satu kali proses *forward pass*. Pendekatan ini memungkinkan deteksi yang sangat cepat karena seluruh gambar diproses sekaligus tanpa tahap proposal wilayah yang terpisah. Sejak versi pertamanya, YOLO telah mengalami evolusi yang pesat melalui berbagai inovasi arsitektur [6]. Tinjauan komprehensif terhadap perkembangan algoritma YOLO menunjukkan bahwa setiap versi baru secara konsisten berupaya meningkatkan keseimbangan antara akurasi deteksi, kecepatan inferensi, dan efisiensi parameter [6, 5] tiga dimensi yang menjadi dasar perbandingan dalam penelitian ini.

2.5 YOLOv8

YOLOv8 dirilis oleh Ultralytics pada Januari 2023 dan menjadi salah satu versi YOLO yang paling banyak diadopsi dalam penelitian terapan hingga

saat ini [7]. YOLOv8 memperkenalkan beberapa inovasi arsitektur utama. Pertama, arsitektur *anchor-free* yang memprediksi pusat objek secara langsung tanpa bergantung pada *anchor box* yang telah dikonfigurasi sebelumnya, sehingga lebih adaptif terhadap objek dengan bentuk dan rasio aspek yang bervariasi [7]. Kedua, modul C2f (*Cross Stage Partial with two convolutions*) yang menggantikan modul C3 pada pendahulunya, memberikan aliran gradien yang lebih kaya dan representasi fitur yang lebih baik. Ketiga, *decoupled head* yang memisahkan jalur prediksi klasifikasi dan regresi *bounding box* secara independen sehingga mengurangi interferensi antar tugas [2]. Pada *benchmark* MS COCO, YOLOv8s mencapai 44.9% AP dengan kecepatan inferensi yang memadai untuk aplikasi *real-time* [7]. Dalam penelitian komparatif pada domain makanan, YOLOv8 menunjukkan performa yang kuat dan menjadi *baseline* yang umum digunakan [11]. YOLOv8 dipilih sebagai salah satu model dalam penelitian ini karena merupakan *baseline* yang paling matang, paling banyak divalidasi oleh komunitas peneliti, dan memiliki ekosistem *tooling* yang stabil [7].

2.6 YOLOv9

YOLOv9 dipresentasikan pada *European Conference on Computer Vision* (ECCV) 2024 dan memperkenalkan dua kontribusi arsitektur utama yang dirancang untuk mengatasi permasalahan fundamental pada jaringan dalam [8]. Pertama, *Programmable Gradient Information* (PGI), yaitu mekanisme yang memungkinkan model mempertahankan informasi gradien yang lengkap sepanjang proses pelatihan dengan cara menyediakan jalur informasi tambahan dari *input* ke *output*. PGI mengatasi masalah *information bottleneck* yang umum terjadi pada jaringan dengan banyak lapisan, di mana informasi penting dari *input* cenderung hilang setelah melewati banyak transformasi [8]. Kedua, *Generalized Efficient Layer Aggregation Network* (GELAN), yaitu arsitektur jaringan yang dirancang berdasarkan perencanaan jalur gradien untuk mengoptimalkan penggunaan parameter dan operasi komputasi secara bersamaan. Kombinasi PGI dan GELAN memungkinkan YOLOv9 mencapai peningkatan performa dibandingkan YOLOv8 dengan jumlah parameter yang lebih sedikit pada *benchmark* MS COCO [8]. Namun, tinjauan komparatif menunjukkan bahwa YOLOv9 cenderung memiliki kecepatan inferensi yang lebih lambat dibandingkan YOLOv8 dan YOLOv11 pada skala model yang setara [12].

2.7 YOLOv11

YOLOv11 dirilis oleh Ultralytics pada Oktober 2024 dan merupakan penerus langsung dari YOLOv8 dalam ekosistem Ultralytics [9]. YOLOv11 memperkenalkan beberapa peningkatan arsitektur yang signifikan. Pertama, blok C3k2 sebagai pengganti C2f, yang menggunakan *kernel* konvolusi 3×3 yang lebih kecil namun lebih banyak untuk meningkatkan aliran gradien dan efisiensi komputasi [9]. Kedua, modul C2PSA (*Convolutional block with Parallel Spatial Attention*) yang mengintegrasikan mekanisme *spatial attention* untuk memperhalus representasi fitur spasial dan meningkatkan kemampuan model dalam memperhatikan bagian citra yang relevan [9]. Ketiga, modul SPPF (*Spatial Pyramid Pooling-Fast*) yang diperbarui untuk menangkap informasi kontekstual pada berbagai skala secara efisien. Peningkatan-peningkatan ini memungkinkan YOLOv11 mencapai akurasi yang lebih tinggi dengan jumlah parameter yang lebih sedikit dibandingkan YOLOv8 pada ukuran model yang setara [9]. Dalam berbagai evaluasi komparatif, YOLOv11 secara konsisten menunjukkan keseimbangan terbaik antara akurasi dan efisiensi komputasi dibandingkan versi YOLO lainnya [12].

2.8 Perbandingan Arsitektur YOLOv8, YOLOv9, dan YOLOv11

Ketiga versi YOLO yang digunakan dalam penelitian ini memiliki perbedaan arsitektur yang mendasar, yang dirangkum pada Tabel 2.1. Perbandingan ini didasarkan pada hasil *benchmark* menggunakan dataset MS COCO sebagai standar evaluasi yang umum digunakan dalam komunitas *object detection* [6].

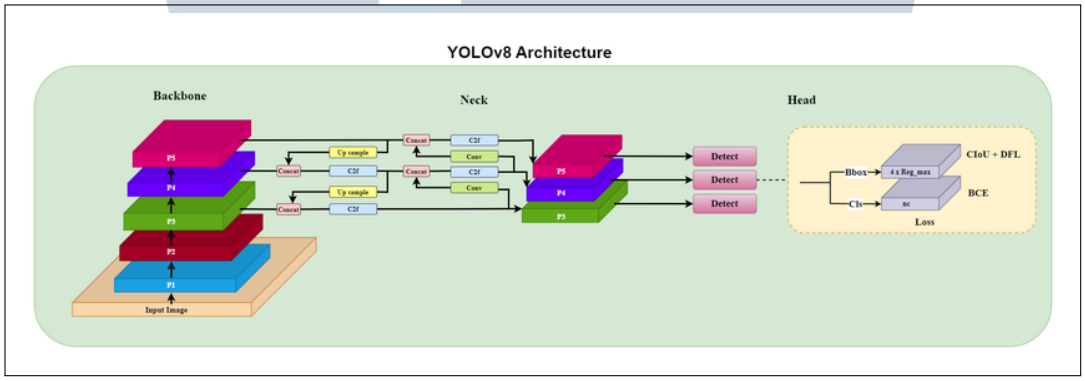
Tabel 2.1. Perbandingan arsitektur YOLOv8s, YOLOv9s, dan YOLOv11s berdasarkan *benchmark* MS COCO

Aspek	**YOLOv8s**	**YOLOv9s**	**YOLOv11s**
Tahun rilis	2023	2024	2024
Modul utama	C2f	GELAN + PGI	C3k2 + C2PSA
Strategi deteksi	*Anchor-free*	*Anchor-free*	*Anchor-free*
Parameter	11.2M	7.1M	9.4M
GFLOPs	28.6	26.4	21.5
mAP@50-95 (COCO)	44.9%	46.8%	47.0%

Sumber: [7, 8, 9]

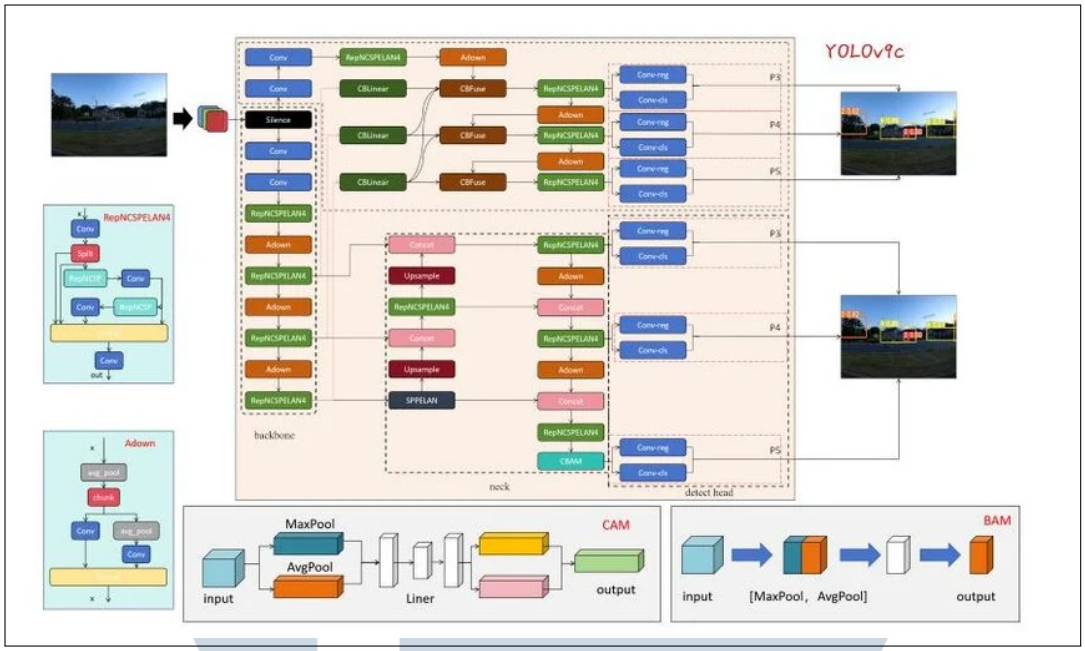
Perbedaan arsitektur ini menunjukkan karakteristik *trade-off* yang berbeda: YOLOv8s sebagai *baseline* yang matang; YOLOv9s dengan akurasi lebih tinggi namun kompleksitas pelatihan yang lebih tinggi; dan YOLOv11s dengan efisiensi parameter terbaik dan akurasi tertinggi [12].

Untuk memahami karakteristik struktural dari masing-masing model secara lebih mendalam, visualisasi diagram arsitektur dari setiap varian dijabarkan sebagai berikut. Arsitektur YOLOv8s yang dikembangkan oleh Ultralytics mengadopsi struktur *backbone* dan *neck* tradisional yang telah diperbarui menggunakan modul C2f (*Cross Stage Partial with two convolutions*), serta mengimplementasikan *decoupled head* untuk memisahkan proses regresi kotak pembatas dan klasifikasi kelas objek secara independen sebagaimana tertera pada Gambar 2.1.



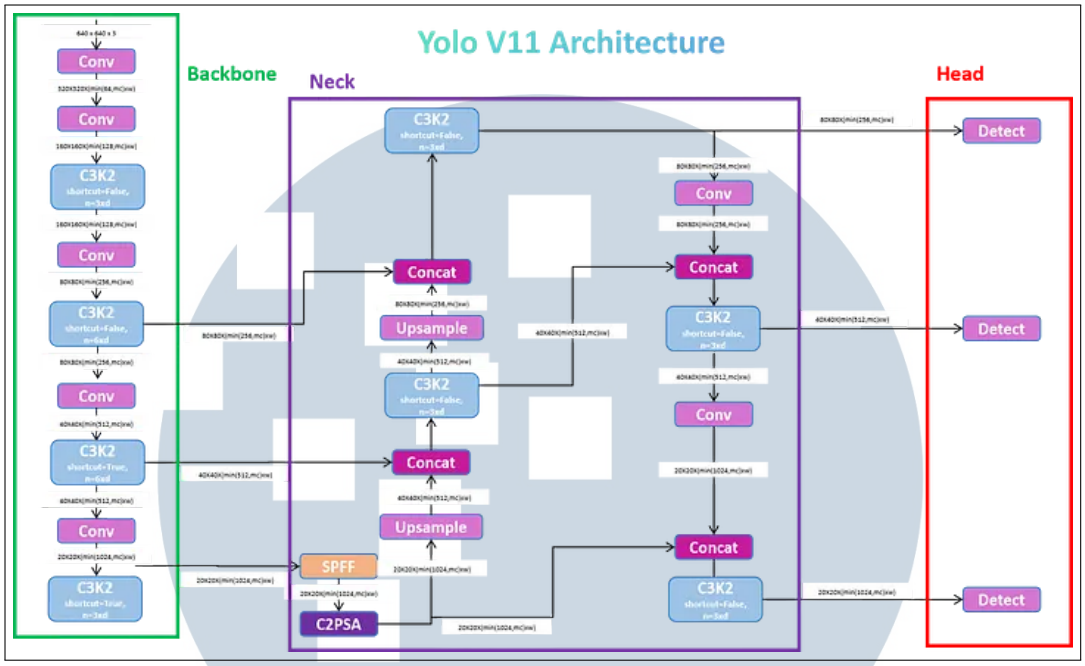
Gambar 2.1. Diagram Struktur Arsitektur Jaringan YOLOv8s

Sementara itu, YOLOv9s memperkenalkan pergeseran paradigma struktural yang signifikan melalui integrasi komponen *Generalized Efficient Layer Aggregation Network* (GELAN) dan mekanisme *Programmable Gradient Information* (PGI). Komponen GELAN mengoptimalkan perencanaan jalur gradien dengan menggabungkan komputasi berbasis blok konvolusi konvensional untuk mempertahankan kelengkapan fitur dari lapisan *input* menuju lapisan *output*. Visualisasi susunan blok arsitektur YOLOv9s beserta integrasi komponen internalnya ditunjukkan pada Gambar 2.2.



Gambar 2.2. Diagram Struktur Arsitektur Jaringan YOLOv9s

Sebagai evolusi mutakhir, YOLOv11s melakukan rekonstruksi pada tingkat efisiensi komputasi dengan mengganti modul C2f menggunakan blok C3k2, yang mengombinasikan ukuran *kernel* lebih ringkas untuk mereduksi parameter komputasi tanpa mendegradasi akurasi ekstraksi. Selain itu, model ini mengintegrasikan modul *Convolutional block with Parallel Spatial Attention* (C2PSA) yang bertindak sebagai mekanisme atensi spasial paralel untuk memperkuat fokus jaringan pada area objek yang relevan di dalam citra. Struktur lengkap dari konfigurasi YOLOv11s tersebut diilustrasikan pada Gambar 2.3.



Gambar 2.3. Diagram Struktur Arsitektur Jaringan YOLOv11s

Melalui analisis spasial pada ketiga diagram arsitektur di atas, dapat ditarik pemahaman teoretis bahwa evolusi dari YOLOv8 menuju YOLOv11 tidak sekadar berfokus pada penambahan kedalaman jaringan, melainkan pada optimalisasi aliran gradien informasi dan penyematan modul atensi. Pemilihan varian *small* (s) pada ketiga model ini menjamin bahwa perbandingan kinerja komparatif yang diuji pada tugas deteksi kondisi peralatan makan restoran berjalan di atas fondasi komputasi yang adil dan setara.

2.9 Deteksi Objek pada Domain Makanan dan Peralatan Makan

Penerapan *deep learning* untuk deteksi dan pengenalan kondisi objek pada domain makanan telah menjadi area penelitian yang berkembang. Alahmari dan Salem mengembangkan sistem pengenalan kondisi makanan (*food state recognition*) menggunakan *deep learning* pada dataset makanan dengan berbagai kondisi visual, dan menunjukkan bahwa CNN mampu membedakan kondisi makanan secara akurat meskipun variasi visual antar kondisi yang relatif kecil [4]. Chrintz-Gath et al. melakukan evaluasi komparatif antara YOLOv7, YOLOv8, dan YOLOv9 untuk pengenalan komponen makanan berdasarkan model piring kesehatan pada dataset 3.707 gambar dengan 42 kelas makanan, dan menemukan bahwa YOLOv8 mengungguli YOLOv7 dan YOLOv9 dalam hal *peak precision* dan

F1-score pada domain tersebut [11]. Temuan ini memperkuat relevansi YOLOv8 sebagai *baseline* yang kuat dalam penelitian ini. Namun, penelitian tersebut tidak menyertakan YOLOv11 yang dirilis setelahnya, dan tidak membandingkan kecepatan inferensi maupun ukuran model secara eksplisit — dua dimensi yang menjadi fokus utama penelitian ini. Kesenjangan ini menegaskan kebutuhan akan studi komparatif yang lebih komprehensif dan terkini pada domain yang lebih spesifik, yaitu peralatan makan di lingkungan restoran.

2.10 Transfer Learning pada Model YOLO

Transfer learning adalah teknik pelatihan model di mana bobot dari model yang telah dilatih pada dataset berskala besar digunakan sebagai inisialisasi awal sebelum *fine-tuning* pada dataset yang lebih kecil dan spesifik [2]. Dalam konteks penelitian ini, ketiga model YOLOv8s, YOLOv9s, dan YOLOv11s diinisialisasi dengan *pretrained weights* dari dataset MS COCO sebelum dilatih ulang pada dataset deteksi kondisi peralatan makan. Pendekatan ini memungkinkan model memanfaatkan representasi fitur visual umum yang telah dipelajari dari jutaan gambar, sehingga konvergensi lebih cepat dan performa pada dataset kustom yang relatif kecil menjadi lebih optimal [2]. Penelitian terdahulu menunjukkan bahwa dengan teknik *transfer learning*, model YOLOv8 mampu mencapai hasil deteksi yang hampir sempurna bahkan dengan dataset pelatihan yang hanya terdiri dari 100 hingga 350 gambar [14]. Selain itu, *fine-tuning* berbasis *transfer learning* memungkinkan pelatihan model deteksi objek menggunakan kurang dari 1.000 gambar per kelas pada perangkat keras komputasi biasa [16].