BAB II

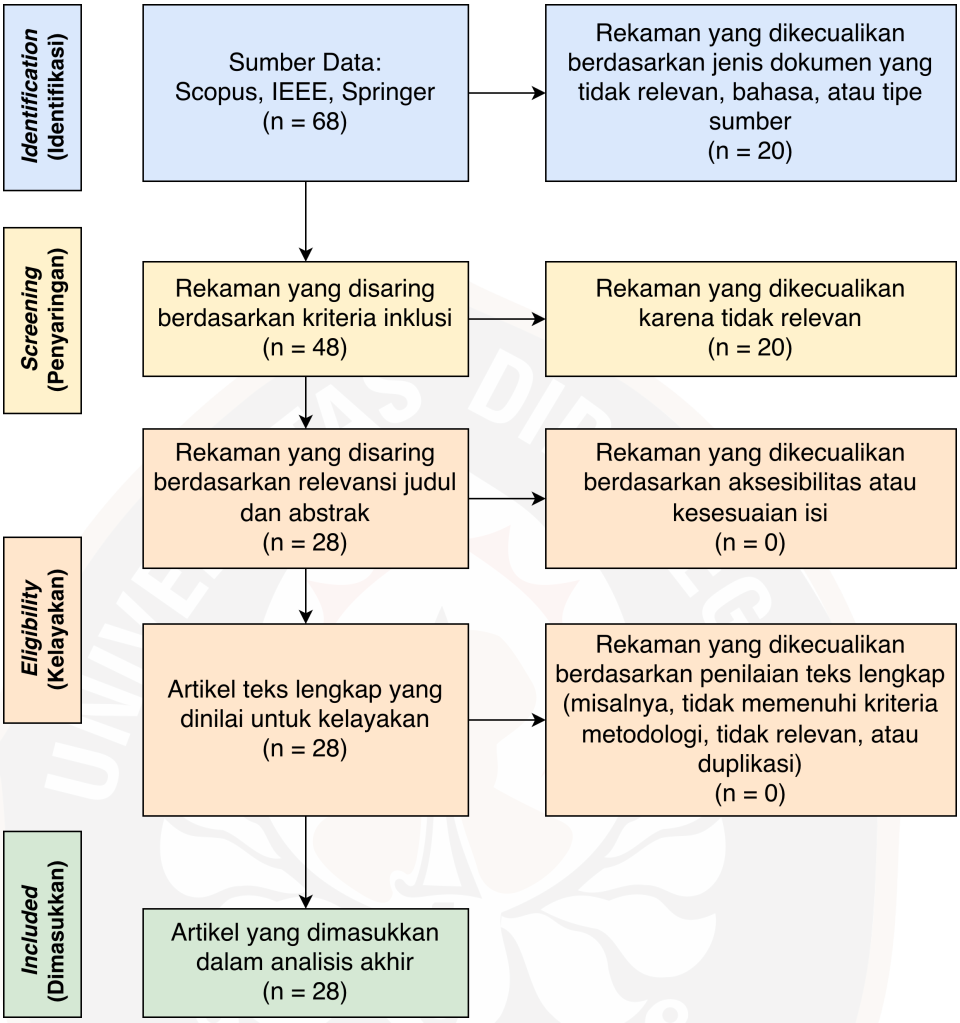
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka ini disusun menggunakan pendekatan *Systematic Literature Review* (SLR) untuk memetakan perkembangan penelitian terkait identifikasi plankton berbasis *deep learning*, termasuk temuan utama, tantangan, dan celah penelitian (Kitchenham et al., 2010; Kitchenham and Charters, 2007; Okoli, 2015). Proses SLR mencakup pencarian pada database bereputasi seperti Scopus, IEEE Xplore, dan SpringerLink, lalu dilanjutkan tahap *screening*, *eligibility*, dan seleksi akhir berdasarkan relevansi terhadap deteksi plankton berbasis citra mikroskop digital. Kata kunci seperti *plankton*, *deep learning*, *classification*, *detection*, dan *phytoplankton* digunakan untuk menjaga cakupan, sementara Gambar 2.1 merangkum alur penyaringan bertahap dari hasil penelusuran awal hingga artikel yang paling sesuai.

Kajian literatur menunjukkan bahwa identifikasi plankton umumnya mencakup tiga tugas utama, yaitu klasifikasi citra tunggal, deteksi objek, dan segmentasi spasial, yang telah diuji pada beragam dataset plankton dan organisme akuatik (Culverhouse et al., 2006; Benfield et al., 2007; Orenstein et al., 2015; Luo et al., 2018; Yue et al., 2023; Irisson et al., 2022). Pemetaan ini juga menyoroti variasi teknik pencitraan, sumber dataset, serta peran augmentasi dan strategi pengendalian variabilitas citra yang memengaruhi stabilitas performa model (Eerola et al., 2024). Dalam konteks segmentasi, U-Net tetap menonjol karena efektif menghasilkan pemetaan struktur yang presisi, khususnya pada plankton dengan kontur halus dan detail internal kompleks (Ronneberger et al., 2015; Ardhi et al., 2024a).

Perkembangan mutakhir memperlihatkan pergeseran dari arsitektur CNN murni menuju model hibrida dan detektor transformer end-to-end seperti DETR, Deformable-DETR, dan RT-DETR yang lebih adaptif terhadap objek kecil serta struktur biologis kompleks (Carion et al., 2020a; Zhu et al., 2021; Guemas et al., 2024; Wang et al., nd; Zhao et al., 2024a). CNN tetap unggul dalam menangkap fitur lokal yang detail, sementara *Vision Transformer* (ViT) memperkaya pemahaman



Gambar 2.1 Tahapan pemilihan studi

konteks global, sehingga integrasi keduanya menjadi tren yang menjanjikan pada domain plankton dan ekologi akuatik (LeCun et al., 2015; Goodfellow et al., 2016; Dosovitskiy et al., 2021; Kyathanahally et al., 2022). Namun, kemiripan morfologi antarspesies, variasi bentuk yang tinggi, perbedaan kualitas pencitraan, serta ketidakseimbangan kelas dataset masih menjadi hambatan utama (Luo et al., 2018; Li et al., 2021; Lumini and Nanni, 2019; MacNeil et al., 2021; Maracani et al., 2023; Eerola et al., 2024). Celah inilah yang menunjukkan bahwa masih terbatas kajian yang mengevaluasi secara sistematis integrasi mekanisme ekstraksi fitur lokal yang lebih adaptif seperti *asymmetric convolution* dengan representasi konteks global berbasis transformer dalam backbone model deteksi modern, termasuk keluarga YOLOv8, terutama dalam menyeimbangkan akurasi, efisiensi komputasi, dan kompleksitas arsitektur (Shorten and Khoshgoftaar, 2019; Taylor and Nitschke,

2018; Gupta et al., 2019; Schröder et al., 2020; Schmarje et al., 2021; Nanni et al., 2023).

2.1.1 Teknik Pencitraan dan Sumber Data dalam Klasifikasi Plankton

Perkembangan identifikasi plankton berbasis kecerdasan buatan sangat bergantung pada kemajuan teknik pencitraan digital serta meningkatnya ketersediaan dataset berskala besar yang dapat digunakan untuk melatih model *deep learning* (MacNeil et al., 2021). Berbagai teknik pencitraan modern memungkinkan pengumpulan citra plankton dengan resolusi tinggi dan volume data yang cukup besar, sehingga mampu mendukung pelatihan model deteksi dan klasifikasi dengan kompleksitas arsitektur yang lebih tinggi. Secara umum, pendekatan pencitraan yang digunakan meliputi *holographic imaging*, *flow-based imaging*, *in situ imaging*, *digital microscopy*, serta sistem pencitraan berbasis deteksi objek, di mana masing-masing teknik memberikan karakteristik tampilan morfologi, kualitas kontras, dan jenis artefak visual yang berbeda (Eerola et al., 2024).

Pendekatan *holographic imaging* seperti Digital Holographic Microscopy (DHM) dan Digital In-line Holographic Microscopy (DIHM) digunakan untuk menangkap citra plankton tanpa preparasi sampel yang kompleks, sekaligus memungkinkan rekonstruksi tiga dimensi struktur organisme (Guo et al., 2021; Lang et al., 2022; MacNeil et al., 2021; Zheng et al., 2017). Teknik ini sangat bermanfaat untuk menganalisis tekstur serta bentuk halus plankton yang sulit ditangkap oleh mikroskop cahaya konvensional. Sistem seperti HoloSea DHM dan berbagai dataset holografi menyediakan citra berkualitas tinggi yang sering dijadikan basis pengembangan model CNN dan transformer, terutama dalam penelitian yang menguji kemampuan model mendeteksi pola kedalaman dan artefak holografi (MacNeil et al., 2021; Schmarje et al., 2021).

Pencitraan berbasis aliran (*flow-based imaging*) seperti Imaging FlowCytobot (IFCB) dan FlowCam Imaging System telah menjadi standar dalam pemantauan komunitas plankton secara otomatis, menghasilkan jutaan citra plankton setiap harinya dalam aliran fluida optik (Kerr et al., 2020; Li et al., 2021; Lumini and Nanni, 2019; Maracani et al., 2023). WHOI-Plankton Dataset, ZooScan Dataset, dan Kaggle Plankton Dataset adalah contoh dataset besar yang berasal dari sistem seperti ini dan menjadi rujukan utama dalam banyak studi klasifikasi dan deteksi (Orenstein et al., 2015; Lumini and Nanni, 2019; Lumini et al., 2020; Wacquet and Lefebvre, 2022). Variasi kondisi lingkungan dan instrumen dalam

dataset tersebut juga memunculkan *domain shift* yang menantang, sehingga relevan bagi desain model yang hendak digeneralisasikan ke perairan tropis Indonesia (Plonus et al., 2021; Pastore et al., 2023).

Teknik *in situ imaging* seperti In Situ Ichthyoplankton Imaging System (ISIIS) dan Zooglider memungkinkan pencitraan plankton langsung di habitatnya, tanpa harus membawa sampel ke laboratorium (Luo et al., 2018; Ellen et al., 2019; Cowen and Guigand, 2008). Pendekatan ini menghasilkan representasi ekologi yang lebih otentik karena citra diambil dalam kondisi air, cahaya, dan turbulensi yang nyata. Sementara itu, *digital microscopy* tetap menjadi tulang punggung dalam penelitian yang membutuhkan analisis detail mikrostruktur, misalnya menggunakan Dual-magnification Scripps Plankton Camera (DSPC) atau kamera CMOS beresolusi tinggi (Kakehi et al., 2021; Kyathanahally et al., 2021, 2022). Pencitraan berbasis mikroskop ini umumnya menghasilkan citra dengan kualitas optik lebih stabil, meskipun cakupan ruang dan waktu lebih terbatas dibanding sistem *in situ* dan *flow-based* (Gupta et al., 2019; Yue et al., 2023).

Pendekatan berbasis deteksi objek juga mengalami perkembangan pesat, terutama pada organisme makroskopik seperti ubur-ubur (*jellyfish*). Model seperti YOLOv4 dan Faster R-CNN telah digunakan untuk mendeteksi fenomena *jellyfish bloom* melalui sistem pencitraan video bawah air, dengan akurasi dan kecepatan yang mendukung pemantauan lapangan secara *real time* (Weihong et al., 2023; Zhang et al., 2024; Ren et al., 2017; Redmon and Farhadi, 2018). Dataset seperti Jellyfish Bloom Dataset memberikan kesempatan untuk menguji model deteksi pada objek biologis yang lebih besar, dinamis, dan bergerak cepat (Weihong et al., 2023; Zhang et al., 2024). Selain itu, dataset lain seperti LOKI, DYB-PlanktonNet, Changjiang, EILAT, RSMAS, EcoTaxa UVP5, dan North Sea VPR menampilkan keberagaman teknik pencitraan dan rentang organisme yang luas, yang dirangkum dalam Lampiran 2 sebagai dasar perbandingan lintas studi (Oldenburg et al., 2023; Schröder et al., 2020; Schmarje et al., 2021; Plonus et al., 2021).

Sumber data lain yang relevan bagi konteks Indonesia adalah Plankton Image Database (cPID) yang dikembangkan oleh Pusat Riset Oseanografi BRIN (Rachman et al., 2022). Dataset ini berisi ribuan citra plankton mikroskopik dari berbagai perairan tropis Indonesia yang diambil menggunakan mikroskop *phase contrast* beresolusi tinggi. Pencitraan *phase contrast* menonjolkan perbedaan fase optik antara struktur internal sel, sehingga menghasilkan pola intensitas yang khas namun kadang tidak stabil pada tepi sel (Grant et al., 2020). Kondisi ini sangat relevan dengan kebutuhan arsitektur yang peka terhadap variasi morfologi dan tekstur,

seperti YOLO dan RT-DETR, serta membuka peluang integrasi dengan pendekatan Quantitative Phase Imaging (QPI) di masa depan (Nguyen et al., 2022; Park et al., 2020).

Selain deteksi objek, pendekatan segmentasi semantik memainkan peran penting dalam memahami bentuk dan struktur fitoplankton secara lebih rinci. Penelitian sebelumnya telah mengembangkan serangkaian model U-Net dengan berbagai *encoder* seperti EfficientNet-B5, MobileNetV2, ResNet50, dan ResNeXt50 untuk segmentasi fitoplankton (Ronneberger et al., 2015; Ardhi et al., 2024a). Ragam *encoder* pada arsitektur U-Net juga dikombinasikan dengan strategi augmentasi data yang kaya seperti rotasi, *optical distortion*, *elastic transform*, dan variasi intensitas (Shorten and Khoshgoftaar, 2019; Taylor and Nitschke, 2018). Hasilnya menunjukkan bahwa U-Net dengan *encoder* MobileNetV2 dan augmentasi *optical distortion* mencapai kinerja unggul dengan Dice 93,69%, IoU 88,14%, Precision 99,89%, dan Recall 100%, sementara kombinasi *encoder* ResNet50 dengan *mixed transform* juga memberikan performa sangat tinggi dengan waktu pelatihan dan pengujian di bawah 250 detik (Ardhi et al., 2024a). Temuan ini menegaskan bahwa pemilihan *encoder* dan desain augmentasi yang tepat sangat menentukan keberhasilan segmentasi semantik, dan sekaligus memberikan dasar empiris bagi perancangan strategi augmentasi pada tugas deteksi plankton dalam disertasi ini.

2.1.2 Penanganan Variabilitas dan Ketidakpastian dalam Identifikasi Plankton

Identifikasi plankton menghadapi tantangan besar akibat variabilitas morfologi, heterogenitas kondisi lingkungan, dan keterbatasan anotasi yang akurat (Eerola et al., 2024). Perbedaan distribusi data antar wilayah, instrumen pencitraan, dan periode pengambilan sampel dapat menyebabkan *dataset shift*, sehingga model yang dilatih pada satu domain mengalami penurunan akurasi ketika diterapkan pada domain lain (Wu and He, 2025). Selain itu, ketidakseimbangan jumlah sampel antar kelas menyebabkan model cenderung bias terhadap spesies mayoritas dan kesulitan mengenali spesies langka (Li et al., 2021; MacNeil et al., 2021). Kondisi ini menuntut strategi pemodelan dan pengelolaan data yang lebih adaptif untuk menjamin stabilitas performa dan kemampuan generalisasi model dalam berbagai skenario operasional (Eerola et al., 2024).

Untuk mengatasi ketidakseimbangan kelas, salah satu pendekatan yang

digunakan adalah *oversampling*, yang mencakup teknik augmentasi data untuk menambah jumlah sampel dari kelas minoritas (Chen et al., 2024). Teknik augmentasi ini dapat mencakup rotasi citra, flipping, pemotongan, dan transformasi geometrik lainnya yang bertujuan memperkaya variasi data pelatihan. *Augmentasi data* ini membantu meningkatkan keragaman sampel sehingga model dapat belajar lebih banyak variasi visual yang muncul pada plankton yang lebih jarang. Metode ini sangat relevan untuk menangani dataset yang memiliki distribusi kelas yang tidak seimbang, yang sering dijumpai pada penelitian plankton (Buda et al., 2018).

Selain itu, *transfer learning* juga diterapkan untuk menangani masalah keterbatasan data pada kelas minoritas. Dengan memanfaatkan model yang telah dilatih pada dataset besar yang lebih umum, seperti ImageNet, dan kemudian melakukan penyesuaian untuk dataset plankton, *transfer learning* memungkinkan model untuk memulai pelatihan dengan bobot yang sudah terlatih dan kemudian menyesuaikannya pada tugas deteksi plankton. Pendekatan ini terbukti efektif dalam mempercepat pelatihan, mengurangi risiko *overfitting*, dan meningkatkan akurasi model pada dataset yang terbatas atau tidak seimbang (Wang et al., 2025; Wu and He, 2025).

Berbagai pendekatan *deep learning* telah dikembangkan untuk mengatasi permasalahan tersebut. Pendekatan *semi-supervised learning* berbasis Fuzzy Overclustering (FOC) mengombinasikan ResNet50 dengan mekanisme *overclustering* untuk mengelola label fuzzy dalam dataset plankton, dan dilaporkan mampu meningkatkan akurasi sekitar 5–10% dibandingkan metode *supervised* konvensional (Schmarje et al., 2021). Pendekatan ini relevan ketika batas antar kelas tidak tegas dan anotasi mengandung ketidakpastian taksonomi, misalnya pada spesies dengan kemiripan morfologi halus. Di sisi lain, Capsule Neural Network (CapsNet) digunakan untuk meningkatkan ketahanan model terhadap *dataset shift*, karena struktur kapsulnya mampu menyandikan hubungan bagian–keseluruhan dan variasi orientasi dengan lebih baik dibandingkan CNN biasa (Plonus et al., 2021). Penerapan CapsNet tercatat dapat mengurangi *error* prediksi hingga sekitar 12% dalam skenario pergeseran distribusi data (Plonus et al., 2021).

Pendekatan *unsupervised learning* berbasis klaster juga digunakan untuk mengakomodasi dinamika taksonomi yang terus berkembang. MorphoCluster memanfaatkan ResNet18 sebagai *feature extractor* dan HDBSCAN untuk mengelompokkan plankton berdasarkan kemiripan morfologi, sehingga jumlah kategori anotasi meningkat dari 65 menjadi 280 kelas (Schröder et al., 2020). Peningkatan jumlah klaster ini memungkinkan identifikasi kandidat spesies

baru dan memperbaiki struktur label dalam dataset besar yang semula sangat agregatif (Schröder et al., 2020). Pendekatan *retrieval-based classification* yang menggunakan SEResNeXt-50 dan *Supervised Contrastive Learning* (SCL) juga menunjukkan kinerja tinggi dengan akurasi 94,18% dan deteksi *out-of-distribution* (OOD) sebesar 86,87%, memperlihatkan bahwa representasi laten yang lebih diskriminatif dapat membantu mengenali plankton baru yang belum teranotasi (Yang et al., 2022).

Integrasi metadata lingkungan menjadi strategi tambahan untuk meningkatkan akurasi identifikasi, terutama bagi spesies langka dan kelas yang secara visual sulit dibedakan. Penggabungan metadata geotemporal, hidrograpik, dan informasi geometri dengan model VGG-16 dilaporkan mampu meningkatkan akurasi dari 87% menjadi 92,3% (Ellen et al., 2019). Metadata tersebut membantu model membedakan spesies yang secara morfologis mirip tetapi menempati *niche* lingkungan berbeda (Ellen et al., 2019). Pendekatan lain memanfaatkan CycleGAN untuk menghasilkan citra sintetis pada kelas minoritas, kemudian menggabungkannya dengan YOLOv3–DenseNet untuk menangani ketidakseimbangan kelas, dan berhasil meningkatkan mAP deteksi plankton langka sekitar 4,02% sekaligus memperbaiki distribusi prediksi antar kelas (Li et al., 2021).

Pendekatan berbasis transformer juga mulai dimanfaatkan untuk mengatasi keterbatasan anotasi dan variabilitas morfologi. β-Variational AutoEncoder (β-VAE) yang menggabungkan CNN dan transformer digunakan untuk mengompresi fitur citra plankton secara adaptif dan melakukan pembelajaran tanpa pengawasan, sehingga struktur laten data dapat dimanfaatkan meskipun label terbatas (Pastore et al., 2023). Studi lain menunjukkan bahwa *transfer learning* yang sistematis, misalnya dari ImageNet22K atau koleksi citra generik berskala besar, memberikan peningkatan akurasi beberapa persen dibanding pelatihan dari nol (Lumini and Nanni, 2019; Maracani et al., 2023; Nanni et al., 2023). Temuan-temuan tersebut memperkuat argumen bahwa penanganan variabilitas dan ketidakpastian memerlukan kombinasi strategi arsitektural, augmentasi, generasi data sintetis, dan pemanfaatan informasi tambahan yang terintegrasi (Eerola et al., 2024).

2.1.3 Peningkatan Akurasi Identifikasi Plankton

Peningkatan akurasi model merupakan salah satu fokus utama dalam pengembangan sistem identifikasi plankton berbasis *deep learning*. Akurasi yang tinggi diperlukan untuk memastikan bahwa hasil klasifikasi dan deteksi dapat digunakan sebagai dasar pengambilan keputusan ilmiah maupun kebijakan pengelolaan lingkungan laut yang andal (Luo et al., 2018; Li et al., 2021). Berbagai strategi telah diusulkan, mulai dari *ensemble learning*, *transfer learning*, optimasi arsitektur CNN dan transformer, hingga augmentasi data dan *probability filtering*. Pendekatan-pendekatan ini umumnya tidak berdiri sendiri, tetapi dikombinasikan secara cermat untuk mencapai keseimbangan antara akurasi, *robustness*, dan kompleksitas komputasi (MacNeil et al., 2021; Eerola et al., 2024).

Pendekatan *ensemble learning* memanfaatkan kelebihan berbagai arsitektur untuk meningkatkan akurasi klasifikasi. Eksperimen yang menggabungkan EfficientNetB7, DenseNet121, MobileNet, ResNet50, dan InceptionV3 pada dataset ZooLake berhasil mencapai akurasi hingga 98% dan F1-score 93%, menunjukkan bahwa kombinasi model dengan kapasitas dan cara representasi fitur berbeda memberikan keuntungan signifikan pada data yang kompleks (Kyathanahally et al., 2021). Studi lain melaporkan bahwa ansambel ResNet50, ResNet18, GoogleNet, dan MobileNetV2 dapat meningkatkan skor F1 klasifikasi foraminifera hingga 90,6%, menegaskan relevansi pendekatan ansambel di domain citra biologis (Nanni et al., 2023).

Transfer learning menjadi strategi penting untuk menghadapi keterbatasan data beranotasi pada dataset plankton. Perbandingan *in-domain transfer learning* menggunakan dataset plankton dengan *out-of-domain transfer learning* menggunakan ImageNet1K dan ImageNet22K menunjukkan bahwa *pre-training* pada ImageNet22K dapat meningkatkan akurasi hingga sekitar 6% dibandingkan skema lainnya, terutama ketika dikombinasikan dengan model transformer seperti BEiT, ViT, Swin Transformer, dan ConvNeXt (Maracani et al., 2023). Pendekatan Ensembles of Data-Efficient Vision Transformers (EDeiTs) juga dilaporkan mampu mencapai kinerja setara *state-of-the-art* pada klasifikasi ekologis, sekaligus mengurangi *overlap* prediksi antar model melalui strategi ansambel yang dirancang secara khusus (Kyathanahally et al., 2022).

Augmentasi data dan *probability filtering* terbukti efektif dalam meningkatkan akurasi sistem identifikasi plankton. Penggunaan SparseConvNet dengan *fractional max-pooling* dan serangkaian teknik augmentasi pada dataset

ISIIS mampu mencapai akurasi sekitar 90% (Luo et al., 2018). Pendekatan lain yang mengombinasikan ShuffleNet V2 dengan *probability filtering* pada citra holografi memperoleh akurasi 93,8–98%, menunjukkan bahwa penyaringan prediksi berprobabilitas rendah dapat meningkatkan keandalan sistem secara signifikan (Guo et al., 2021). Implementasi DIHM dengan *transfer learning* berbasis CNN juga dilaporkan mencapai F1-score lebih dari 89,8%, menegaskan bahwa informasi holografi dapat dimanfaatkan secara optimal tanpa rekonstruksi penuh apabila arsitektur dan strategi pelatihannya dirancang dengan tepat (MacNeil et al., 2021).

Optimasi arsitektur CNN tetap menjadi fokus penting dalam peningkatan akurasi identifikasi plankton. Penggunaan Multiple Kernel Learning (MKL) dan SVM untuk menggabungkan berbagai jenis fitur dalam klasifikasi plankton dilaporkan mampu meningkatkan *recall* pada spesies dengan morfologi kompleks (Zheng et al., 2017). Kombinasi ResNet50 dan SVM pada sistem ZOOVIS menghasilkan akurasi 94,52% dengan *recall* 94,13% pada dataset plankton *in situ*, sekaligus mengurangi kebutuhan komputasi dibanding CNN murni (Cheng et al., 2019). DeepLOKI berbasis DINO dengan pendekatan *self-supervised learning* dan ResNet18 menunjukkan potensi besar untuk mengurangi ketergantungan pada anotasi manual dalam skenario klasifikasi plankton berskala besar (Oldenburg et al., 2023).

Pendekatan segmentasi semantik berbasis U-Net menambahkan dimensi lain dalam upaya peningkatan akurasi, karena menyediakan informasi spasial yang kaya mengenai bentuk dan batas objek (Ronneberger et al., 2015; Ardhi et al., 2024a). Studi U-Net untuk segmentasi fitoplankton dengan *encoder* seperti EfficientNet-B5, MobileNetV2, ResNet50, dan ResNeXt50 serta strategi augmentasi kompleks menunjukkan bahwa kombinasi *encoder* MobileNetV2 dengan augmentasi *optical distortion* menghasilkan Dice 93,69%, IoU 88,14%, Precision 99,89%, dan Recall 100% (Ardhi et al., 2024a). Kombinasi *encoder* ResNet50 dengan *mixed transform* juga memberikan performa tinggi dengan waktu pelatihan dan inferensi yang efisien (Ardhi et al., 2024a). Temuan ini mengindikasikan bahwa informasi segmen spasial dapat dimanfaatkan untuk merancang strategi augmentasi dan desain *backbone* yang lebih peka terhadap variasi morfologi, sebagaimana menjadi salah satu pertimbangan dalam pengembangan ACViT-YOLO pada penelitian ini.

2.1.4 Pemrosesan Cepat dan Efisien dalam Identifikasi Plankton

Kecepatan pemrosesan dan efisiensi komputasi merupakan komponen penting dalam sistem identifikasi plankton, terutama ketika digunakan untuk pemantauan ekologis jangka panjang atau aplikasi *real time* yang membutuhkan respons cepat (Zhang et al., 2024). Model yang memiliki akurasi tinggi tetapi latensi besar akan sulit diterapkan pada perangkat tepi atau sistem otomatis yang memerlukan *throughput* stabil sepanjang waktu (Lang et al., 2022). Penelitian terkini menunjukkan bahwa optimasi arsitektur dapat dilakukan melalui pemilihan *backbone* yang lebih ringan, desain blok konvolusi yang efisien, serta optimasi presisi numerik untuk mengurangi beban komputasi (Cheng et al., 2019). Pendekatan ini membuka ruang untuk mencapai keseimbangan antara akurasi deteksi, penggunaan sumber daya, dan kecepatan proses inferensi dalam konteks planktonologi modern (Weihong et al., 2023).

Pada arsitektur CNN, JF-YOLO muncul sebagai varian YOLOv4 yang dirancang khusus untuk deteksi ubur-ubur dalam kondisi visual yang menantang, dan dilaporkan mampu mencapai sekitar 41 FPS dengan mAP lebih dari 92% (Zhang et al., 2024). Kinerja tersebut jauh lebih cepat dibandingkan Faster R-CNN yang hanya mencapai sekitar 8 FPS pada dataset yang sama, menegaskan bahwa modifikasi arsitektur YOLO dapat memberikan manfaat signifikan untuk deteksi organisme laut bergerak (Zhang et al., 2024). Integrasi citra holografi digital dengan YOLOv4 melalui mekanisme fusi citra juga menghasilkan mAP sebesar 97,69%, menunjukkan bahwa kualitas sinyal masukan yang lebih baik dapat mendukung percepatan proses tanpa mengorbankan akurasi (Lang et al., 2022).

Kombinasi CNN dengan metode klasik turut berkontribusi pada peningkatan efisiensi komputasi. Sistem ZOOVIS memanfaatkan ResNet50 sebagai ekstraktor fitur dan SVM sebagai pengklasifikasi akhir, menghasilkan akurasi 94,52% sekaligus menurunkan beban komputasi dibandingkan CNN penuh (Cheng et al., 2019). Teknik *adaptive thresholding* dalam pipeline ZOOVIS mempercepat penyaringan kandidat objek sehingga meningkatkan respons sistem secara keseluruhan (Cheng et al., 2019). Optimalisasi numerik juga memberikan kontribusi signifikan; pemanfaatan *half precision* pada Faster R-CNN terbukti mampu meningkatkan kecepatan inferensi sambil mempertahankan mAP di atas 93% (Weihong et al., 2023). Strategi ini memanfaatkan kemampuan GPU modern untuk menangani format presisi rendah tanpa kehilangan stabilitas prediksi (Weihong et al., 2023).

Dalam konteks plankton, efisiensi sangat relevan karena citra mikroskopik umumnya beresolusi tinggi dan memuat objek dengan variasi morfologi halus yang membutuhkan pemrosesan fitur komprehensif (MacNeil et al., 2021; Eerola et al., 2024). Sistem yang lambat akan kesulitan menangani volume data berkelanjutan dari perangkat pencitraan otomatis seperti *imaging flow cytometer* atau *in situ plankton camera*. Temuan dari berbagai penelitian tersebut memberikan pijakan teoritis kuat bahwa model yang efisien diperlukan untuk memastikan keberlanjutan pemantauan ekologis laut (Luo et al., 2018; Lang et al., 2022). Oleh karena itu, penelitian ini mengusulkan arsitektur hibrida yang secara eksplisit dirancang untuk menyeimbangkan akurasi, kompleksitas arsitektur, dan kecepatan inferensi, sehingga lebih siap untuk diadopsi dalam sistem pemantauan plankton berbasis citra mikroskopik berskala besar.

2.1.5 Deteksi Objek End-to-End dan Relevansi RT-DETR sebagai Pembanding

Arsitektur deteksi objek dalam beberapa tahun terakhir menunjukkan pergeseran signifikan dari pendekatan *anchor-based* menuju paradigma *end-to-end* yang memformulasikan deteksi sebagai prediksi himpunan terstruktur (Carion et al., 2020b). RT-DETR merupakan detektor real-time berbasis transformer yang menghilangkan kebutuhan *Non Maximum Suppression* (NMS) melalui penggunaan *Hungarian matching* untuk mengasosiasikan prediksi dan *ground truth* secara optimal (Zhao et al., 2024b). Mekanisme *hybrid encoder* memberikan efisiensi tinggi dalam pemrosesan fitur multiskala, sementara pemilihan *object query* berdasarkan tingkat ketidakpastian menghasilkan representasi yang lebih informatif tanpa menambah beban komputasi (Zhao et al., 2024b). Model ini menunjukkan keseimbangan kuat antara akurasi dan kecepatan, sehingga menjadi detektor transformer yang kompetitif dengan YOLO generasi terbaru pada dataset COCO (Zhao et al., 2024b).

Performa RT-DETR telah diuji pada berbagai domain biologis dan nonbiologis yang memiliki karakteristik objek kecil dan struktur kompleks. Pada bidang medis, penerapan RT-DETR untuk deteksi empat spesies *Plasmodium* pada preparat apusan darah menunjukkan bahwa model ini mampu mencapai akurasi tingkat pasien sekitar 79,4% sambil tetap memenuhi standar kompetensi diagnostik WHO untuk perangkat berbiaya rendah (Guemas et al., 2024). Pada bidang keamanan dan pengenalan fasilitas lalu lintas, RT-DETR yang ditingkatkan dengan

modul seperti BiFPN dan RepGFPN mampu melampaui performa YOLO-World dengan mAP mencapai 82,3% serta penurunan kompleksitas model hingga lebih dari 50% (Wan et al., 2024). Temuan tersebut menegaskan bahwa kemampuan RT-DETR dalam menangani objek kecil dan padat sangat relevan untuk domain yang menuntut presisi tinggi dan efisiensi komputasi (Guemas et al., 2024; Wan et al., 2024).

Dalam konteks penelitian ini, RT-DETR-L dipilih sebagai *baseline* pembanding karena mewakili detektor transformer efisien yang telah divalidasi pada berbagai struktur biologis halus. Pemilihan ini memungkinkan penilaian yang lebih objektif terhadap keunggulan dan kelemahan arsitektur ACViT-YOLO, khususnya dalam hal *trade-off* antara akurasi deteksi, kecepatan inferensi, dan kompleksitas komputasi (Zhao et al., 2024b). Kehadiran subbab ini memastikan bahwa pembandingan model tidak terbatas pada arsitektur CNN tradisional, tetapi juga mencakup detektor transformer generasi terbaru yang relevan secara metodologis dan empiris. Dengan demikian, bagian ini menjadi jembatan penting dari tinjauan pustaka menuju analisis metodologi pada Bab III.

2.2 Keaslian Penelitian

Keaslian penelitian ini terletak pada perancangan dan evaluasi pendekatan deteksi plankton berbasis *deep learning* yang mengintegrasikan penguatan fitur spasial-lokal dan pemodelan konteks spasial global dalam satu kerangka deteksi satu-tahap yang efisien. Integrasi tersebut dirancang untuk menjawab karakter citra plankton mikroskopik *phase contrast* yang menampilkan variasi intensitas internal, perbedaan skala objek, kemiripan morfologi antarspecies, serta potensi kemunculan multiobjek dalam satu bidang pandang. Dengan demikian, penelitian ini tidak hanya menekankan kemampuan lokalisasi objek, tetapi juga menargetkan ketahanan model terhadap variasi bentuk dan tekstur yang bersifat *fine-grained*.

Dibandingkan studi terdahulu yang dominan berfokus pada klasifikasi citra tunggal atau deteksi pada domain visual lain, penelitian ini secara eksplisit menempatkan keragaman morfologi plankton sebagai tantangan utama deteksi, sehingga kebutuhan penguatan representasi fitur menjadi lebih relevan (Luo et al., 2018; Lang et al., 2022; Zhang et al., 2024). Celah penelitian juga terlihat pada masih terbatasnya kajian komparatif yang menyeimbangkan kinerja deteksi, efisiensi komputasi, dan kompleksitas arsitektur dalam satu kerangka evaluasi yang benar-benar sebanding, terutama ketika membandingkan detektor satu-tahap lintas generasi dengan pendekatan berbasis transformer (Maracani et al., 2023;

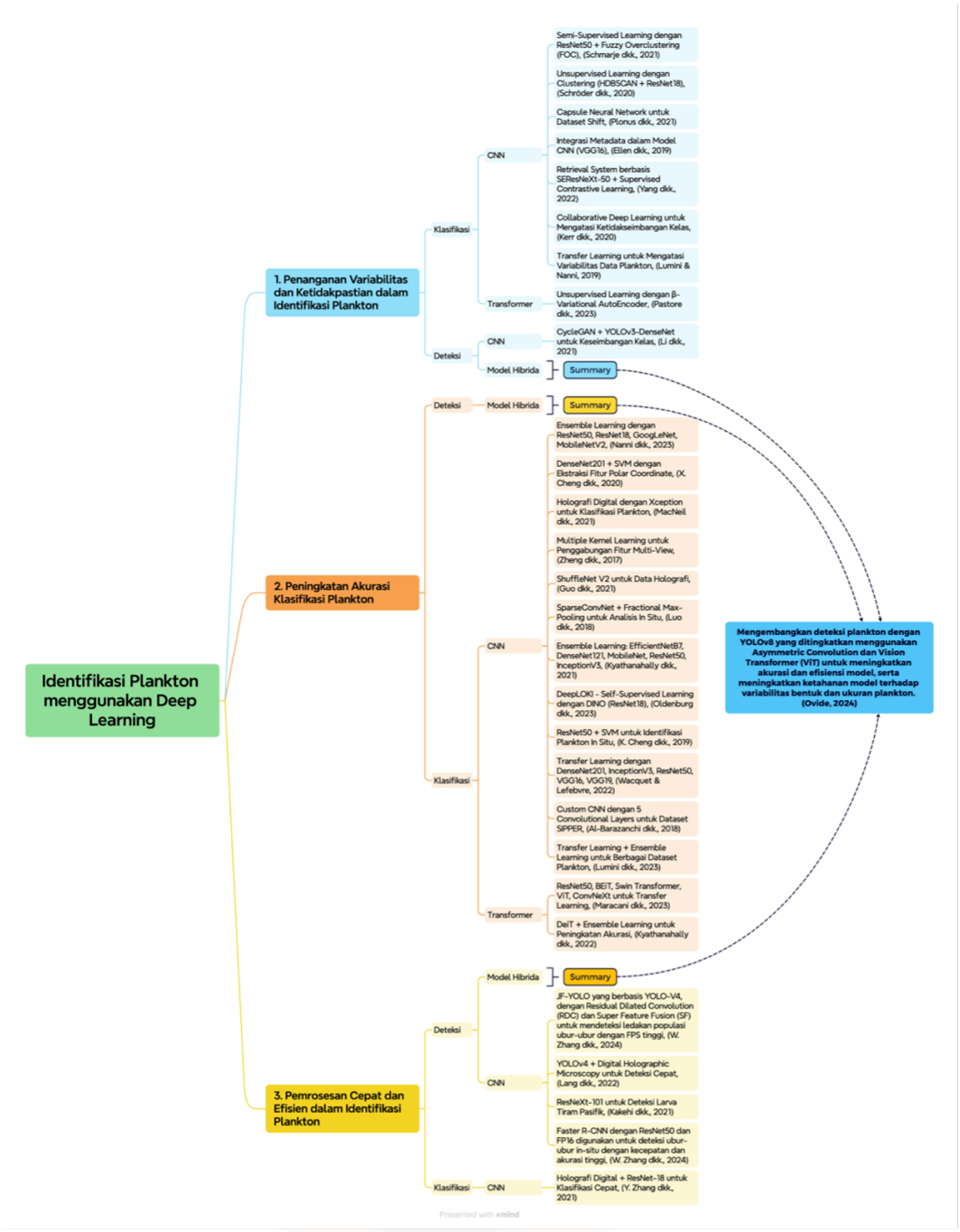
Kyathanahally et al., 2022; Zhao et al., 2024b). Relasi antara celah penelitian dan pendekatan yang diusulkan dirangkum pada Gambar 2.2 sebagai peta konseptual kontribusi yang diajukan.

Untuk memastikan kebaruan tersebut dapat dipertanggungjawabkan secara empiris, penelitian ini menerapkan evaluasi terukur menggunakan metrik kinerja deteksi (misalnya Precision, Recall, mAP, dan F1-score) serta indikator efisiensi komputasi (misalnya waktu inferensi, FLOPs, ukuran model, dan jumlah parameter). Evaluasi dilakukan melalui skema *hold-out* dan *stratified k-fold* untuk menilai stabilitas performa pada berbagai pembagian data, sekaligus menguji dampak ketidakseimbangan kelas dan strategi penyiapan data yang umum digunakan pada studi sebelumnya (Kerr et al., 2020; Li et al., 2021; Schmarje et al., 2021). Di luar aspek model inti, penelitian ini juga memosisikan pemanfaatan *Large Language Model* sebagai mekanisme pascaproses untuk memperkaya interpretasi hasil deteksi plankton dalam bentuk narasi yang lebih mudah dipahami, yang masih jarang dibahas secara eksplisit dalam konteks identifikasi plankton (Pastore et al., 2023). Ringkasan posisi penelitian ini dibandingkan studi terdahulu disajikan pada Lampiran 1.

2.3 Landasan Teori

Bab ini menyajikan landasan teori yang menjadi fondasi konseptual dalam pengembangan sistem deteksi plankton berbasis citra mikroskopik menggunakan *deep learning*. Penyusunan landasan teori diarahkan untuk membangun *Body of Knowledge* (BOK) yang koheren dan relevan dengan permasalahan penelitian, mulai dari karakter biologis plankton sebagai objek visual, sifat dan tantangan citra mikroskopik, hingga prinsip komputasional deteksi objek modern. Kerangka BOK yang merangkum keterkaitan antar konsep tersebut ditampilkan pada Gambar 2.3 sebagai penghubung logis antara dasar teoritis, perancangan metodologi, dan analisis hasil pada bab-bab berikutnya.

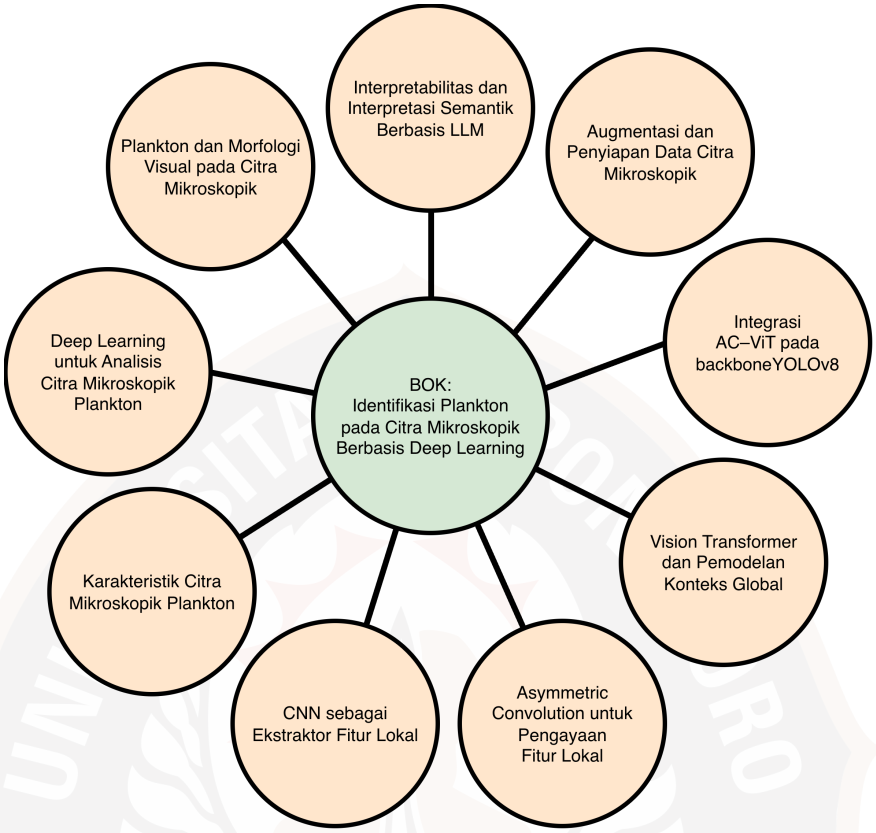
Pembahasan diawali dengan konsep dasar plankton dan identifikasi morfologi berbasis citra mikroskopik. Plankton diposisikan sebagai objek visual yang memiliki variasi bentuk, ukuran, dan struktur internal yang tinggi, serta dapat muncul sebagai lebih dari satu objek dalam satu bidang pandang. Karakteristik ini menjadikan identifikasi plankton sebagai permasalahan visual yang kompleks, sehingga membutuhkan pendekatan komputasional yang mampu menangani variasi morfologi, kemiripan antarspecies, dan kondisi multi-objek secara simultan.



Gambar 2.2 Keaslian penelitian berupa hubungan antara celah penelitian dan pendekatan yang diusulkan

Pemahaman biologis tersebut menjadi dasar dalam merumuskan kebutuhan sistem deteksi yang sesuai.

Selanjutnya, landasan teori membahas karakteristik citra mikroskopik plankton, khususnya yang dihasilkan melalui teknik *brightfield* dan *phase contrast*.



Gambar 2.3 Body of Knowledge (BOK) penelitian identifikasi plankton berbasis *deep learning*

Variasi intensitas internal, batas objek yang tidak selalu tegas, artefak optik, serta potensi tumpang tindih antarobjek dibahas sebagai faktor yang memengaruhi kualitas representasi visual. Karakteristik ini memiliki implikasi langsung terhadap proses ekstraksi fitur dan pemilihan strategi penyiapan data, sehingga pembahasan citra mikroskopik diposisikan sebagai jembatan antara aspek biologis dan pendekatan komputasional.

Landasan teori kemudian menguraikan formulasi identifikasi plankton sebagai masalah deteksi objek. Pada citra yang memuat lebih dari satu plankton dengan variasi ukuran dalam satu frame, pendekatan klasifikasi citra tunggal menjadi kurang memadai karena tidak menyediakan informasi lokalisasi. Oleh karena itu, deteksi objek dipahami sebagai formulasi yang lebih representatif, karena mampu mengidentifikasi lokasi dan kelas setiap objek secara bersamaan. Formulasi ini menjadi dasar pemilihan kerangka deteksi berbasis *deep learning* dalam penelitian.

Pada ranah komputasional, Bab 2.3 membahas prinsip dasar *deep learning*

untuk analisis citra, dengan fokus pada peran Convolutional Neural Network (CNN) sebagai ekstraktor fitur lokal dan Vision Transformer (ViT) sebagai pemodel konteks global. CNN dipahami efektif dalam menangkap pola tepi, tekstur, dan bentuk lokal, sedangkan ViT memberikan kemampuan untuk memodelkan hubungan spasial jarak jauh antarbagian citra. Pembahasan ini dilanjutkan dengan penjelasan mengenai deteksi objek berbasis YOLO sebagai kerangka efisien, serta RT-DETR sebagai pendekatan pembanding berbasis transformer, guna memberikan konteks arsitektural yang seimbang.

Sebagai puncak landasan teori, dibahas rasional integrasi CNN, *Asymmetric Convolution*, dan Vision Transformer dalam satu kerangka deteksi. Integrasi ini dijelaskan secara konseptual sebagai upaya memadukan keunggulan ekstraksi fitur lokal dan pemodelan konteks global untuk menghadapi karakter citra plankton yang bervariasi ukuran, memiliki kemiripan morfologi, dan dapat muncul sebagai multi-objek dalam satu frame. Pembahasan ini mengunci kerangka teoritis yang mendasari pengembangan arsitektur dalam penelitian, tanpa memasuki detail implementatif yang akan dijelaskan pada bab metodologi.

Bab ini ditutup dengan pembahasan teori evaluasi kinerja dan efisiensi model deteksi, serta konsep interpretabilitas dan interpretasi semantik berbasis *Large Language Model* (LLM). Metrik evaluasi diposisikan sebagai alat objektif untuk memvalidasi kualitas dan kelayakan model, sedangkan interpretabilitas dan LLM dipahami sebagai mekanisme pendukung untuk memperkaya pemahaman terhadap keluaran deteksi. Dengan alur ini, landasan teori tidak hanya merangkum konsep, tetapi juga menegaskan keterkaitan antara persoalan empiris dan pendekatan metodologis yang dipilih dalam penelitian.

2.3.1 Plankton dan Morfologi Visual pada Citra Mikroskopik

Plankton merupakan organisme akuatik yang berperan fundamental dalam ekosistem perairan dan, dalam konteks analisis citra mikroskopik, diposisikan sebagai objek visual yang dapat dikenali melalui ciri morfologi. Identifikasi plankton secara konvensional masih sangat bergantung pada pengamatan manual oleh ahli taksonomi, yang bersifat subjektif, memerlukan waktu lama, serta sulit diskalakan. Keterbatasan tersebut mendorong pemanfaatan pendekatan otomatis berbasis citra dan *deep learning* untuk meningkatkan efisiensi, konsistensi, dan objektivitas proses identifikasi plankton (Irisson et al., 2022; Eerola et al., 2024).

Secara biologis, plankton sering dikelompokkan ke dalam kategori

taksonomi seperti diatom, dinoflagellata, dan sianobakteri, yang masing-masing memiliki kecenderungan morfologi khas pada citra mikroskopik. Namun, untuk kebutuhan analisis visual dan perancangan sistem deteksi, pengelompokan berbasis ciri visual menjadi lebih relevan, khususnya ketika mempertimbangkan variasi bentuk global, skala objek, orientasi, serta kemunculan satu atau lebih objek dalam satu bidang pandang. Perspektif ini membantu menjelaskan kompleksitas visual yang dihadapi model deteksi, termasuk variasi ukuran dan kemiripan antarkelas dalam satu citra (Irisson et al., 2022; Eerola et al., 2024).

Morfologi dalam penelitian ini dipahami sebagai karakter visual yang tampak pada citra mikroskopik, mencakup bentuk kontur, karakter tepi, serta pola intensitas internal akibat proses pencitraan. Pada citra mikroskopik *phase contrast*, batas objek dapat terlihat jelas, tetapi variasi intensitas internal, artefak optik, perbedaan fokus, dan tumpang tindih antarobjek sering menimbulkan ambiguitas visual. Kondisi tersebut menjadikan identifikasi plankton sebagai permasalahan *fine-grained visual recognition*, terutama ketika perbedaan morfologi antarspecies bersifat halus (Irisson et al., 2022; Eerola et al., 2024).

Dalam praktik pemantauan perairan, identifikasi plankton juga dikaitkan dengan konteks ekologis seperti fenomena *harmful algal blooms* (HABs), di mana beberapa taksa tertentu dilaporkan berasosiasi dengan kejadian *bloom*. Pada penelitian ini, klasifikasi HAB dan non-HAB diposisikan sebagai konteks biologis pendukung interpretasi kelas, bukan sebagai penentuan toksisitas berbasis citra semata. Literatur menunjukkan bahwa taksa seperti *Dinophysis*, *Prorocentrum*, dan *Pseudo-nitzschia* kerap dibahas dalam konteks HAB, sementara banyak diatom lain merupakan komponen umum fitoplankton tanpa implikasi ekologis yang merugikan pada kondisi normal (Caballero et al., 2020; Hill et al., 2020; Rolton et al., 2022; Bu et al., 2023). Tingginya kemiripan morfologi pada sejumlah pasangan spesies, sebagaimana dirangkum pada Tabel 2.1, menegaskan bahwa tantangan utama penelitian ini adalah deteksi dan identifikasi multiobjek dalam kondisi visual yang kompleks (Irisson et al., 2022; Eerola et al., 2024).

Tabel 2.1 Kemiripan morfologi plankton

Spesies	Mirip dengan	Jenis kemiripan	Karakter pembeda
Chaetoceros curvisetus	*Skeletonema costatum*	Sama-sama rantai sel kecil pada latar berpartikel	*Chaetoceros* sedikit melengkung, *Skeletonema* lebih lurus dan rapat
Chaetoceros curvisetus	*Eucampia zodiacus*	Sama-sama rantai melengkung	*Eucampia* memiliki segmen lebih lebar dan seperti setengah lingkaran
Odontella mobiliensis	*Ornithocercus thumii*	Pinggir bersirip atau bertonjolan	*Odontella* lebih poligonal, *Ornithocercus* lebih bundar dan berjari-jari
Podolampas bipes	*Protoperidinium oceanicum*	Sama-sama piriform atau ovoid dengan isi sel padat	*Podolampas* memiliki ujung basal lebih jelas, *Protoperidinium* lebih simetris
Podolampas bipes	*Noctiluca scintillans*	Massa sel besar dengan isi di tengah	*Noctilluca* bulat, *Podolampas* bertangkai ke bawah
Prorocentrum micans	*Podolampas bipes*	Oval tebal dengan isi granular	*Prorocentrum* lebih pipih dengan tepi yang lebih jelas
Proboscia alata	*Pseudo-nitzschia*	Keduanya garis memanjang	*Proboscia* memiliki penebalan ujung, *Pseudo-nitzschia* tidak
Ceratium furca	*Trichodesmium erythraeum*	Sama-sama memanjang bila *Ceratium* terpotong	*Ceratium* memiliki tanduk, *Trichodesmium* tidak
Coscinodiscus oculus-iridis	*Asteromphalus hyalinus*; *Planktoniella sol*	Sama-sama diskus radial	*Coscinodiscus* paling padat di bagian pusat
Dinophysis caudata	*Dinophysis miles*	Badan segitiga dengan ekor	*D. caudata* memiliki ekor lebih panjang dan tegak

2.3.2 Karakteristik Citra Mikroskopik Plankton

Citra mikroskopik plankton memiliki karakteristik visual yang berbeda dari citra objek makroskopik pada umumnya karena dipengaruhi oleh teknik pencitraan, kondisi optik mikroskop, dan sifat fisik objek yang diamati. Dalam penelitian ini, citra plankton diperoleh menggunakan mikroskop dengan teknik *phase contrast* untuk meningkatkan visibilitas objek transparan tanpa pewarnaan. Teknik tersebut efektif menonjolkan kontur dan struktur visual, namun sekaligus dapat memunculkan tantangan khas seperti variasi intensitas internal, halo optik, dan ketidakstabilan batas objek yang berpengaruh terhadap proses pembelajaran model *deep learning* (Irisson et al., 2022; Eerola et al., 2024).

Salah satu karakteristik yang paling menentukan adalah variasi kontras dan tekstur internal yang tidak seragam. Pada citra *phase contrast*, perbedaan indeks bias antara objek dan medium menghasilkan pola intensitas kompleks di dalam objek, sehingga bagian internal plankton dapat tampak lebih terang atau lebih gelap dibandingkan latar. Kondisi ini menyebabkan batas objek tidak selalu tegas dan dapat berubah bergantung pada fokus, orientasi, serta pencahayaan, sehingga pelokalan objek menjadi lebih menantang dibandingkan citra dengan latar yang benar-benar homogen (Guo et al., 2021; Irisson et al., 2022).

Karakteristik berikutnya adalah keberagaman skala objek dan kemunculan lebih dari satu objek dalam satu *frame*. Plankton pada dataset penelitian berukuran beragam, mulai dari objek kecil hingga besar, dan variasi tersebut dapat muncul pada bidang pandang yang sama. Selain itu, satu citra dapat memuat multiobjek dengan jarak berdekatan atau tumpang tindih, sehingga permasalahan yang dihadapi bersifat *multi-scale* dan *multi-object* dan tidak dapat disederhanakan menjadi klasifikasi citra tunggal (Irisson et al., 2022; Eerola et al., 2024).

Citra mikroskopik plankton juga rentan terhadap artefak visual yang berasal dari sistem optik dan proses akuisisi data. Pada *phase contrast*, artefak halo, noise optik, variasi fokus, dan ketidakrataan latar dapat memiliki intensitas serta tekstur yang menyerupai bagian objek. Akibatnya, model berpotensi mengalami *false positive*, pergeseran *bounding box*, atau kebingungan antara objek dan latar, khususnya pada area tepi objek dan region dengan kontras lemah (Guo et al., 2021; Eerola et al., 2024). Ringkasan contoh karakteristik visual tersebut disajikan pada Tabel 2.2.

Selain faktor teknis, orientasi dan pose objek plankton pada citra bersifat acak dan tidak terkontrol, sehingga ciri morfologi yang sama dapat tampil pada

konfigurasi visual yang berbeda. Variasi rotasi dan sudut pandang ini memperkaya keragaman dataset, tetapi juga meningkatkan kompleksitas pembelajaran karena ciri yang relevan tidak selalu muncul dalam posisi yang konsisten (Irisson et al., 2022). Implikasi praktisnya adalah model deteksi perlu memiliki representasi fitur yang stabil terhadap variasi orientasi, skala, dan artefak, agar mampu mempertahankan kinerja pada kondisi akuisisi yang beragam.

Berdasarkan uraian tersebut, karakteristik citra mikroskopik plankton menuntut pendekatan analisis yang mampu menangani variasi kontras, skala, orientasi, serta multiobjek secara simultan. Pemahaman terhadap sifat visual citra ini menjadi landasan penting dalam merancang strategi prapemrosesan, augmentasi data, dan pemilihan arsitektur deteksi berbasis *deep learning* yang sesuai (Irisson et al., 2022; Eerola et al., 2024).

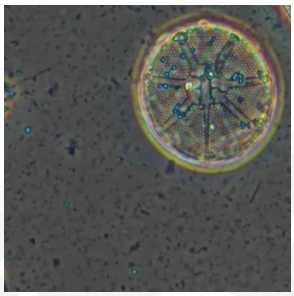
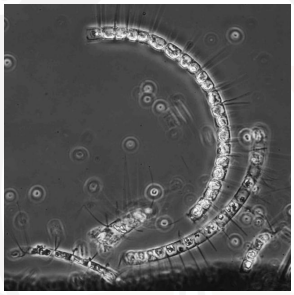
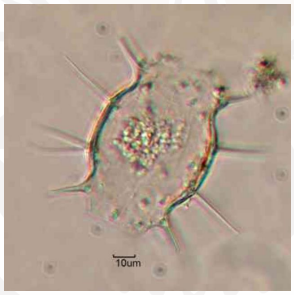
2.3.3 Formulasi Masalah Identifikasi Plankton sebagai Deteksi Objek

Identifikasi plankton berbasis citra mikroskopik dapat diformulasikan ke dalam beberapa paradigma permasalahan dalam *computer vision*, antara lain klasifikasi citra, deteksi objek, dan segmentasi. Pemilihan formulasi yang tepat menjadi aspek krusial karena secara langsung menentukan jenis informasi yang dapat diekstraksi dari citra serta relevansi hasil analisis terhadap kondisi nyata akuisisi data. Pada pendekatan klasifikasi citra, satu citra diasumsikan hanya merepresentasikan satu kelas objek utama, sehingga keluaran model terbatas pada label kelas tanpa informasi lokasi objek di dalam citra (Irisson et al., 2022).

Dalam konteks citra mikroskopik plankton, asumsi tersebut sering kali tidak terpenuhi. Satu bidang pandang dapat memuat lebih dari satu objek plankton dengan ukuran, orientasi, dan jarak antarbenda yang bervariasi, bahkan dari kelas yang berbeda. Selain itu, objek plankton dapat muncul sebagian, saling tumpang tindih, atau berada pada skala yang sangat berbeda dalam satu citra. Kondisi ini menjadikan formulasi klasifikasi citra tunggal kurang memadai karena tidak mampu merepresentasikan struktur spasial dan distribusi objek yang sebenarnya (Eerola et al., 2024; Irisson et al., 2022).

Pendekatan segmentasi, baik *semantic* maupun *instance segmentation*, secara teoritis mampu memberikan informasi spasial yang lebih rinci dengan memisahkan area objek dari latar belakang. Namun, segmentasi menuntut anotasi tingkat piksel yang jauh lebih mahal dan kompleks, terutama pada citra plankton yang memiliki batas objek kabur, tekstur internal kompleks, dan artefak optik khas

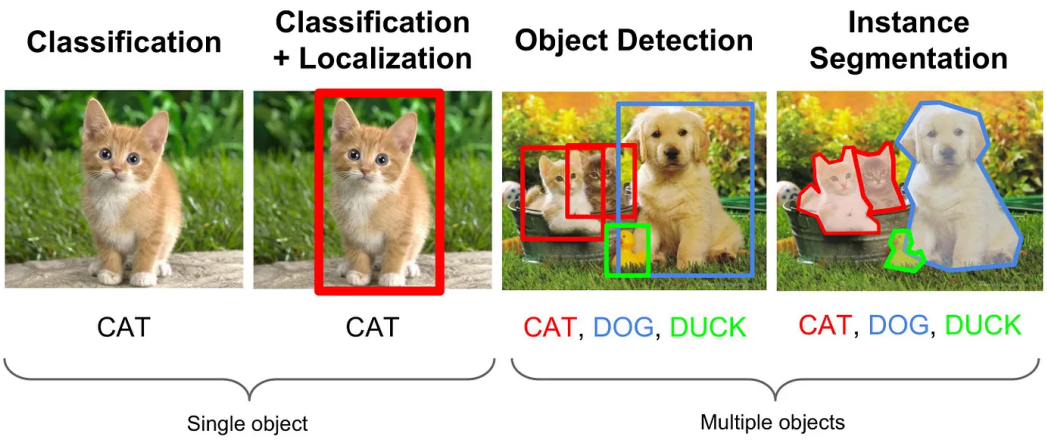
Tabel 2.2 Contoh karakteristik citra mikroskopik plankton yang relevan untuk perancangan dan evaluasi deteksi berbasis *deep learning*.

Kategori	**Contoh citra**	**Keterangan singkat**
Artefak *phase contrast* (halo/noise)		Terlihat efek halo/tepi semu di sekitar objek yang dapat mengganggu batas objek dan memicu deteksi berlebih atau pergeseran *bounding box*.
Variasi skala (kecil–besar)		Objek tampil pada ukuran relatif berbeda; menuntut deteksi multi-skala dan fitur yang stabil terhadap perubahan ukuran dan orientasi.
Variasi kontras & tekstur internal		Pola intensitas/tekstur internal tidak homogen; berpotensi menyulitkan pemisahan antarspesies yang memiliki bentuk global mirip.
Multiobjek dalam satu frame		Lebih dari satu objek muncul dalam satu bidang pandang; mempertegas formulasi sebagai deteksi objek, bukan klasifikasi citra tunggal.

phase contrast. Kebutuhan anotasi detail ini sering menjadi kendala praktis dalam skala dataset yang terbatas atau tidak seimbang (Schmarje et al., 2021; Guo et al., 2021).

Berdasarkan pertimbangan tersebut, penelitian ini memformulasikan identifikasi plankton sebagai permasalahan deteksi objek. Formulasi deteksi objek memungkinkan model untuk secara simultan menentukan lokasi (*bounding box*) dan kelas setiap plankton yang muncul dalam satu citra. Pendekatan ini lebih selaras dengan karakteristik citra mikroskopik plankton yang bersifat *multi-object*, *multi-scale*, dan heterogen, serta tetap mempertahankan kebutuhan anotasi yang relatif lebih efisien dibandingkan segmentasi (Capinha et al., 2021; Maracani et al., 2023). Perbedaan formulasi antara klasifikasi citra, deteksi objek, dan segmentasi instans ditunjukkan secara konseptual pada Gambar 2.4. Klasifikasi citra mengasumsikan keberadaan satu objek dominan dalam satu citra dan tidak menyediakan informasi lokalisasi, sehingga kurang sesuai untuk citra plankton yang sering memuat lebih dari satu objek. Segmentasi instans memberikan representasi spasial yang lebih rinci, namun menuntut anotasi yang jauh lebih kompleks dan mahal. Oleh karena itu, deteksi objek dipilih sebagai formulasi masalah yang paling seimbang untuk mengakomodasi karakter multiobjek, variasi skala, serta keterbatasan anotasi pada citra mikroskopik plankton.

Selain itu, deteksi objek menyediakan kerangka kerja yang fleksibel untuk analisis lanjutan. Informasi lokasi memungkinkan evaluasi distribusi spasial plankton, analisis kepadatan objek, serta integrasi dengan metode interpretabilitas visual seperti *Class Activation Map* (CAM). Dalam konteks penelitian ini, formulasi deteksi juga mendukung integrasi pascaproses berbasis *Large Language Models* (LLM) untuk menghasilkan deskripsi semantik berdasarkan hasil deteksi tanpa menjadikan LLM sebagai bagian dari arsitektur inti model (Pastore et al., 2023). Di luar tantangan visual akibat kemiripan morfologi antarpesies, penelitian identifikasi plankton juga dihadapkan pada konsistensi taksonomi dan penamaan ilmiah, karena perubahan klasifikasi, keberadaan sinonim, serta variasi ejaan nama spesies dapat memicu inkonsistensi label yang berdampak pada kualitas *ground truth*, validitas evaluasi, dan reprodusibilitas hasil. Oleh karena itu, basis data taksonomi terkurasi seperti AlgaeBase berperan sebagai rujukan standar untuk memastikan kesesuaian nama ilmiah dengan klasifikasi terkini, sekaligus mendukung keterlacakan ilmiah dan konsistensi pelabelan pada studi berbasis citra mikroskopik (Guiry and Guiry, 2026).



Gambar 2.4 Ilustrasi konseptual perbedaan formulasi masalah antara klasifikasi citra, deteksi objek, dan segmentasi instans.
Sumber: (Zylapp, 2021).

2.3.4 Deep Learning untuk Analisis Citra Mikroskopik Plankton

Dalam ranah *Artificial Intelligence* (AI), *machine learning* (ML) berkembang sebagai pendekatan yang memungkinkan sistem mempelajari pola dari data secara otomatis, sedangkan *deep learning* merupakan cabang khusus dari ML yang memanfaatkan jaringan saraf berlapis banyak untuk mempelajari representasi data secara hierarkis. Pada analisis citra, *deep learning* menjadi pendekatan dominan karena kemampuannya mengekstraksi fitur visual langsung dari data mentah tanpa bergantung pada perancangan fitur manual. Karakteristik ini menjadikan *deep learning* sangat relevan untuk domain bioimaging, di mana struktur visual objek sering kompleks, bervariasi, dan sulit dirumuskan secara eksplisit (LeCun et al., 2015; Panaïotis et al., 2025).

Pada citra mikroskopik plankton, pendekatan berbasis *deep learning* didorong oleh variasi bentuk, ukuran, tekstur, dan orientasi objek, serta perbedaan kualitas citra akibat kondisi optik dan proses akuisisi. Ciri pembeda antarspesies plankton sering bersifat halus dan tidak selalu konsisten antarframe, sehingga metode berbasis aturan atau fitur buatan tangan cenderung memiliki keterbatasan dalam mempertahankan generalisasi. Melalui pembelajaran representasi bertingkat, *deep learning* memungkinkan model menangkap pola visual yang relevan secara adaptif pada berbagai tingkat abstraksi, mulai dari fitur lokal hingga konteks visual yang lebih luas (Irisson et al., 2022; Eerola et al., 2024).

Dalam konteks analisis citra mikroskopik, *deep learning* dipahami sebagai kerangka komputasional umum yang menyediakan mekanisme pembelajaran

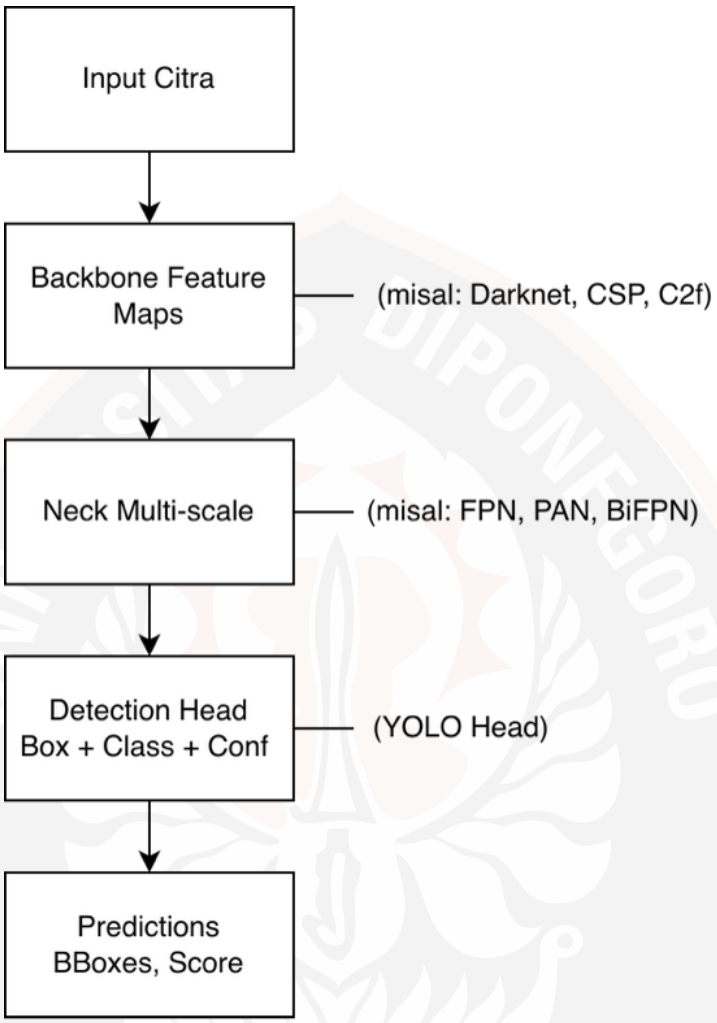
representasi visual secara otomatis dari data, sementara bentuk tugas analisis seperti klasifikasi citra, deteksi objek, atau segmentasi ditentukan berdasarkan karakteristik visual data dan kebutuhan analisis yang ingin dicapai. Pendekatan ini memungkinkan pemodelan pola visual yang kompleks dan bervariasi tanpa ketergantungan pada perancangan fitur manual, sehingga sesuai untuk citra plankton yang dicirikan oleh variasi bentuk, tekstur internal, orientasi, serta perbedaan kualitas akibat kondisi akuisisi. Perkembangan terbaru juga menunjukkan bahwa penerapan *deep learning* pada analisis plankton tidak terbatas pada citra mikroskopik konvensional, tetapi meluas ke berbagai skema pencitraan non-konvensional seperti *digital holography* dan pendekatan optik lainnya, yang menegaskan bahwa tantangan utama dalam analisis plankton terletak pada kemampuan model dalam mempelajari representasi visual yang robust terhadap variasi kondisi akuisisi dan kompleksitas objek biologis, bukan semata-mata pada jenis citra yang digunakan (Xu et al., 2025).

2.3.5 Arsitektur Deteksi Objek Modern Berbasis *Deep Learning*

Perkembangan arsitektur deteksi objek berbasis *deep learning* ditandai oleh upaya integrasi proses ekstraksi fitur, pelokalan objek, dan klasifikasi kelas ke dalam satu kerangka pemodelan yang efisien dan terukur. Secara umum, arsitektur deteksi objek dapat diklasifikasikan ke dalam tiga paradigma utama, yaitu *two-stage detectors*, *one-stage detectors*, dan arsitektur *end-to-end* berbasis transformer. Pendekatan *two-stage*, seperti Faster R-CNN, memisahkan proses proposal wilayah dan klasifikasi sehingga mampu mencapai akurasi tinggi, namun dengan konsekuensi kompleksitas komputasi dan latensi inferensi yang lebih besar. Oleh karena itu, dalam konteks penelitian ini, pembahasan difokuskan pada dua paradigma yang lebih relevan untuk pengolahan citra mikroskopik plankton berskala besar, yaitu *one-stage detector* dan *end-to-end detector*.

Pendekatan *one-stage detector* dirancang untuk menghilangkan tahapan proposal wilayah terpisah dengan memformulasikan deteksi objek sebagai regresi langsung dari citra masukan ke koordinat *bounding box* dan probabilitas kelas. Keluarga arsitektur YOLO (*You Only Look Once*) merupakan representasi paling menonjol dari paradigma ini, dengan keunggulan utama pada efisiensi inferensi dan kesederhanaan struktur (Redmon et al., 2016; Bochkovskiy et al., 2020). Secara konseptual, kerangka YOLO terdiri dari tiga komponen utama, yaitu *backbone* untuk ekstraksi fitur, *neck* untuk penggabungan fitur multiskala, dan *detection head*

untuk prediksi lokasi dan kelas objek, sebagaimana ditunjukkan pada Gambar 2.5.



Gambar 2.5 Arsitektur umum YOLO yang terdiri dari *backbone*, *neck*, dan *detection head*.

Evolusi YOLO dari versi awal hingga generasi modern menunjukkan pergeseran fokus dari sekadar kecepatan menuju keseimbangan antara akurasi, stabilitas pelatihan, dan efisiensi komputasi. Dalam konteks penelitian ini, YOLOv8 diposisikan sebagai titik acuan (*baseline*) karena merepresentasikan desain YOLO modern yang telah matang dan banyak diadopsi, terutama melalui karakter *anchor-free* dan *decoupled head* yang memisahkan jalur klasifikasi dan regresi untuk meningkatkan kualitas optimisasi. Selain itu, penggunaan blok C2f sebagai pengembangan dari pendekatan CSP menekankan efisiensi aliran gradien dan representasi fitur, yang relevan untuk objek kecil serta detail morfologi halus (Hidayatullah et al., 2025; Yaseen, 2024). Arsitektur YOLOv8 yang menjadi

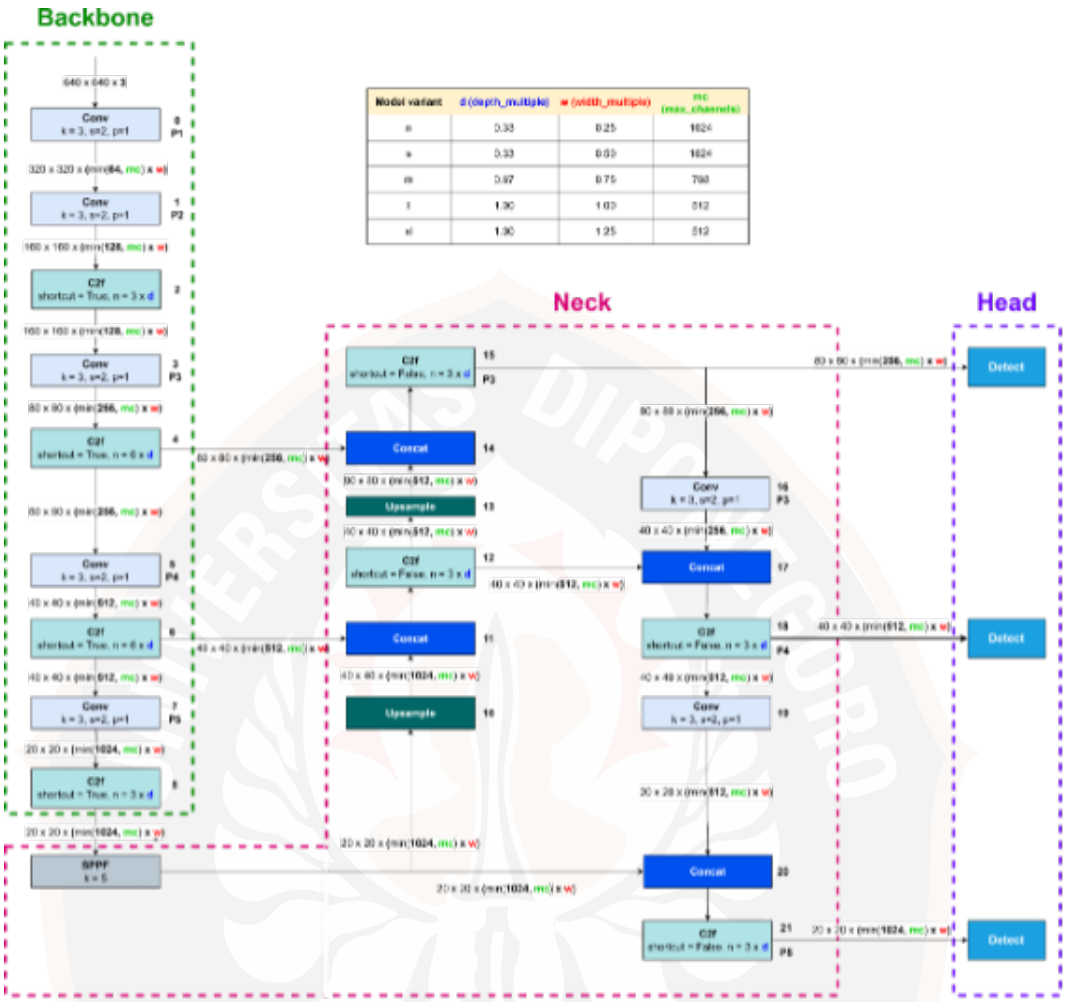
baseline dalam penelitian ini ditunjukkan pada Gambar 2.6.

Pengembangan YOLO setelah YOLOv8 (misalnya YOLOv9 hingga YOLOv12) pada prinsipnya dapat dipahami sebagai kelanjutan dari agenda yang sama, yakni memperkuat representasi multiskala, menyelaraskan pembelajaran tugas klasifikasi dan regresi, serta meningkatkan stabilitas pelatihan dan efisiensi inferensi. Oleh karena itu, pembahasan teori pada bagian ini menekankan fondasi desain YOLOv8 sebagai kerangka rujukan, sedangkan versi lanjutan diposisikan sebagai konteks evolusi arsitektur untuk menunjukkan bahwa keluarga YOLO terus berkembang dengan arah optimisasi yang konsisten pada deteksi objek modern (Wang et al., 2024b,a; Khanam and Hussain, 2024; Tian et al., 2025).

Sebagai pembanding dari paradigma *one-stage*, RT-DETR merepresentasikan pendekatan *end-to-end* berbasis transformer yang ditunjukkan pada Gambar 2.7. Berbeda dengan YOLO yang umumnya memprediksi kandidat objek melalui regresi langsung dan kemudian melakukan seleksi prediksi melalui pascaproses seperti NMS, RT-DETR mengadopsi pendekatan *set prediction* dengan penugasan satu-ke-satu berbasis *Hungarian matching*. Mekanisme tersebut mendorong prediksi yang lebih konsisten karena setiap *query* diarahkan untuk merepresentasikan satu objek, sementara desain encoder–decoder yang efisien serta strategi seleksi *query* yang informatif ditujukan untuk menjaga latensi tetap *real-time*, khususnya pada skenario objek kecil dan tumpang tindih.

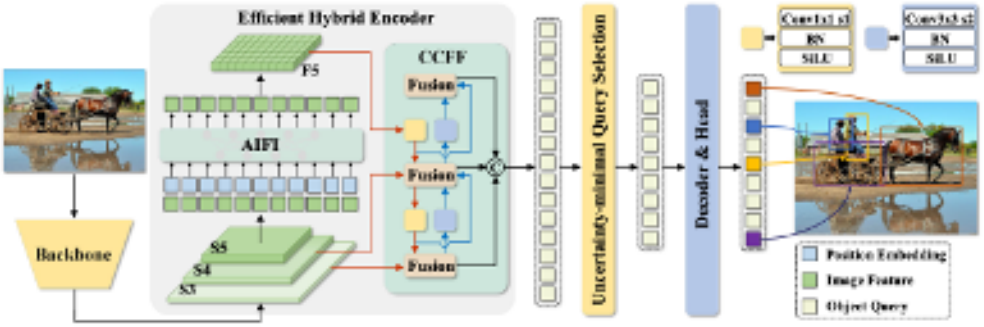
Perbandingan karakteristik antara YOLO dan RT-DETR disajikan pada Tabel 2.3, yang menegaskan perbedaan mendasar dalam strategi pemodelan, mekanisme prediksi, dan implikasi komputasi. YOLO menekankan kesederhanaan struktur dan kecepatan inferensi melalui formulasi *one-stage* yang praktis, sedangkan RT-DETR menitikberatkan konsistensi prediksi dan integrasi konteks global melalui mekanisme transformer dan *set prediction*. Kerangka perbandingan ini menjelaskan alasan pemilihan RT-DETR sebagai pembanding di luar keluarga YOLO dalam eksperimen penelitian.

Untuk melengkapi perspektif tersebut, ringkasan perbandingan struktural dan arah pengembangan keluarga YOLO dari versi YOLOv8 hingga YOLOv12 serta RT-DETR disajikan pada Tabel 2.4. Ringkasan tersebut menunjukkan bahwa evolusi YOLO modern tidak hanya bertumpu pada peningkatan ukuran jaringan atau kedalaman arsitektur, melainkan semakin menekankan penguatan representasi multiskala, peningkatan stabilitas pelatihan, serta penyelarasan komponen klasifikasi dan regresi. Inovasi seperti *task-aligned learning*, jalur fusi fitur yang lebih adaptif, dan pengayaan representasi pada level fitur dirancang



Gambar 2.6 Struktur arsitektur YOLOv8 sebagai *one-stage detector* dengan *backbone*, *neck*, dan *decoupled detection head*.
Sumber: (Hidayatullah et al., 2025).

untuk meningkatkan sensitivitas terhadap objek kecil dan detail morfologi halus, tanpa mengorbankan efisiensi inferensi yang menjadi karakter utama paradigma *one-stage*.



Gambar 2.7 Arsitektur RT-DETR sebagai detektor *end-to-end* berbasis transformer.
Sumber: (Zhao et al., 2024b).

Tabel 2.3 Perbandingan karakteristik arsitektur deteksi objek modern berbasis *deep learning*.

Aspek	**YOLO (One-stage Detector)**	**RT-DETR (End-to-end Transformer)**
Paradigma deteksi	One-stage, prediksi lokasi dan kelas secara langsung	End-to-end, berbasis *transformer decoder*
Pemodelan fitur	Berbasis CNN dengan fitur multiskala	Berbasis *self-attention* dan konteks global
Kebutuhan komponen tambahan	Umumnya menggunakan pascaproses *non-maximum suppression*; mayoritas varian modern *anchor-free*	Tidak memerlukan *anchor* dan NMS
Kemampuan konteks global	Implisit melalui fitur multiskala	Eksplisit melalui mekanisme perhatian global
Efisiensi komputasi	Sangat efisien dan cocok untuk inferensi cepat	Lebih komputasional, namun stabil secara struktural
Kesesuaian untuk citra plankton	Efektif untuk multiobjek dan variasi skala	Kuat untuk objek dengan kemiripan morfologi tinggi
Peran dalam penelitian	Kerangka deteksi utama dan pembanding berbasis CNN	Kerangka pembanding berbasis transformer

Tabel 2.4 Perbandingan struktural YOLOv8 hingga YOLOv12 dan RT-DETR

Aspek	YOLOv8	YOLOv9	YOLOv10	YOLOv11	YOLOv12	RT-DETR (v1)
Tahun rilis	2023	2024	2024	2024	2025	2023
Paradigma utama	One-stage, anchor-free	One-stage, anchor-free	One-stage, anchor-free	One-stage, anchor-free	One-stage, anchor-free	End-to-end set prediction (Transformer), tanpa NMS
Pipeline inferensi	Dense prediction + post-process (umumnya NMS)	Dense prediction + post-process	Dense prediction + post-process	Dense prediction + post-process	Dense prediction + post-process	Query-based prediction, keluaran set objek langsung
Backbone / encoder	C2f-based	C2f + PGI	C2f ringan (efisiensi)	C2f + optimisasi stabilitas (mis. aktivasi/normalisasi)	C2f + extractor ditingkatkan + attention	Backbone CNN + encoder Transformer ringan multiskala
Neck / fusi fitur	PAN–FPN	PAN–FPN + PRB	PAN–FPN optimal (latency-aware)	Dynamic FPN	Dynamic FPN v2 + scale-adaptive fusion	FPN multiskala terintegrasi ke encoder Transformer
Head / decoder	Decoupled head	Dual head + TAL	Stabilized decoupled head	Dynamic task alignment head	Unified adaptive head	Transformer decoder efisien (query terbatas)
Seleksi kandidat / query	N/A (dense)	N/A (dense)	N/A (dense)	N/A (dense)	N/A (dense)	IoU-aware query selection (memilih proposal bernilai IoU tinggi)
Label assignment / matching	Dense (anchor-free)	TAL (task-aligned)	Optimisasi assignment untuk stabilitas/efisiensi	Dynamic alignment	Adaptive coupling cls-reg	Hungarian matching satu-ke-satu (set-based)
Kebutuhan NMS	Umumnya ya	Umumnya ya	Umumnya ya	Umumnya ya	Umumnya ya	Tidak perlu NMS

Tabel 2.4 (Lanjutan) Perbandingan struktural YOLOv8 hingga YOLOv12 dan RT-DETR

Aspek	**YOLOv8**	**YOLOv9**	**YOLOv10**	**YOLOv11**	**YOLOv12**	**RT-DETR (v1)**
Inovasi ringkas	C2f, efisiensi backbone	PGI, PRB, TAL (objek kecil)	Latency-aware design	Stabilitas training dan generalisasi	Adaptive fusion + attention multiskala	Real-time DETR dengan query selection dan desain efisien
Fokus optimisasi	Keseimbangan akurasi–kecepatan	Performa objek kecil dan sulit	Efisiensi deployment	Robust pada data kecil/variatif	Representasi multiskala lebih kaya	End-to-end real-time, mengurangi overhead post-process
Kompleksitas relatif	Sedang	Tinggi	Rendah–sedang	Sedang	Tinggi	Sedang–tinggi (Transformer, tetapi efisien)
Kesesuaian citra mikroskopik	Baseline kuat	Sangat baik untuk objek kecil/halus	Baik untuk real-time ringan	Baik untuk data kecil dan variatif	Sangat baik untuk tekstur kompleks	Menarik untuk skenario multiobjek padat tanpa NMS, bergantung konfigurasi query

2.3.6 CNN sebagai Ekstraktor Fitur Lokal

Convolutional Neural Network (CNN) merupakan arsitektur *deep learning* yang banyak digunakan dalam analisis citra karena kemampuannya mempelajari representasi visual secara bertingkat langsung dari data. Pada konteks citra mikroskopik plankton, CNN berperan sebagai mekanisme utama untuk mengekstraksi ciri visual yang berkaitan dengan bentuk, tepi, dan pola internal objek. Proses ini memungkinkan citra mentah dipetakan menjadi peta fitur yang semakin abstrak dan bermakna seiring bertambahnya kedalaman jaringan (Litjens et al., 2017).

Secara konseptual, operasi konvolusi dapat dipahami sebagai proses penggabungan informasi lokal pada citra melalui fungsi pembobotan, yang secara umum dapat dituliskan sebagai

$$y(i,j) = \sum_{m} \sum_{n} x(i+m, j+n)\, w(m,n), \tag{2.1}$$

dengan x menyatakan citra masukan, w menyatakan bobot pembentuk pola, dan y adalah respons fitur yang dihasilkan. Persamaan ini menegaskan bahwa CNN secara alami sensitif terhadap pola lokal, seperti tepi dan tekstur, yang menjadi dasar pembeda visual pada citra mikroskopik.

Struktur hierarkis ekstraksi fitur pada CNN diilustrasikan pada Gambar 2.8. Lapisan awal CNN umumnya mengekstraksi fitur tingkat rendah seperti tepi, garis, dan tekstur mikro, yang merupakan komponen dasar untuk membedakan kontur objek plankton. Lapisan menengah membentuk fitur tingkat menengah berupa pola morfologi yang menggabungkan beberapa komponen lokal, seperti konfigurasi bentuk atau susunan struktur yang lebih kompleks. Lapisan akhir menghasilkan fitur tingkat tinggi atau *semantic features* yang merepresentasikan bentuk global objek, sehingga memungkinkan pembedaan kelas berdasarkan karakter morfologi secara keseluruhan (Eerola et al., 2024). Hirarki ini menunjukkan keselarasan CNN dengan prinsip identifikasi plankton berbasis morfologi visual yang telah dibahas pada subbab sebelumnya.

Meskipun efektif dalam menangkap ciri lokal, CNN memiliki sejumlah keterbatasan yang relevan pada citra mikroskopik plankton. Sensitivitas terhadap orientasi dan rotasi objek dapat menyebabkan perbedaan aktivasi fitur ketika objek muncul dalam sudut pandang yang bervariasi (Krizhevsky et al., 2017; Redmon et al., 2016). Selain itu, artefak optik khas mikroskop *phase contrast*, seperti

halo di sekitar objek, homogenitas intensitas internal, serta noise dan blur, dapat mengaburkan batas morfologi yang bersifat diagnostik. Kondisi ini meningkatkan risiko ambiguitas fitur, terutama pada kelas plankton yang memiliki kemiripan morfologi tinggi.

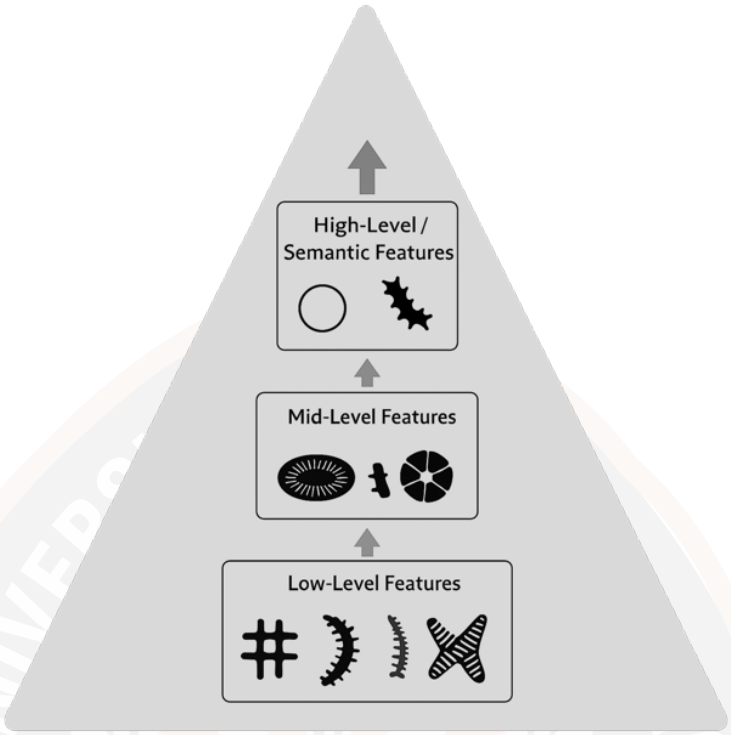
Contoh keterbatasan CNN pada citra plankton ditunjukkan pada Gambar 2.9. Panel-panel pada gambar tersebut memperlihatkan bagaimana variasi orientasi, artefak halo, homogenitas internal, serta gangguan noise dapat memengaruhi konsistensi ekstraksi fitur oleh CNN, sehingga berpotensi menurunkan stabilitas prediksi (Cheng et al., 2019; Prakasa et al., 2021).

Dalam kerangka analisis citra mikroskopik, CNN berfungsi sebagai fondasi ekstraksi fitur yang memetakan struktur visual plankton menjadi representasi numerik yang siap digunakan oleh model deteksi lanjutan. Arsitektur deteksi modern memanfaatkan peta fitur hierarkis ini untuk melakukan pelokalan objek dan penentuan kelas secara simultan (Redmon et al., 2016; Rezatofighi et al., 2019). Oleh karena itu, pemahaman terhadap kekuatan dan keterbatasan CNN menjadi landasan penting sebelum membahas pendekatan pelengkap pada subbab berikutnya, yang dirancang untuk memperkaya representasi lokal dan memperluas pemodelan konteks visual.

2.3.7 Asymmetric Convolution sebagai Penguatan Ekstraksi Fitur CNN

Konvolusi konvensional pada jaringan saraf konvolusional umumnya menggunakan kernel simetris seperti 3×3 untuk mengekstraksi fitur lokal. Pada citra mikroskopik organisme halus, struktur visual sering bersifat anisotropik, misalnya bentuk memanjang, berduri, bercabang, atau berorientasi tertentu, sehingga pola tepi dan kontur dapat lebih dominan pada arah horizontal atau vertikal. Kernel simetris tetap mampu menangkap pola tersebut, namun responsnya tidak selalu optimal ketika arah fitur sangat kuat dan tidak seragam. Untuk menguatkan sensitivitas arah tanpa mengubah prinsip dasar konvolusi, *Asymmetric Convolution* (AC) memperkenalkan kernel non-persegi seperti $1 \times N$ dan $N \times 1$ sebagai komponen pengaya fitur lokal (Ding et al., 2019). Intinya, jaringan diberikan kemampuan untuk merespons pola garis dan kontur pada dua orientasi utama secara lebih eksplisit, sebelum representasi tersebut digabungkan sebagai fitur spasial-lokal yang lebih kaya, sebagaimana diilustrasikan pada Gambar 2.10.

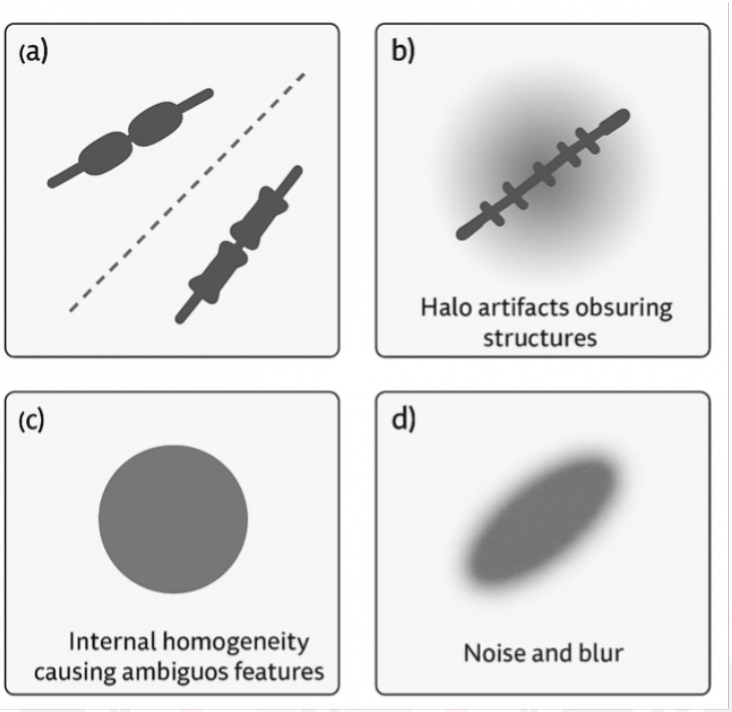
Pada implementasi AC yang umum, kernel asimetris tidak berdiri sendiri, tetapi digunakan sebagai cabang tambahan yang melengkapi kernel simetris.



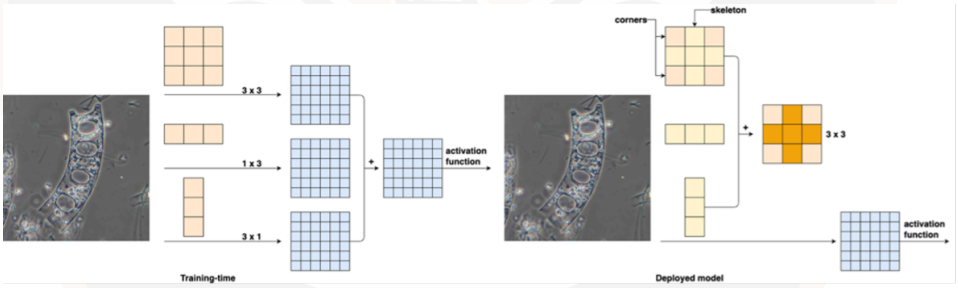
Gambar 2.8 Hirarki peta fitur pada *Convolutional Neural Network* (CNN), mulai dari fitur tingkat rendah hingga representasi semantik tingkat tinggi.

Selama pelatihan, beberapa cabang konvolusi dapat dijalankan paralel, misalnya cabang $N \times N$, $1 \times N$, dan $N \times 1$, kemudian hasilnya dijumlahkan menjadi satu peta fitur. Skema multi-cabang ini sejalan dengan ilustrasi pada Gambar 2.10, yakni pengayaan respon arah dilakukan melalui pemisahan kernel menjadi komponen horizontal dan vertikal. Keuntungan praktisnya adalah jaringan memperoleh respon arah yang lebih tajam, sedangkan biaya inferensi dapat tetap efisien jika parameter cabang-cabang tersebut direparametrisasi menjadi satu kernel ekuivalen pada saat inferensi (Ding et al., 2019). Dengan demikian, pembahasan efisiensi AC lebih tepat dipahami sebagai *pengayaan representasi arah dengan overhead yang terkendali*, bukan sekadar klaim penurunan kompleksitas secara otomatis pada semua konfigurasi.

Secara formal, misalkan $I \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ adalah citra atau peta fitur masukan, dan $Y \in \mathbb{R}^{H' \times W' \times C_{\text{out}}}$ adalah keluaran konvolusi. Konvolusi 2D standar dengan



Gambar 2.9 Contoh keterbatasan CNN pada citra mikroskopik plankton: (a) sensitivitas terhadap orientasi/rotasi, (b) artefak halo pada *phase contrast*, (c) homogenitas intensitas internal, dan (d) pengaruh noise atau blur.



Gambar 2.10 Konsep *asymmetric convolution* dengan kernel $1 \times N$ dan $N \times 1$ sebagai pengaya respon arah.
Sumber: (Ding et al., 2019).

stride s dan padding p dapat dituliskan sebagai:

$$Y(x, y, c_{\text{out}}) = \sum_{c_{\text{in}}=1}^{C_{\text{in}}} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} K_{N\times N}(i, j, c_{\text{in}}, c_{\text{out}})\, I(x \cdot s + i - p,\ y \cdot s + j - p,\ c_{\text{in}}), \tag{2.2}$$

dengan keterangan: $x \in \{0, \ldots, H' - 1\}$ dan $y \in \{0, \ldots, W' - 1\}$ adalah indeks spasial keluaran, c_{in} dan c_{out} adalah indeks kanal masukan dan keluaran, N adalah ukuran kernel, $K_{N\times N}$ adalah bobot kernel simetris, s adalah stride, dan p

adalah padding (diasumsikan sama untuk tinggi dan lebar agar ringkas). Notasi $I(\cdot)$ mengikuti konvensi *zero padding* ketika indeks keluar dari batas citra.

Pada *asymmetric convolution*, kernel non-persegi digunakan untuk menguatkan respon arah. Untuk cabang horizontal $1 \times N$ dan cabang vertikal $N \times 1$, keluaran masing-masing cabang dapat dituliskan sebagai:

$$Y_{1\times N}(x, y, c_{\text{out}}) = \sum_{c_{\text{in}}=1}^{C_{\text{in}}} \sum_{j=0}^{N-1} K_{1\times N}(0, j, c_{\text{in}}, c_{\text{out}})\, I(x \cdot s - p,\ y \cdot s + j - p,\ c_{\text{in}}),$$
$$Y_{N\times 1}(x, y, c_{\text{out}}) = \sum_{c_{\text{in}}=1}^{C_{\text{in}}} \sum_{i=0}^{N-1} K_{N\times 1}(i, 0, c_{\text{in}}, c_{\text{out}})\, I(x \cdot s + i - p,\ y \cdot s - p,\ c_{\text{in}}). \tag{2.3}$$

dengan keterangan: $K_{1\times N}$ adalah kernel asimetris horizontal dan $K_{N\times 1}$ adalah kernel asimetris vertikal, sedangkan simbol lain sama dengan Persamaan (2.2). Intuisi dari dua cabang ini konsisten dengan ilustrasi pada Gambar 2.10, yakni masing-masing cabang menonjolkan respon pada orientasi tertentu untuk menangkap pola garis, kontur, atau struktur anisotropik yang dominan.

Jika AC diterapkan sebagai pengaya berbasis multi-cabang, maka peta fitur keluaran dapat digabungkan, misalnya dengan penjumlahan linear dan bias:

$$Y_{\text{AC}}(x, y, c_{\text{out}}) = Y_{N\times N}(x, y, c_{\text{out}}) + Y_{1\times N}(x, y, c_{\text{out}}) + Y_{N\times 1}(x, y, c_{\text{out}}) + b(c_{\text{out}}), \tag{2.4}$$

dengan keterangan: $Y_{N\times N}$ adalah keluaran cabang kernel simetris (bila digunakan), b adalah bias per kanal keluaran. Skema penggabungan ini membuat jaringan memperoleh representasi fitur lokal yang lebih kaya dan lebih peka terhadap orientasi, yang relevan untuk objek mikroskopik dengan bentuk anisotropik. Pada tahap desain model, implikasi AC dapat dibahas sebagai trade-off antara peningkatan sensitivitas fitur lokal dan overhead parameter atau komputasi yang dapat ditekan melalui strategi implementasi (misalnya pemilihan posisi blok, ukuran kernel, serta reparametrisasi pada fase inferensi) (Ding et al., 2019).

2.3.8 Vision Transformer dan Representasi Konteks Global

Vision Transformer (ViT) diperkenalkan sebagai pendekatan pemrosesan citra berbasis arsitektur Transformer yang sebelumnya sukses pada pemodelan urutan, terutama karena kemampuannya menangkap dependensi jarak jauh melalui mekanisme *self-attention* (Vaswani et al., 2017; Dosovitskiy et al., 2021). Pada identifikasi fitoplankton, objek dapat muncul pada posisi acak,

berorientasi beragam, serta berpotensi saling tumpang tindih, sementara perbedaan antarspesies sering bersifat *fine-grained* dan tidak selalu ditentukan oleh tepi lokal saja. Model CNN memang efektif mengekstraksi fitur spasial-lokal, namun pemahaman relasi antarbagian citra dalam skala luas umumnya bergantung pada pertambahan *receptive field* secara bertahap seiring kedalaman jaringan. ViT menawarkan mekanisme yang lebih langsung, yakni menilai hubungan antarpotongan citra (*patch*) secara global sehingga bagian yang berjauhan tetap dapat saling memengaruhi pembentukan representasi. Dengan demikian, peran ViT pada konteks ini dapat dipahami sebagai penguatan *konteks global/konten*, yaitu pemodelan relasi bagian-ke-bagian dan koherensi bentuk keseluruhan yang membantu membedakan spesies dengan kemiripan morfologi tinggi.

Arsitektur dasar ViT ditunjukkan pada Gambar 2.11. Citra masukan $X \in \mathbb{R}^{H\times W\times C}$ dibagi menjadi *patch* berukuran tetap $P \times P$, sehingga jumlah *patch* menjadi $N = \frac{HW}{P^2}$. Setiap *patch* diratakan (*flatten*) lalu diproyeksikan secara linear menjadi *patch embedding* berdimensi D, menghasilkan urutan token $E \in \mathbb{R}^{N\times D}$. Agar informasi urutan spasial tidak hilang, *positional embedding* $E_{\text{pos}} \in \mathbb{R}^{N\times D}$ ditambahkan sehingga token yang diproses encoder menjadi $Z_0 = E + E_{\text{pos}}$. Pada tahap klasifikasi, sebuah token khusus [CLS] ditambahkan di awal urutan sehingga panjang urutan menjadi $L = N + 1$ dan representasi masuk encoder dinyatakan sebagai $Z_0 \in \mathbb{R}^{L\times D}$. Susunan ini membuat setiap token mewakili bagian citra (patch) beserta posisinya, sehingga relasi antarbagian citra dapat dipelajari secara eksplisit oleh encoder.

Gambar 2.12 memperlihatkan arsitektur Transformer secara umum yang terdiri dari encoder dan decoder. Dalam ViT untuk klasifikasi citra, komponen yang digunakan secara utama adalah *Transformer Encoder* untuk membangun representasi global dari seluruh token (Dosovitskiy et al., 2021). Encoder tersusun atas beberapa blok yang memuat *Multi-Head Self-Attention* (MHSA), normalisasi, dan *feed-forward network* (MLP). Inti dari MHSA adalah menghitung bobot perhatian antar token sehingga setiap token dapat mengagregasi informasi dari token lain, termasuk token yang jauh secara spasial. Mekanisme ini selaras dengan kebutuhan identifikasi plankton, karena beberapa ciri morfologi penting dapat tersebar pada bagian citra yang berbeda dan tidak selalu dominan pada fitur lokal tunggal.

Mekanisme *self-attention* dihitung menggunakan matriks *query*, *key*, dan *value* sebagaimana Persamaan (2.5). Misalkan keluaran sebuah blok sebelum perhatian adalah $Z \in \mathbb{R}^{L\times D}$. Proyeksi linear membentuk $Q = ZW^Q$, $K = ZW^K$,

dan $V = ZW^V$, dengan $Q, K \in \mathbb{R}^{L\times d_k}$ dan $V \in \mathbb{R}^{L\times d_v}$, di mana $W^Q \in \mathbb{R}^{D\times d_k}$, $W^K \in \mathbb{R}^{D\times d_k}$, dan $W^V \in \mathbb{R}^{D\times d_v}$ adalah parameter yang dipelajari. Matriks $QK^\top \in \mathbb{R}^{L\times L}$ membentuk skor afinitas antar token; pembagian dengan $\sqrt{d_k}$ menjaga skala numerik stabil; dan fungsi *softmax* diterapkan per baris agar bobot perhatian untuk setiap token membentuk distribusi probabilitas (Vaswani et al., 2017). Dengan cara ini, satu token dapat "memperhatikan" token lain yang relevan sehingga relasi global dapat terbangun tanpa bergantung pada perluasan *receptive field* konvolusi.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{2.5}$$

Untuk meningkatkan kapasitas representasi, perhatian dihitung pada beberapa subruang melalui *multi-head*. Jika jumlah *head* adalah h, maka pada *head* ke-i digunakan proyeksi W_i^Q, W_i^K, dan W_i^V, sehingga $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. Hasil semua *head* kemudian digabungkan dengan konkatenasi dan diproyeksikan kembali melalui W^O sebagaimana Persamaan (2.6). Secara dimensi, jika setiap *head* menghasilkan keluaran berdimensi d_v, maka operasi $\text{Concat}(\cdot)$ menghasilkan representasi $\mathbb{R}^{L\times(h\cdot d_v)}$ yang kemudian diproyeksikan oleh $W^O \in \mathbb{R}^{(h\cdot d_v)\times D}$ agar kembali ke dimensi model D. Mekanisme ini memungkinkan model mempelajari beberapa pola relasi secara paralel, misalnya relasi kontur, tekstur internal, atau konfigurasi bentuk global yang menjadi pembeda penting pada kasus morfologi yang mirip.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q,\, KW_i^K,\, VW_i^V), \end{aligned} \tag{2.6}$$

Setelah MHSA, encoder menggunakan *feed-forward network* (MLP) untuk memperkaya non-linearitas representasi, seperti pada Persamaan (2.7). Secara umum, x merepresentasikan vektor fitur token pada satu posisi, W_1 dan W_2 adalah bobot MLP, serta b_1 dan b_2 adalah bias. Fungsi aktivasi non-linear $\sigma(\cdot)$ lazim menggunakan GELU atau ReLU, yang membantu membentuk pemetaan non-linear pada ruang fitur sehingga representasi menjadi lebih diskriminatif (Dosovitskiy et al., 2021).

$$\text{MLP}(x) = \sigma(xW_1 + b_1)\, W_2 + b_2, \tag{2.7}$$

Pada tahap klasifikasi, token [CLS] berfungsi sebagai agregator representasi global seluruh token setelah melalui beberapa lapisan encoder. Jika $Z_{\text{CLS}} \in \mathbb{R}^D$ menyatakan vektor keluaran token [CLS], maka prediksi kelas diperoleh melalui lapisan linear dan *softmax* sebagaimana Persamaan (2.8). Parameter $W_c \in \mathbb{R}^{D\times K}$ dan $b_c \in \mathbb{R}^K$ adalah bobot dan bias klasifier, dengan K jumlah kelas.

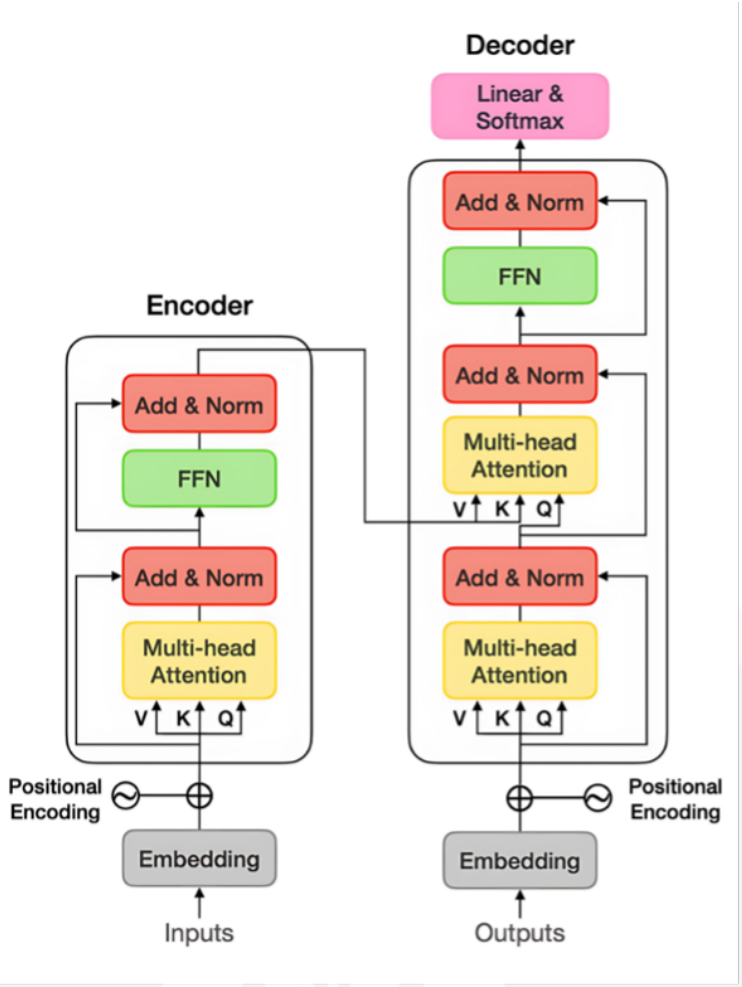
$$\hat{y} = \text{softmax}\left(Z_{\text{CLS}} W_c + b_c\right), \tag{2.8}$$

Dibandingkan CNN, ViT memiliki keunggulan dalam menangkap dependensi jarak jauh karena setiap token dapat berinteraksi langsung dengan token lain melalui matriks perhatian. Keunggulan ini penting untuk membedakan spesies plankton yang secara lokal tampak serupa namun berbeda pada konfigurasi bentuk global atau relasi bagian internalnya. Di sisi lain, komputasi perhatian bergantung pada panjang urutan token L karena melibatkan matriks $L \times L$, sehingga strategi seperti pemilihan ukuran *patch*, dimensi embedding, dan jumlah *head* perlu dipertimbangkan agar trade-off akurasi dan efisiensi tetap proporsional. Walaupun ViT sering diasosiasikan dengan kebutuhan data besar, pendekatan *pre-training* dan *transfer learning* memungkinkan adaptasi pada dataset berukuran terbatas, termasuk pada citra mikroskopik. Dengan karakter tersebut, ViT menjadi fondasi yang relevan untuk memperkuat pemodelan konteks global sebagai pelengkap penguatan fitur spasial-lokal, yang selanjutnya dirancang terintegrasi dalam arsitektur hibrida ACViT–YOLO yang diusulkan.

2.3.9 Rasional Integrasi CNN–AC–ViT pada Deteksi Plankton

Citra mikroskopik plankton memiliki karakter yang tidak sederhana. Objek plankton sering kali memiliki perbedaan bentuk yang sangat halus antarspesies, muncul dalam berbagai ukuran dalam satu citra, serta dipengaruhi oleh kondisi pencahayaan dan artefak optik. Dalam situasi seperti ini, pendekatan deteksi yang hanya mengandalkan satu jenis mekanisme representasi visual berisiko tidak mampu menangkap seluruh informasi penting yang dibutuhkan. Oleh karena itu, integrasi beberapa pendekatan pembelajaran visual dipandang sebagai langkah konseptual yang masuk akal untuk membangun sistem deteksi plankton yang lebih seimbang.

Dalam arsitektur YOLO, ekstraksi fitur visual pada dasarnya dilakukan menggunakan *Convolutional Neural Network* (CNN). CNN bekerja dengan mengekstraksi pola-pola lokal dari citra, seperti tepi, tekstur, dan bentuk dasar objek, melalui proses konvolusi bertingkat yang dapat dinyatakan secara umum pada



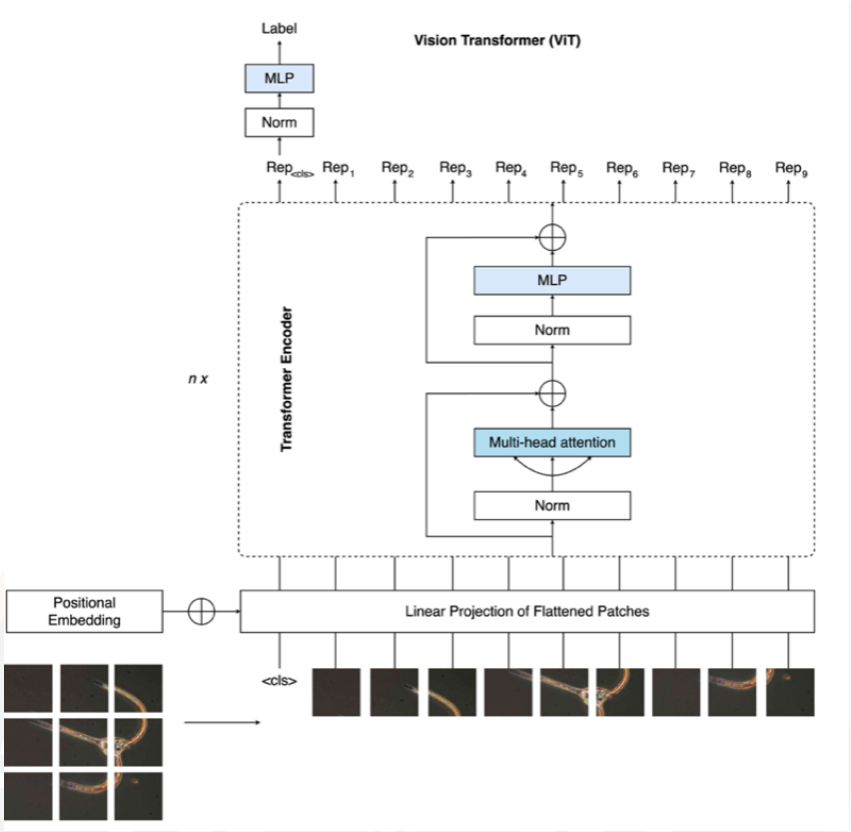
Gambar 2.11 Detail arsitektur Transformer yang terdiri dari encoder dan decoder.
Sumber: (Ruan et al., 2022).

Persamaan (2.9).

$$\mathbf{F}_{l+1} = \text{Conv}(\mathbf{F}_l). \tag{2.9}$$

Pendekatan ini efektif untuk mengenali ciri-ciri visual tingkat rendah hingga menengah, yang sangat penting dalam pengenalan morfologi plankton. Namun, karena CNN memproses citra dalam area lokal yang terbatas dan dengan pola yang relatif seragam ke segala arah, informasi tentang arah bentuk tertentu atau susunan global objek sering kali belum tertangkap secara optimal.

Untuk memperkuat kemampuan model dalam menangkap detail lokal yang memiliki arah tertentu, *Asymmetric Convolution* (AC) diperkenalkan sebagai pengayaan terhadap konvolusi standar. Berbeda dengan konvolusi konvensional yang menggunakan kernel simetris, AC memproses arah horizontal dan vertikal secara terpisah. Secara sederhana, perbedaan ini dapat dipahami melalui formulasi



Gambar 2.12 Arsitektur dasar Vision Transformer (ViT) dengan pembentukan *patch embedding*, *positional embedding*, dan *Transformer Encoder*.
dimodifikasi dari (Zhang et al., 2023).

pada Persamaan (2.10).

$$\mathcal{B}_{\text{CNN}}(\mathbf{X}) \approx \text{Conv}_{3\times3}(\mathbf{X}), \qquad \mathcal{B}_{\text{AC}}(\mathbf{X}) = \text{Conv}_{3\times1}\big(\text{Conv}_{1\times3}(\mathbf{X})\big). \tag{2.10}$$

Pendekatan ini membuat model lebih peka terhadap bentuk memanjang, berlekuk, atau tidak simetris, yang sering dijumpai pada plankton seperti rantai diatom atau struktur bercabang. Dalam kerangka YOLO, gagasan ini diwujudkan dengan mengganti blok `C2f` standar menjadi `C2fx` pada tahap awal backbone, sehingga pengayaan bentuk lokal dilakukan sejak fitur masih memiliki resolusi tinggi.

Di sisi lain, citra plankton tidak hanya menuntut ketelitian pada detail lokal, tetapi juga pemahaman hubungan antarbagian objek secara keseluruhan, bahkan antarobjek dalam satu citra. Untuk kebutuhan ini, *Vision Transformer* (ViT) menawarkan mekanisme yang memungkinkan model melihat citra secara lebih menyeluruh. Melalui interaksi antarbagian fitur, ViT membantu model memahami susunan global dan konteks spasial yang lebih luas. Dalam arsitektur YOLO,

peran ini direalisasikan melalui blok `C3TR` yang ditempatkan pada tahap akhir backbone, ketika fitur telah cukup abstrak untuk merepresentasikan makna bentuk secara global.

Penting untuk ditekankan bahwa AC dan ViT memiliki tujuan yang berbeda. AC berfokus pada penguatan detail lokal yang memiliki arah tertentu, sedangkan ViT berfokus pada penggabungan informasi global dari seluruh bagian citra. Jika masing-masing pendekatan ini digunakan secara terpisah, maka model secara konseptual akan lebih menekankan satu aspek representasi tertentu, baik lokal maupun global. Pada data seperti plankton, yang membutuhkan keduanya secara bersamaan, pendekatan terpisah tersebut berpotensi menyisakan keterbatasan pada aspek yang tidak diperkuat.

Atas dasar pemikiran tersebut, penelitian ini memposisikan integrasi AC dan ViT bukan sebagai upaya perbaikan parsial satu per satu, melainkan sebagai strategi penggabungan dua mekanisme yang saling melengkapi. AC ditempatkan pada tahap awal backbone untuk memperkaya representasi bentuk lokal sejak awal, sementara ViT ditempatkan pada tahap akhir untuk memperkuat pemahaman konteks global. Perbedaan konfigurasi backbone antar model dalam penelitian ini dapat diringkas sebagai berikut:

$$\textbf{YOLOv8: } [\text{C2f}]_{\text{awal}} + [\text{C2f}]_{\text{akhir}}, \tag{2.11}$$

$$\textbf{AC-YOLO: } [\text{C2fx}]_{\text{awal}} + [\text{C2f}]_{\text{akhir}}, \tag{2.12}$$

$$\textbf{ViT-YOLO: } [\text{C2f}]_{\text{awal}} + [\text{C3TR}]_{\text{akhir}}, \tag{2.13}$$

$$\textbf{ACViT-YOLO: } [\text{C2fx}]_{\text{awal}} + [\text{C3TR}]_{\text{akhir}}. \tag{2.14}$$

Konfigurasi ini menunjukkan bahwa AC-YOLO dan ViT-YOLO masing-masing menekankan satu jenis pengayaan fitur, sedangkan ACViT-YOLO dirancang untuk menggabungkan penguatan detail lokal dan pemahaman konteks global dalam satu kerangka yang utuh. Dengan demikian, ACViT-YOLO diposisikan sebagai pendekatan yang secara konseptual lebih sesuai untuk karakter citra plankton yang bersifat multi-skala, kaya variasi bentuk, dan sering memuat lebih dari satu objek dalam satu citra.

Untuk memperjelas perbedaan rancangan backbone pada masing-masing model, Tabel 2.5 merangkum konfigurasi utama backbone berdasarkan pengaturan berkas *configuration* (cfg) YAML. Perbedaan difokuskan pada jenis blok yang digunakan pada tahap awal dan tahap akhir backbone, karena kedua tahap tersebut merepresentasikan pengayaan fitur lokal dan pemodelan konteks global.

Tabel 2.5 Perbandingan konfigurasi backbone pada empat model YOLO berdasarkan berkas YAML.

Model	**Tahap Awal Backbone**	**Tahap Akhir Backbone**	**Fokus Representasi Utama**
YOLOv8	C2f (konvolusi standar berbasis CNN)	C2f (konvolusi standar berbasis CNN)	Ekstraksi fitur lokal umum
AC-YOLO	C2fx (konvolusi asimetris)	C2f (konvolusi standar)	Penguatan detail lokal berarah
ViT-YOLO	C2f (konvolusi standar)	C3TR (Vision Transformer)	Pemodelan konteks global
ACViT-YOLO	C2fx (konvolusi asimetris)	C3TR (Vision Transformer)	Detail lokal dan konteks global

2.3.10 Augmentasi dan Penyiapan Data Citra Mikroskopik

Penyiapan data merupakan tahapan kunci dalam sistem deteksi citra mikroskopik karena kualitas dan konsistensi citra secara langsung memengaruhi stabilitas pembelajaran model *deep learning*. Pada citra plankton, variasi iluminasi, perbedaan intensitas antar sesi akuisisi, serta variasi orientasi objek tidak dapat sepenuhnya dihindari. Variasi tersebut dapat membuat model belajar pada artefak pencahayaan atau perbedaan kontras, alih-alih pada ciri morfologi yang relevan. Oleh karena itu, prapemrosesan dan augmentasi diperlukan untuk menormalkan karakteristik dasar citra sekaligus menambah keragaman sampel pelatihan secara terkontrol agar representasi yang dipelajari lebih stabil dan representatif.

Tahap prapemrosesan diawali dengan normalisasi intensitas piksel untuk mengurangi pengaruh perbedaan pencahayaan dan rentang nilai intensitas antar citra. Normalisasi skala linier dilakukan dengan memetakan intensitas I ke rentang $[0, 1]$ sebagaimana dirumuskan pada Persamaan (2.15).

$$I' = \frac{I - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}} \tag{2.15}$$

dengan keterangan: I adalah nilai intensitas piksel (atau nilai kanal pada peta fitur) sebelum normalisasi, I' adalah intensitas setelah normalisasi, I_{min} dan I_{max} berturut-turut adalah nilai minimum dan maksimum intensitas pada citra yang sama (atau pada kanal yang sama). Notasi ini dapat dipahami berlaku per piksel $I(x, y)$

pada koordinat spasial (x, y), sehingga $I'(x, y)$ adalah hasil normalisasi pada lokasi yang sama. Pemetaan ini menjaga urutan relatif intensitas dan menyetarakan skala masukan agar kompatibel dengan proses optimisasi.

Untuk kebutuhan tertentu, transformasi ke rentang simetris $[-1, 1]$ digunakan agar distribusi nilai piksel lebih seimbang di sekitar nol, sebagaimana ditunjukkan pada Persamaan (2.16).

$$I' = 2 \times \left(\frac{I - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}} \right) - 1 \tag{2.16}$$

dengan keterangan: simbol I, I_{min}, dan I_{max} sama seperti pada Persamaan (2.15), sedangkan I' menyatakan intensitas yang telah dipetakan ke rentang simetris. Transformasi ini sering berguna ketika model atau fungsi aktivasi lebih stabil jika masukan berpusat di sekitar nol.

Selain normalisasi, standardisasi berbasis rerata μ dan simpangan baku σ diterapkan untuk menekan variasi intensitas yang muncul akibat perbedaan pengaturan mikroskop, karakteristik sensor kamera, atau kondisi lingkungan saat pengambilan citra. Standardisasi ini dirumuskan pada Persamaan (2.17) dan membantu model memfokuskan pembelajaran pada pola bentuk dan tekstur, bukan pada perbedaan iluminasi semata.

$$I_{\text{norm}} = \frac{I - \mu}{\sigma} \tag{2.17}$$

dengan keterangan: I adalah intensitas piksel sebelum standardisasi (dapat dipandang sebagai $I(x, y)$), I_{norm} adalah intensitas setelah standardisasi, μ adalah rerata intensitas citra (atau rerata per kanal), dan σ adalah simpangan baku intensitas citra (atau per kanal). Dalam praktik, μ dan σ dihitung pada ruang intensitas yang sama dengan I dan digunakan konsisten agar skala masukan antar citra lebih seragam.

Setelah normalisasi/standardisasi, penyeragaman ukuran citra dilakukan karena model deteksi objek memerlukan dimensi masukan yang tetap. Teknik *letterbox resizing* digunakan untuk menyesuaikan citra ke resolusi target tanpa mengubah rasio aspek asli objek. Faktor skala s serta ukuran hasil penskalaan (W', H') ditentukan menggunakan Persamaan (2.18).

$$s = \min\left(\frac{W_t}{W}, \frac{H_t}{H} \right), \quad W' = \lfloor sW \rfloor, \quad H' = \lfloor sH \rfloor \tag{2.18}$$

dengan keterangan: W dan H adalah lebar dan tinggi citra asli, W_t dan H_t adalah lebar dan tinggi target masukan model, s adalah faktor skala yang dipilih agar citra hasil skala tidak melebihi dimensi target pada salah satu sumbu, dan W' serta H' adalah lebar dan tinggi citra setelah penskalaan. Operator $\lfloor \cdot \rfloor$ menyatakan pembulatan ke bawah untuk menjaga koordinat piksel berbasis bilangan bulat.

Selanjutnya, padding horizontal dan vertikal (p_x, p_y) digunakan untuk menempatkan citra hasil skala pada kanvas berukuran target, sebagaimana didefinisikan pada Persamaan (2.19). Dengan langkah ini, bentuk plankton tetap proporsional dan tidak terdistorsi akibat penskalaan.

$$p_x = \frac{W_t - W'}{2}, \quad p_y = \frac{H_t - H'}{2} \tag{2.19}$$

dengan keterangan: p_x dan p_y adalah padding pada arah lebar dan tinggi, sedangkan W_t, H_t, W', H' sama seperti pada Persamaan (2.18). Padding dapat direalisasikan sebagai penambahan piksel bernilai konstan (misalnya 0) pada sisi kiri-kanan dan atas-bawah secara simetris agar objek tidak mengalami perubahan rasio aspek.

Selain penyeragaman ukuran, peningkatan keterbacaan tekstur mikro dilakukan menggunakan *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Teknik ini meningkatkan kontras secara lokal sehingga detail halus seperti ornamen permukaan, garis radial, atau struktur internal plankton menjadi lebih jelas pada tiap wilayah citra. Karena peningkatan kontras lokal berpotensi memperkuat *noise*, CLAHE membatasi penguatan histogram melalui mekanisme *clip limit*. Batas klip tersebut dinyatakan pada Persamaan (2.20).

$$\text{LimClip} = \mathcal{X}(x, y)\, T_{gl} \left(1 + \frac{\phi}{\phi_{\text{max}}} (h_{\text{max}} - 1) \right) \tag{2.20}$$

dengan keterangan: $\mathcal{X}(x, y)$ menyatakan jumlah piksel pada satu blok lokal (tile) CLAHE yang memuat posisi (x, y), T_{gl} adalah jumlah level intensitas (misalnya $T_{gl} = 256$ untuk citra 8-bit), ϕ adalah parameter pengendali batas klip (*clip factor*) yang dipilih dalam rentang tertentu, ϕ_{max} adalah nilai maksimum parameter batas klip pada konfigurasi yang digunakan, dan h_{max} menyatakan nilai puncak histogram yang diizinkan (atau faktor pembatas puncak histogram) sehingga penguatan kontras tidak berkembang tanpa kendali. Dengan formulasi ini, *clip limit* ditentukan oleh ukuran tile dan konfigurasi pembatas sehingga redistribusi histogram tetap stabil.

Setelah pembatasan histogram, redistribusi intensitas dilakukan melalui fungsi distribusi kumulatif (*cumulative distribution function*, CDF) pada setiap blok

lokal, sebagaimana ditunjukkan pada Persamaan (2.21).

$$\mathcal{J} = \frac{T_{gl} - 1}{\mathcal{X}(x, y)} \sum_{l=0}^{L} CDF(l) \tag{2.21}$$

dengan keterangan: $\mathcal{J}$ adalah nilai intensitas keluaran (atau fungsi pemetaan intensitas) setelah ekualisasi lokal, T_{gl} adalah jumlah level intensitas, $\mathcal{X}(x, y)$ adalah jumlah piksel pada tile terkait, l adalah indeks level intensitas, $CDF(l)$ adalah nilai distribusi kumulatif pada level l, dan L adalah level intensitas maksimum yang dipertimbangkan (umumnya $L = T_{gl} - 1$). Dengan demikian, pemetaan CDF mengubah intensitas berdasarkan distribusi lokal agar kontras pada tile meningkat secara adaptif.

Karena CLAHE bekerja pada blok-blok lokal, interpolasi bilinear digunakan untuk menghaluskan transisi antarblok agar tidak menimbulkan artefak batas tile. Mekanisme interpolasi ini dinyatakan pada Persamaan (2.22).

$$I'(x, y) = \alpha\beta\, T_A + (1 - \beta)\, T_B + (1 - \alpha)\beta\, T_C + (1 - \alpha)(1 - \beta)\, T_D \tag{2.22}$$

dengan keterangan: $I'(x, y)$ adalah intensitas hasil interpolasi pada posisi (x, y), T_A, T_B, T_C, T_D adalah nilai hasil pemetaan CLAHE pada empat tile tetangga yang mengapit posisi (x, y), sedangkan $\alpha \in [0, 1]$ dan $\beta \in [0, 1]$ adalah bobot jarak relatif (fraksional) terhadap sumbu x dan y di dalam sel interpolasi. Interpolasi ini memastikan perubahan kontras lokal tidak menciptakan diskontinuitas visual pada perbatasan tile.

Setelah seluruh tahap prapemrosesan dilakukan, augmentasi data diterapkan untuk meningkatkan keragaman sampel pelatihan secara sintetis. Augmentasi dirancang untuk merepresentasikan variasi alami plankton, seperti perubahan orientasi, posisi, dan variasi tampilan akibat perbedaan iluminasi atau pengaturan optik. Agar augmentasi tetap terkontrol dan tidak mengubah identitas biologis plankton, teknik yang digunakan dibedakan menjadi augmentasi geometrik dan augmentasi fotometrik.

Augmentasi geometrik berfokus pada perubahan orientasi, posisi, dan komposisi spasial objek tanpa mengubah karakter biologisnya. Teknik seperti *horizontal flip*, *vertical flip*, rotasi terbatas, *shear transformation*, dan *random cropping* digunakan untuk mensimulasikan orientasi plankton di kolom air serta variasi posisi preparat pada bidang pandang mikroskop. Pendekatan ini membantu model menjadi lebih toleran terhadap perubahan pose dan variasi skala objek,

sehingga deteksi lebih robust ketika objek muncul pada orientasi yang beragam.

Sebaliknya, augmentasi fotometrik berfokus pada perubahan intensitas dan kontras tanpa mengubah struktur geometris objek. Pada penelitian ini, augmentasi fotometrik mencakup normalisasi intensitas pada Persamaan (2.15) dan Persamaan (2.16), standardisasi pada Persamaan (2.17), serta peningkatan kontras lokal menggunakan CLAHE pada Persamaan (2.20) hingga Persamaan (2.21). Kategori ini membantu menstabilkan variasi tampilan citra mikroskopik sehingga model tidak mudah terpengaruh oleh fluktuasi pencahayaan atau perbedaan pengaturan akuisisi.

Ringkasan kategori dan jenis augmentasi yang digunakan dalam penelitian ini disajikan pada Tabel 2.6.

Tabel 2.6 Kategori dan jenis augmentasi pada citra mikroskopik plankton.

Kategori Augmentasi	**Jenis Augmentasi**	**Tujuan Utama**
Geometrik	*Horizontal flip*, *vertical flip*, rotasi terbatas ($\pm 10°$ hingga $\pm 20°$), *shear* ringan pada sumbu-x dan sumbu-y, serta *cropping* diikuti *resizing*	Mensimulasikan variasi orientasi dan posisi plankton di dalam medium cair, serta menyesuaikan bidang pandang dan skala objek tanpa mengubah karakteristik morfologi utama.
Fotometrik	Penyesuaian intensitas pencahayaan, kontras, dan saturasi dalam rentang terbatas, serta peningkatan kontras lokal menggunakan CLAHE	Mengurangi pengaruh fluktuasi iluminasi mikroskopik dan meningkatkan visibilitas struktur morfologi halus sehingga fitur visual lebih konsisten untuk proses ekstraksi ciri.

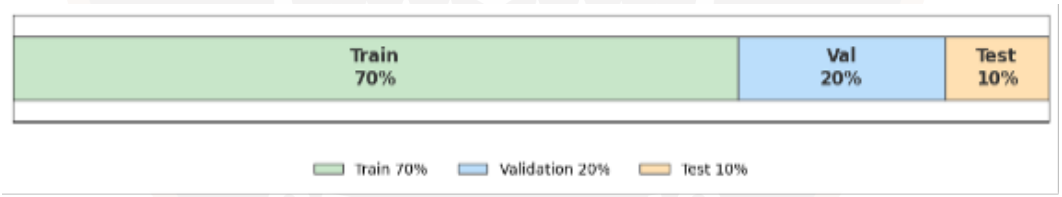
Secara keseluruhan, rangkaian penyiapan dan augmentasi data ini dirancang untuk membentuk masukan citra yang konsisten, kaya variasi, dan tetap mempertahankan karakter biologis plankton. Normalisasi, standardisasi, dan penyeragaman ukuran menjaga keseragaman format masukan, sedangkan CLAHE memperbaiki keterbacaan tekstur mikro secara lokal dengan kontrol penguatan yang mencegah peningkatan *noise*. Di sisi lain, augmentasi geometrik dan fotometrik memperluas ruang variasi pelatihan secara terstruktur tanpa mengubah identitas visual utama plankton. Dengan kombinasi prapemrosesan yang terkontrol dan augmentasi yang sistematis, model deteksi diharapkan memiliki kemampuan

generalisasi yang lebih baik terhadap variasi kondisi citra mikroskopik pada aplikasi nyata.

2.3.11 Evaluasi Kinerja, Efisiensi, dan Validasi Statistik

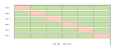
Validasi pada model deteksi objek bertujuan untuk menilai kemampuan generalisasi ketika sistem dihadapkan pada distribusi data baru, sekaligus mengidentifikasi potensi *overfitting* dan sensitivitas terhadap variasi visual (Bengio, 2012). Pada citra mikroskopik, kebutuhan validasi menjadi semakin krusial karena objek berukuran kecil, bertekstur halus, serta memiliki kemiripan morfologi antarkelas. Dalam konteks dataset tidak seimbang, validasi juga memastikan bahwa peningkatan kinerja tidak semata-mata didorong oleh dominasi kelas mayoritas, melainkan mencerminkan kemampuan model dalam mengenali kelas minoritas secara konsisten.

Dua strategi validasi yang umum digunakan adalah *Stratified Hold-Out* (Gambar 2.13) dan *Stratified K-Fold* (Gambar 2.14). Pada *Stratified Hold-Out*, data dipisahkan menjadi subset pelatihan, validasi, dan pengujian dengan proporsi kelas yang dipertahankan serupa, sehingga memberikan estimasi cepat terhadap perilaku model pada satu konfigurasi pembagian data (Ünalan et al., 2024). Sebaliknya, *Stratified K-Fold Cross Validation* (Gambar 2.14) melakukan evaluasi berulang pada beberapa lipatan, sehingga variansi estimasi dapat ditekan dan sensitivitas model terhadap variasi distribusi data dapat diamati lebih stabil (T R et al., 2023).



Gambar 2.13 Ilustrasi pembagian data menggunakan *Stratified Hold-Out*.

Evaluasi deteksi objek menilai dua aspek utama secara simultan, yaitu ketepatan pengenalan kelas dan ketepatan lokalisasi objek (Padilla et al., 2020). Precision merefleksikan proporsi *false positive*, sedangkan recall merefleksikan proporsi *false negative*. Keseimbangan keduanya dirangkum melalui F1-score. Untuk memahami struktur kesalahan antarkelas secara rinci, analisis dilakukan melalui *confusion matrix* sebagaimana ditunjukkan pada Gambar 2.15. Elemen diagonal merepresentasikan prediksi benar per kelas, sedangkan elemen nondiagonal mengindikasikan pola salah-klasifikasi.



Gambar 2.14 Ilustrasi evaluasi menggunakan *Stratified K-Fold Cross Validation*.

Kualitas lokalisasi objek ditentukan oleh ukuran tumpang tindih antara kotak prediksi dan *ground truth*, yang menjadi dasar penentuan apakah prediksi dianggap benar pada ambang tertentu. Ringkasan performa agregat biasanya dinyatakan melalui konsep Average Precision dan *mean* Average Precision (mAP), yang dirumuskan lebih lanjut pada Bab III agar selaras dengan implementasi eksperimen.

Selain strategi validasi, konfigurasi pelatihan memengaruhi dinamika optimisasi parameter (Goodfellow et al., 2016; LeCun et al., 2015). Secara umum, pembaruan parameter berbasis *gradient descent* dinyatakan pada Persamaan 2.23.

$$W_{t+1} = W_t - \alpha \nabla L(W_t), \tag{2.23}$$

Pada deteksi objek, fungsi kerugian merupakan kombinasi komponen lokalisasi, *objectness*, dan klasifikasi, sehingga perilaku konvergensi dipengaruhi oleh keseimbangan antar komponen tersebut.

Pada skenario pelatihan skala besar, *Automatic Mixed Precision* (AMP) digunakan untuk mempercepat komputasi sekaligus menjaga stabilitas numerik. Secara konseptual, mekanisme ini dinyatakan pada Persamaan 2.24.

$$W_{t+1} = W_t - \alpha \cdot \text{scale} \cdot \frac{\nabla L(W_t)}{\text{scale}}, \tag{2.24}$$

Fungsi aktivasi berperan dalam menjaga aliran gradien dan kualitas representasi. Aktivasi SiLU dinyatakan pada Persamaan 2.25, sedangkan fungsi sigmoid untuk pembentukan probabilitas ditunjukkan pada Persamaan 2.26. Pada blok Transformer, GELU digunakan sebagaimana pada Persamaan 2.27, dan distribusi probabilitas akhir dibentuk melalui Softmax pada Persamaan 2.28.

$$\text{SiLU}(x) = x \cdot \sigma(x), \tag{2.25}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2.26}$$

$$\text{GELU}(x) = x\Phi(x), \tag{2.27}$$

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}, \tag{2.28}$$

Pada kondisi ketidakseimbangan kelas, *Focal Loss* digunakan untuk menurunkan kontribusi contoh mudah dan menekankan contoh sulit. Formulasinya dinyatakan pada Persamaan 2.29.

$$\text{FL}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t), \tag{2.29}$$

Sebagai bentuk regulerisasi, *L2 weight decay* diterapkan sebagaimana ditunjukkan pada Persamaan 2.30.

$$L_{\text{reg}} = L + \lambda\|W\|_2^2, \tag{2.30}$$

Untuk memastikan bahwa perbedaan performa antarmodel tidak sekadar akibat fluktuasi sampel, digunakan pendekatan inferensial nonparametrik. Nilai-p pada uji permutasi dinyatakan pada Persamaan 2.31.

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(T^{(b)} \geq T_{\text{obs}})}{B+1}, \tag{2.31}$$

Pada desain berblok, prosedur Freedman–Lane membentuk respons semu sebagaimana pada Persamaan 2.32.

$$y^* = \hat{y} + e^*, \tag{2.32}$$

Untuk perbandingan berpasangan, statistik rerata selisih dinyatakan pada Persamaan 2.33.

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad d_i = x_i - y_i, \tag{2.33}$$

Koreksi multipengujian menggunakan prosedur Holm yang dirumuskan pada Persamaan 2.34.

$$p_{(j)} \leq \frac{\alpha}{m-j+1}, \tag{2.34}$$

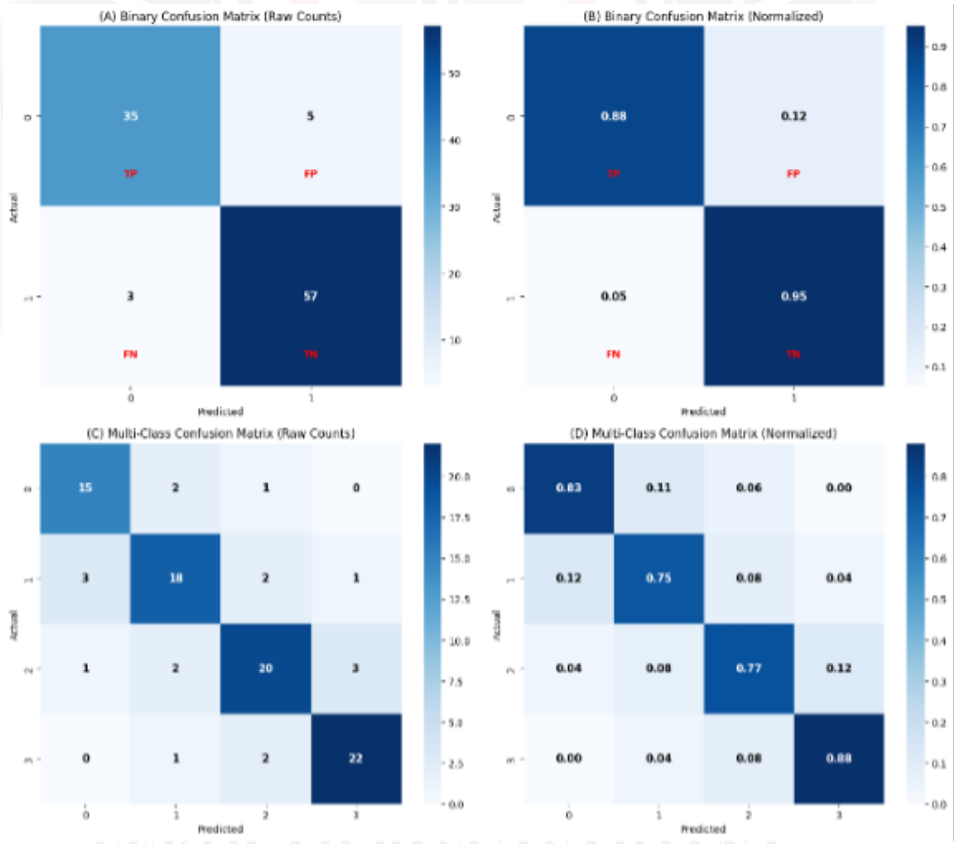
Interpretabilitas model dilakukan melalui pendekatan berbasis CAM. Perbandingan visual Grad-CAM dan EigenCAM ditunjukkan pada Gambar 2.16. Secara matematis, pembentukan EigenCAM mengikuti Persamaan 2.35–2.37.

$$\Sigma = X^{\top} X, \tag{2.35}$$

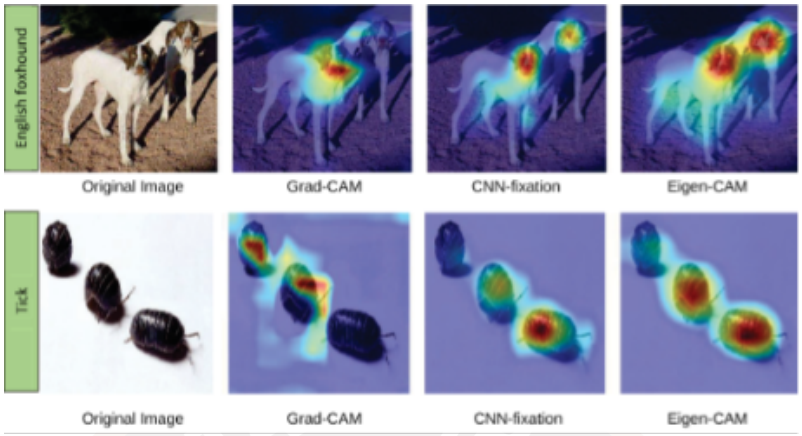
$$\Sigma v = \lambda v, \tag{2.36}$$

$$A_{\text{EigenCAM}} = X v_{\max}, \tag{2.37}$$

Dengan demikian, kerangka evaluasi deteksi objek dalam penelitian ini mencakup validasi terstratifikasi (Gambar 2.13 dan Gambar 2.14), analisis pola kesalahan melalui confusion matrix (Gambar 2.15), pemodelan optimisasi parameter (Persamaan 2.23–2.30), uji statistik nonparametrik (Persamaan 2.31–2.34), serta interpretabilitas berbasis CAM (Persamaan 2.35–2.37).



Gambar 2.15 *Confusion matrix* untuk menunjukkan pola kesalahan klasifikasi antarkelas.



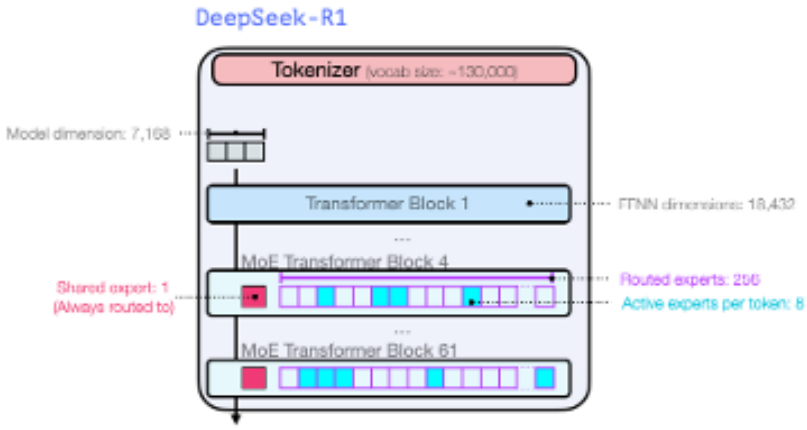
Gambar 2.16 Perbandingan visual pendekatan Grad-CAM dan EigenCAM untuk interpretabilitas model.

2.3.12 Sistem Pendukung Keputusan Visual Semantik Berbasis Large Language Models

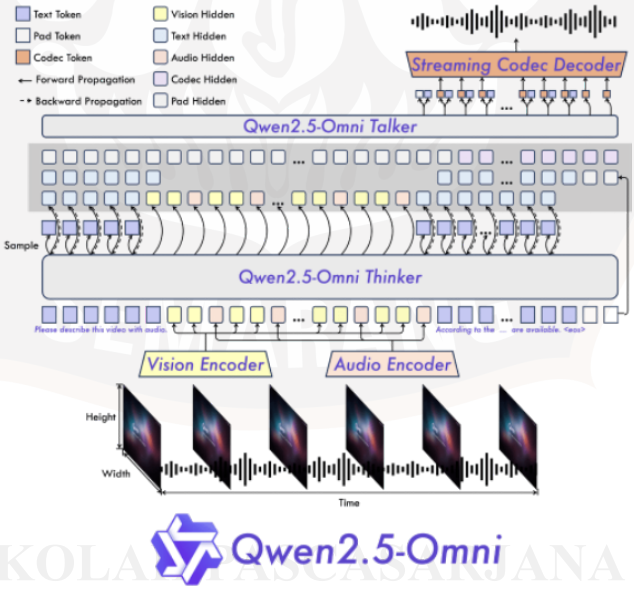
Sebagai tindak lanjut dari pengembangan model deteksi plankton, penelitian ini menghasilkan prototipe sistem pendukung keputusan berbasis web yang mengintegrasikan deteksi objek berbasis YOLO (YOLOv8m dan ACViT–YOLO) dengan kemampuan penjelasan berbasis *Large Language Model* (LLM). Integrasi ini memungkinkan sistem menghasilkan keluaran multimodal dalam satu alur kerja, yaitu prediksi visual berupa lokasi dan label spesies plankton, serta uraian taksonomi dalam bahasa alami. Untuk komponen penjelasan, prototipe memanfaatkan model LLM multimodal modern, termasuk Gemini sebagai model general-purpose multimodal (Google DeepMind, 2023), Qwen-VL sebagai model vision–language yang mendukung pemahaman visual dan deskripsi terstruktur (Bai et al., 2023), serta DeepSeek-VL yang dirancang untuk generalisasi visual-linguistik yang kuat (DeepSeek-VL Team, 2023; Wadekar, 2025a). Pendekatan ini menempatkan sistem tidak hanya sebagai alat identifikasi visual, tetapi juga sebagai media interpretatif yang membantu pengguna non-ahli memahami hasil deteksi secara kontekstual.

Arsitektur sistem dibangun dengan pola *client–server*, di mana *frontend* berbasis React menyediakan antarmuka interaktif, sedangkan *backend* berbasis Flask berperan sebagai API yang mengelola unggah citra, inferensi model YOLO, serta pemanggilan LLM untuk menghasilkan penjelasan berbasis hasil deteksi. Setelah proses inferensi, sistem menampilkan keluaran terpadu berupa citra dengan *bounding box* dan label spesies, disertai ringkasan prediksi serta deskripsi taksonomi berbasis LLM. Untuk menjaga fokus pembahasan pada alur kerja dan integrasi

sistem, variasi keluaran penjelasan dari masing-masing LLM tidak diuraikan secara rinci pada subbab ini dan disajikan pada Lampiran 8. Secara keseluruhan, prototipe ini merepresentasikan integrasi *computer vision* dan *vision–language models* dalam satu sistem pendukung keputusan multimodal yang aplikatif dan interpretatif. Gambar ilustrasi arsitektur DeepSeeks dan Gemini disajikan pada Gambar 2.17 dan Gambar 2.18.



Gambar 2.17 Ilustrasi arsitektur konseptual DeepSeek R1 untuk *reasoning* kontekstual.
Sumber: (Wadekar, 2025a).



Gambar 2.18 Arsitektur konseptual Qwen 2.5 untuk integrasi informasi visual–teks.
Sumber: (Wadekar, 2025b).