BAB II
TINJAUAN LITERATUR

2.1 Penelitian Terkait

Perkembangan teknologi *computer vision*, khususnya pada algoritma *You Only Look Once* (*YOLO*), menawarkan solusi yang tepat dalam mendeteksi objek secara akurat dan *real-time*. Algoritma *YOLO* dikenal memiliki kinerja yang baik dalam mengenali berbagai objek bergerak di lingkungan sekitar, sehingga mendukung proses deteksi objek secara akurat dan efisien dalam berbagai kondisi(Simbolon, 2025). Kemampuan tersebut mendorong banyak penelitian untuk menerapkan *YOLO* dalam berbagai sistem deteksi objek.

Salah satu penelitian dengan judul "Sistem Deteksi Objek Manusia Menggunakan Algoritma *YOLOv8* Berbasis Kamera *Depth* Sensor (Studi Kasus: Cv. Ateri Global Teknologi)" (Zulkarnaen, 2024) menunjukkan bahwa *YOLO*v8 mampu digunakan secara efektif dalam deteksi objek manusia dengan tingkat akurasi yang tinggi. Dukungan informasi kedalaman terbukti meningkatkan kinerja sistem, dengan hasil pengujian mencapai akurasi sebesar 92%, meskipun dipengaruhi oleh sudut pengambilan gambar.

Selain diterapkan pada deteksi objek manusia, *YOLO* juga banyak digunakan dalam konteks pemantauan lingkungan berbasis video(Zulkiflie, 2021). Penelitian terkait lainnya dengan judul "Implementasi *YOLO* untuk Menghitung Kepadatan Kendaraan Tempat Parkir" (Hidayat, Umbara and Ilyas, 2025) menerapkan algoritma *YOLOv5* untuk mendeteksi dan menghitung kendaraan guna memantau kepadatan lalu lintas kampus secara *real-time*. Hasil pengujian

menunjukkan kinerja deteksi yang baik, dengan *precision* hingga 1.00 dan *recall* mencapai 0.90, sehingga sistem mampu menyajikan informasi kepadatan kendaraan secara cepat dan akurat.

Adapun penelitian terdahulu dengan judul "*ZoeDepth*: *Zero-shot transfer by Combining Relative and Metric Depth*" (Bhat *et al.*, 2023) menunjukkan bahwa penerapan *ZoeDepth* mampu memberikan kinerja estimasi jarak yang kompetitif pada berbagai *dataset* standar. Hasil pengujian menunjukkan peningkatan performa yang signifikan pada *dataset NYU Depth v2*, serta kemampuan model dalam melakukan pengujian pada *dataset* yang belum pernah digunakan sebelumnya tanpa penurunan kinerja yang berarti.

Adapun Penelitian terkait dengan judul "*Depth-aware Image-Space Fogging Algorithm for Background Privacy Protection using ZoeDepth Framework*" cepat(Liang, Wang and Pan, 2025) menggunakan *ZoeDepth* untuk memperkirakan jarak objek dari citra tunggal, sehingga sistem dapat membedakan area latar depan dan latar belakang berdasarkan jaraknya. Informasi jarak tersebut dimanfaatkan untuk menerapkan efek fogging pada latar belakang, sehingga detail objek utama tetap terlihat sementara informasi latar belakang menjadi tidak jelas.

Pada penelitian yang berjudul "*Benchmark on Monocular Metric Depth Estimation in Wildlife Setting*" oleh (Niccoli *et al.*, 2025). Penelitian ini mengevaluasi beberapa metode *Monocular Depth Estimation* pada 93 citra camera *trap* untuk mendukung pemantauan satwa liar. Hasil penelitian menunjukkan bahwa *Depth Anything V2* memberikan akurasi terbaik, sedangkan *ZoeDepth* mengalami penurunan performa pada lingkungan luar ruang. Selain itu, metode

ekstraksi kedalaman berbasis median terbukti lebih akurat dibandingkan mean, sehingga direkomendasikan untuk digunakan. Penelitian ini juga menyarankan pengembangan *dataset* yang lebih sesuai dengan lingkungan satwa liar serta optimalisasi model agar akurasi estimasi kedalaman dapat ditingkatkan.

Pada penelitian terdahulu yang lain oleh (Andres *et al.*, 2026) dengan judul "*Advancing Offshore Safety: Monocular Depth Estimation from 360-Degree Images for Enhanced Oil Platform Inspection*" yang menunjukkan perbandingan beberapa metode *Monocular Depth Estimation* pada citra 360° *platform* minyak lepas pantai. Hasilnya menunjukkan bahwa *Depth Anything V2 Indoor* memberikan performa terbaik dengan *RMSE* 3,0 m (*Platform* A) dan 1,0 m (*Platform* B), sedangkan *ZoeDepth* Indoor memperoleh *RMSE* 3,5 m dan 1,3 m. Penelitian juga menyimpulkan bahwa kesalahan prediksi meningkat pada objek yang lebih jauh sehingga diperlukan *fine-tuning* agar akurasi estimasi kedalaman meningkat.

2.2 Deteksi Objek

Deteksi objek merupakan teknik dalam *computer vision* yang digunakan untuk mengenali jenis objek sekaligus menentukan lokasi objek tersebut pada gambar atau video. Hasil deteksi berupa posisi objek, label kelas, dan nilai kepercayaan terhadap objek yang terdeteksi(Ravpreet Kaur, 2023; Apridiansyah *et al.*, 2025). Dengan memanfaatkan informasi antar*frame* untuk meningkatkan akurasi dan kestabilan deteksi objek. Informasi temporal pada video membantu mengurangi kesalahan deteksi yang dapat muncul jika setiap *frame* diproses secara independen(Jiao L, Zhang R, Liu F, Yang S, Hou B, Li L, 2022).

Dalam *computer vision*, terdapat beberapa metode yang digunakan untuk memahami informasi visual, yaitu *image classification*, *object detection*, dan *object segmentation*. *Image classification* merupakan metode yang bertujuan mengidentifikasi kategori dari suatu gambar secara keseluruhan tanpa mengetahui posisi objek yang terdapat di dalamnya. Berbeda dengan metode tersebut, *object detection* mampu mengenali keberadaan satu atau lebih objek sekaligus menentukan letaknya pada gambar dengan memberikan *bounding box* pada setiap objek yang terdeteksi(Apridiansyah *et al.*, 2025). Adapun *object segmentation* memiliki kemampuan yang lebih detail karena tidak hanya mengenali dan menentukan lokasi objek, tetapi juga memisahkan setiap objek berdasarkan piksel penyusunnya. Dengan demikian, metode ini dapat menghasilkan bentuk objek secara lebih presisi sehingga lebih efektif digunakan pada aplikasi yang memerlukan informasi detail(Liang, Wang and Pan, 2025).

2.3 Estimasi Jarak

Depth Estimation merupakan teknik dalam *computer vision* yang digunakan untuk memperkirakan kedalaman atau jarak suatu objek terhadap kamera berdasarkan informasi visual dari sebuah citra. Salah satu model yang banyak diterapkan untuk tugas ini adalah *ZoeDepth*, yaitu model berbasis *deep learning* yang dirancang khusus untuk melakukan estimasi kedalaman secara *monocular*, sehingga hanya memerlukan satu gambar *RGB* sebagai masukan(Takahashi and Hayashi, 2026). Dalam proses estimasinya, *ZoeDepth* memanfaatkan kerangka kerja *MiDaS* untuk menghasilkan informasi kedalaman relatif dari citra, kemudian mengintegrasikan hasil tersebut dengan modul *metric*

bin untuk mengonversinya menjadi nilai kedalaman metrik yang lebih akurat. Kombinasi kedua proses tersebut memungkinkan *ZoeDepth* menghasilkan estimasi jarak yang lebih presisi, sehingga mampu memberikan kinerja yang baik pada berbagai kondisi, termasuk pada lingkungan dengan tingkat kompleksitas tinggi maupun variasi pencahayaan yang berbeda(Bini and Manni, 2025). Secara matematis, estimasi jarak objek dapat dinyatakan sebagai:

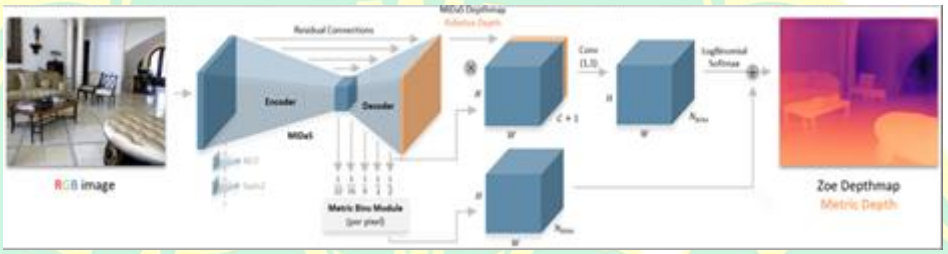
$$Z = D(x_c, y_c)$$

YOLO terlebih dahulu digunakan untuk mendeteksi objek pada citra atau frame video dan menghasilkan bounding box yang menunjukkan posisi objek. Dari bounding box tersebut ditentukan titik pusat objek (x_c, y_c) yang diperoleh dari nilai rata-rata koordinat sudut kiri atas dan sudut kanan bawah bounding box. Selanjutnya, koordinat pusat tersebut digunakan sebagai posisi acuan untuk mengambil nilai kedalaman pada depth map yang dihasilkan oleh ZoeDepth. Dengan demikian, $D(x_c, y_c)$ menunjukkan nilai kedalaman pada titik pusat objek, sedangkan Z merupakan nilai estimasi kedalaman objek terhadap kamera.(Driver and Systems, 2022).

2.4 *ZoeDepth*

ZoeDepth merupakan model *Monocular Depth Estimation* berbasis *deep learning* yang bekerja memanfaatkan kerangka kerja *MiDaS* untuk mengekstraksi fitur multi-skala(Faseeh *et al.*, 2024). Fitur *multi-scale decoder* kemudian dihubungkan ke model *metric* bins untuk memperkirakan pusat bin kedalaman pada setiap piksel. Memungkinkan *ZoeDepth* menghasilkan kedalaman absolut

dengan akurasi yang lebih stabil. Keunggulan utama *ZoeDepth* terletak pada kemampuannya memanfaatkan informasi multi-skala, fleksibilitas penggunaan *backbone transformer*, serta kemampuannya menghasilkan estimasi kedalaman metrik yang sesuai. Model ini mampu melakukan generalisasi pada berbagai kondisi lingkungan tanpa memerlukan proses pelatihan ulang secara khusus (*zero-shot transfer*)(Katiyar *et al.*, 2024)(Sunoto *et al.*, 2025).



Gambar 2. 1 Arsitektur *ZoeDepth* (Bhat *et al.*, 2023)

Secara matematis, zoedepth dapat dinyatakan sebagai:

$$D(x, y) = f\,\theta(I)$$

menunjukkan bahwa ZoeDepth menghasilkan nilai kedalaman D(x,y) pada setiap piksel citra berdasarkan citra RGB I sebagai input. Fungsi $f\,\theta(I)$ merupakan model deep learning ZoeDepth dengan parameter θ yang telah dilatih untuk mengestimasi kedalaman. Dalam penelitian ini, hasil depth map digunakan untuk memperoleh nilai jarak objek yang telah dideteksi oleh YOLO, dengan mengambil nilai kedalaman pada titik tengah bounding box objek.

2.5 *YOLO V10*

YOLOv10 dikembangkan untuk menghasilkan model deteksi objek yang memiliki keseimbangan lebih baik antara tingkat akurasi dan efisiensi komputasi. Peningkatan tersebut dicapai melalui berbagai penyempurnaan pada arsitektur model serta optimalisasi proses setelah deteksi (*post-processing*), sehingga performa sistem secara keseluruhan menjadi lebih baik (Wang and Liao, 2024). Salah satu pembaruan penting pada *YOLOv10* adalah kemampuannya mengatasi keterbatasan yang masih ditemukan pada versi *YOLO* sebelumnya, khususnya ketergantungan terhadap proses *Non-Maximum Suppression* (*NMS*). Pengurangan ketergantungan terhadap *NMS* membuat proses inferensi menjadi lebih cepat, sekaligus meningkatkan efisiensi komputasi tanpa mengurangi kualitas hasil deteksi(Hassani *et al.*, 2026).

Selain itu, *YOLOv10* menerapkan desain arsitektur yang dirancang untuk mengoptimalkan akurasi dan efisiensi secara bersamaan. Model ini juga memperkenalkan mekanisme consistent dual assignment yang mendukung penerapan *NMS-free training protocol*, sehingga proses pelatihan dan prediksi menjadi lebih sederhana. Dengan pendekatan tersebut, waktu yang diperlukan pada tahap pascapemrosesan dapat dikurangi, latensi inferensi menjadi lebih rendah, dan model tetap mampu memberikan hasil deteksi yang akurat serta memiliki performa yang kompetitif dibandingkan versi sebelumnya(Kim, 2024).

2.6 *COCO val2017*

COCO val2017 merupakan bagian dari *COCO Detection* 2017 yang digunakan sebagai *dataset* validasi (validation *dataset*) dalam proses pengembangan dan evaluasi model deteksi objek. *Dataset* ini terdiri atas 5.000 citra yang memuat 36.781 objek beranotasi dari 80 kategori objek. Setiap objek telah dilengkapi dengan anotasi berupa kelas objek dan *bounding box*, sehingga dapat digunakan sebagai acuan dalam mengukur akurasi hasil deteksi. Keberagaman kategori objek, variasi ukuran, posisi, serta kondisi lingkungan pada setiap citra menjadikan *COCO val2017* sebagai salah satu *benchmark* yang banyak dimanfaatkan untuk menguji, mengevaluasi, dan membandingkan performa berbagai model deteksi objek, termasuk model *YOLO*(Tomasz and Jakub, 2022)(Kar *et al.*, 2023).

Dataset Microsoft Common Objects in Context (MS COCO) menyediakan koleksi citra yang beragam dengan objek-objek yang berada pada kondisi lingkungan nyata. Variasi kategori objek, ukuran, posisi, serta tingkat kompleksitas latar belakang pada *dataset* ini memungkinkan model mempelajari karakteristik objek secara lebih menyeluruh. Dengan demikian, model *YOLO* memiliki kemampuan yang lebih baik dalam mengenali dan mendeteksi berbagai jenis objek pada beragam kondisi dan skenario pengujian(Diwan, Anirudh and Tembhurne, 2023).

2.7 *Dataset iBims-1*

Dataset iBims-1 (*Independent Benchmark Images and matched Scans v1*) merupakan *dataset benchmark* yang dikembangkan secara khusus untuk

mengevaluasi performa metode Single Image *Depth Estimation* (*SIDE*). *Dataset* ini terdiri atas pasangan citra *RGB* dan *ground truth depth* yang diperoleh melalui proses akuisisi menggunakan kamera dan pemindaian laser berpresisi tinggi. Selain menyediakan data kedalaman sebagai nilai acuan (*ground truth*), *iBims-1* juga dilengkapi dengan berbagai anotasi pendukung yang memungkinkan evaluasi estimasi kedalaman dilakukan secara objektif dan komprehensif(Liu *et al.*, 2025). *Dataset iBims-1* memiliki beberapa keunggulan yang menjadikannya banyak digunakan sebagai *benchmark* dalam penelitian estimasi kedalaman.

Pada penelitian ini, *dataset iBims-1* digunakan sebagai *dataset* evaluasi untuk mengukur akurasi estimasi jarak yang dihasilkan oleh model *ZoeDepth*. *Dataset* ini dipilih karena menyediakan pasangan citra *RGB* dan *ground truth Depth* yang diperoleh melalui pemindaian laser berpresisi tinggi sehingga dapat digunakan sebagai nilai acuan dalam menghitung *Root Mean Square Error* (*RMSE*)(Technology, no date).

2.8 *Python*

Python merupakan salah satu bahasa pemrograman yang paling banyak digunakan dalam pengembangan *machine learning* dan *deep learning* karena memiliki sintaks yang mudah dipahami serta didukung oleh ekosistem yang lengkap. Berbagai *library* seperti *PyTorch, OpenCV*, dan *Ultralytics* menyediakan beragam fungsi yang memudahkan proses pengembangan model, mulai dari pengolahan data, pemrosesan citra dan video, hingga pelatihan, inferensi, serta evaluasi model(Jalolov, 2023). Selain itu, *Python* juga menawarkan fleksibilitas yang tinggi dalam mengimplementasikan berbagai algoritma. Sebuah tinjauan

komprehensif menunjukkan bahwa library *deep learning* berbasis *Python*, seperti *PyTorch*, memiliki kemampuan optimasi yang baik serta mendukung berbagai arsitektur *deep learning* untuk beragam aplikasi, seperti klasifikasi citra, deteksi objek, dan *Natural Language Processing* (*NLP*). Oleh karena itu, pemahaman terhadap karakteristik dan keunggulan masing-masing library menjadi penting agar peneliti maupun pengembang dapat memilih perangkat yang paling sesuai dengan kebutuhan penelitian atau pengembangan aplikasi. (Mohialden *et al.*, 2024)(Puspadini and Zen, no date).