

Moving Object Detection Using The Yolo Method And Video-Based Depth Estimation

Yuza Reswan ¹⁾; Erwin Dwika Putra ²⁾; Yovi Apridiansyah ³⁾; Dwi Bagas Prakoso ⁴⁾

^{1,2,3,4)}Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Bengkulu

Email: ¹⁾ yuzareswan@umb.ac.id

How to Cite :

Reswan, Y., Putra, E. D., Apridiansyah, Y., Prakoso, D. B. (2026). Moving Object Detection Using The Yolo Method and Video-Based Depth Estimation. Jurnal Media Computer Science, 5(4).

ARTICLE HISTORY

Received [24 Agustus 2026]

Revised [25 Oktober 2026]

Accepted [27 Oktober 2026]

KEYWORDS

Object Detection, Distance

Estimation, YOLOv10, ZoeDepth.

This is an open access article under the
[CC-BY-SA](#) license



ABSTRAK

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence) khususnya bidang computer vision telah mendorong perkembangan pengolahan citra dan video untuk memahami kondisi lingkungan secara otomatis. Salah satu penerapannya adalah deteksi objek, tetapi metode deteksi umumnya hanya memberikan informasi jenis dan lokasi objek dalam bounding box tanpa informasi jarak terhadap kamera. Keterbatasan tersebut menyebabkan informasi yang dihasilkan belum cukup untuk mendukung sistem yang membutuhkan pemahaman lingkungan secara lebih lengkap. Oleh karena itu, diperlukan pendekatan yang menggabungkan deteksi objek dan estimasi jarak menggunakan kamera RGB tanpa perangkat keras tambahan. Penelitian ini bertujuan mengembangkan sistem deteksi objek dan estimasi jarak berbasis video menggunakan kombinasi YOLOv10 dan ZoeDepth. YOLOv10 digunakan untuk mendeteksi objek dan menentukan posisi pada setiap frame video, sedangkan ZoeDepth digunakan untuk melakukan estimasi kedalaman (Depth Estimation) guna memperoleh perkiraan jarak objek terhadap kamera. Metode penelitian meliputi pengumpulan data, prapemrosesan, implementasi model, dan evaluasi sistem. Data yang digunakan berupa 250 sampel video CCTV, terdiri dari 162 data sekunder dan 88 data primer. Seluruh data digunakan sebagai data pengujian karena YOLOv10 merupakan model pre-trained yang telah dilatih menggunakan dataset COCO. Evaluasi deteksi objek menggunakan metrik mAP, sedangkan estimasi jarak menggunakan Root Mean Square Error (RMSE) dengan dataset iBims-1 sebagai acuan ground truth Depth. Hasil pengujian menunjukkan sistem mampu mendeteksi beberapa objek, yaitu car, bus, truck, dan person, serta memberikan estimasi jarak pada objek yang terdeteksi. Hasil evaluasi deteksi objek memperoleh nilai mAP sebesar 81,03%, sedangkan evaluasi estimasi jarak menggunakan ZoeDepth menghasilkan nilai RMSE sebesar 4,2248 meter. Berdasarkan hasil tersebut, kombinasi YOLOv10 dan ZoeDepth berhasil menghasilkan sistem yang mampu melakukan deteksi objek sekaligus memberikan informasi estimasi jarak berbasis video menggunakan kamera RGB. Sistem ini dapat menjadi salah satu pendekatan dalam pengembangan aplikasi computer vision yang membutuhkan informasi objek dan jarak secara bersamaan. Pendekatan ini dapat meningkatkan kelengkapan informasi visual untuk mendukung pemantauan lingkungan secara otomatis.

ABSTRACT

The development of artificial intelligence (Artificial Intelligence), particularly in the field of computer vision, has driven advances in image and video processing to automatically understand environmental conditions. One of its applications is object detection; however, detection methods generally only provide

information about the type and location of objects within bounding boxes without information about their distance from the camera. This limitation means that the information generated is insufficient to support systems requiring a more comprehensive understanding of the environment. Therefore, an approach that combines object detection and distance estimation using an RGB camera without additional hardware is required. This study aims to develop a video-based object detection and distance estimation system using a combination of YOLOv10 and ZoeDepth. YOLOv10 is used to detect objects and determine their positions in each video frame, while ZoeDepth is used to perform depth estimation to obtain the estimated distance of objects from the camera. The research methodology includes data collection, preprocessing, model implementation, and system evaluation. The data consist of 250 CCTV video samples, comprising 162 secondary data and 88 primary data. All data were used as testing data because YOLOv10 is a pre-trained model previously trained using the COCO dataset. Object detection performance was evaluated using the mAP metric, while distance estimation was evaluated using Root Mean Square Error (RMSE), with the iBims-1 dataset serving as the ground-truth depth reference. The test results show that the system is capable of detecting several objects, namely cars, buses, trucks, and persons, while also providing distance estimates for the detected objects. The object detection evaluation achieved an mAP value of 81.03%, while the distance estimation evaluation using ZoeDepth produced an RMSE value of 4.2248 meters. Based on these results, the combination of YOLOv10 and ZoeDepth successfully produces a system capable of performing object detection while simultaneously providing video-based distance estimation using an RGB camera. This system can serve as an approach for developing computer vision applications that require both object and distance information. This approach can improve the completeness of visual information to support automated environmental monitoring.

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan AI (*Artificial Intelligence*) telah mendorong kemajuan yang signifikan dalam bidang pengolahan citra dan video. Salah satu bidang yang berkembang pesat adalah *computer vision*, yang berfokus pada kemampuan sistem komputer dalam memahami dan menafsirkan kondisi lingkungan melalui data visual (Rohim, 2025). Deteksi objek merupakan salah satu penerapan utama dalam *computer vision* yang bertujuan untuk mengenali dan mengidentifikasi objek bergerak di lingkungan sekitar, seperti manusia, kendaraan, dan objek statis lainnya. Teknologi ini telah banyak diterapkan dalam berbagai aplikasi nyata, termasuk sistem pengawasan, robotika, kendaraan pintar, dan sistem bantuan navigasi. Penelitian terdahulu menunjukkan bahwa pendekatan berbasis *deep learning* mampu menghasilkan kinerja deteksi objek yang lebih akurat dan cepat dibandingkan metode tradisional yang masih bergantung pada fitur manual (Raisha Agma, 2025).

Algoritma YOLO (*You Only Look Once*) dikenal sebagai metode deteksi objek yang memiliki kecepatan tinggi dan mampu bekerja pada video karena proses deteksi dilakukan dalam satu tahap jaringan saraf (Nafis Alfarizi *et al.*, 2023). Karakteristik tersebut menjadikan YOLO sesuai untuk diterapkan pada sistem deteksi objek berbasis video karena mampu memberikan keseimbangan antara kecepatan proses deteksi dan akurasi hasil. Namun, YOLO tidak menyediakan informasi kedalaman atau jarak objek. Di sisi lain, ZoeDepth merupakan metode estimasi kedalaman berbasis single image *Depth Estimation* yang mampu menghasilkan peta kedalaman dari satu citra tanpa memerlukan sensor tambahan dan tanpa pelatihan ulang khusus. ZoeDepth dinilai fleksibel dan mampu bekerja pada berbagai kondisi lingkungan, meskipun tidak memiliki kemampuan untuk mengenali jenis objek secara langsung (Bhat *et al.*, 2023). Kemampuan tersebut menjadikan

1702 | Yuza Reswan, Erwin Dwika Putra, Yovi Apridiansyah, Dwi Bagas Prakoso; Moving Object Detection Using The Yolo Method ...

ZoeDepth sesuai untuk diintegrasikan dengan *YOLO* dalam menghasilkan informasi estimasi jarak menggunakan kamera *RGB* tanpa memerlukan perangkat keras tambahan.

Namun demikian, sebagian besar sistem deteksi objek yang dikembangkan saat ini masih terbatas pada kemampuan mengenali jenis objek dan menentukan posisi objek pada citra dua dimensi. Informasi jarak antara kamera dan objek belum menjadi fokus utama dalam banyak penelitian, padahal informasi tersebut memiliki peranan penting dalam sistem yang bekerja di lingkungan nyata. Pada aplikasi seperti navigasi robot, sistem keselamatan, dan penghindaran tabrakan, ketepatan estimasi jarak objek sangat memengaruhi proses pengambilan keputusan sistem. Beberapa penelitian menunjukkan bahwa sistem deteksi objek konvensional masih mengalami kesulitan dalam memperkirakan jarak objek secara akurat (Bachtiyar, 2024)(Ramadhani, Muttakin and Irsyad, 2026).

Berbagai pendekatan telah dikembangkan untuk mengatasi permasalahan estimasi jarak objek dalam sistem deteksi objek bergerak lingkungan sekitar.

Beberapa penelitian memanfaatkan sensor tambahan seperti *LiDAR* atau kamera *stereo* untuk memperoleh informasi kedalaman yang lebih presisi (Pinasthika and Utaminingrum, 2022). Meskipun pendekatan tersebut mampu meningkatkan akurasi, penggunaan sensor tambahan berdampak pada meningkatnya biaya dan kompleksitas sistem, sehingga kurang sesuai untuk penerapan pada sistem berbasis *real-time* dan perangkat berskala kecil. Pendekatan lain menggunakan metode estimasi kedalaman berbasis citra tunggal (*Monocular Depth Estimation*), namun metode ini masih menghadapi tantangan terkait akurasi, kestabilan hasil, serta kebutuhan komputasi yang tinggi(Katiyar *et al.*, 2024).

Berdasarkan permasalahan dan pendekatan yang telah dijelaskan, penelitian ini diharapkan mampu menggabungkan keunggulan metode *YOLO* dan *ZoeDepth* dalam membangun sistem deteksi objek bergerak di lingkungan sekitar dan estimasi jarak berbasis video. *YOLO* digunakan untuk mendeteksi serta menentukan lokasi objek, sedangkan *ZoeDepth* dimanfaatkan untuk memperoleh informasi kedalaman yang digunakan dalam proses perhitungan jarak objek dari kamera. Penelitian ini diharapkan dapat menghasilkan sistem yang lebih efektif, efisien, dan aplikatif, serta memberikan kontribusi dalam pengembangan sistem berbasis *computer vision* yang mampu bekerja pada perangkat dengan spesifikasi terbatas.

LANDASAN TEORI

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 antarframe 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).

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:

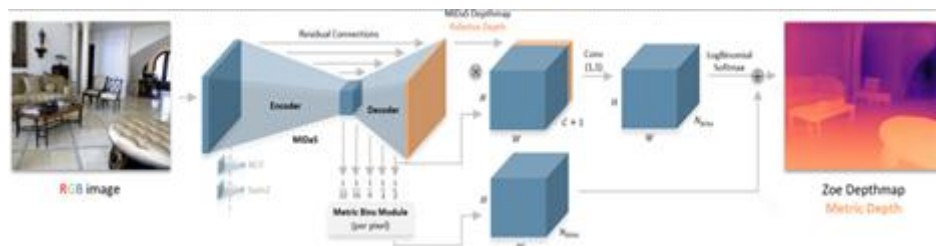
$$Z = D(x_c, y_c)$$

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).

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 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.

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).

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).

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).

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).

METODE PENELITIAN

Metode Penelitian

Penelitian ini dirancang untuk mengembangkan sistem deteksi objek dan estimasi jarak menggunakan kombinasi algoritma *YOLO* dan *ZoeDepth* berbasis video. *YOLOv10* digunakan untuk mendeteksi objek dan menentukan posisi pada setiap frame video, sedangkan *ZoeDepth* digunakan untuk melakukan estimasi kedalaman (Depth Estimation) guna memperoleh perkiraan jarak objek terhadap kamera. Metode penelitian meliputi pengumpulan data, prapemrosesan, implementasi model, dan evaluasi sistem. Data yang digunakan berupa 250 sampel video CCTV, terdiri dari 162 data sekunder dan 88 data primer. Seluruh data digunakan sebagai data pengujian karena *YOLOv10* merupakan model pre-trained yang telah dilatih menggunakan dataset COCO. Evaluasi deteksi objek menggunakan metrik mAP, sedangkan estimasi jarak menggunakan Root Mean Square Error (RMSE) dengan dataset iBims-1 sebagai acuan ground truth Depth.

HASIL DAN PEMBAHASAN

Hasil

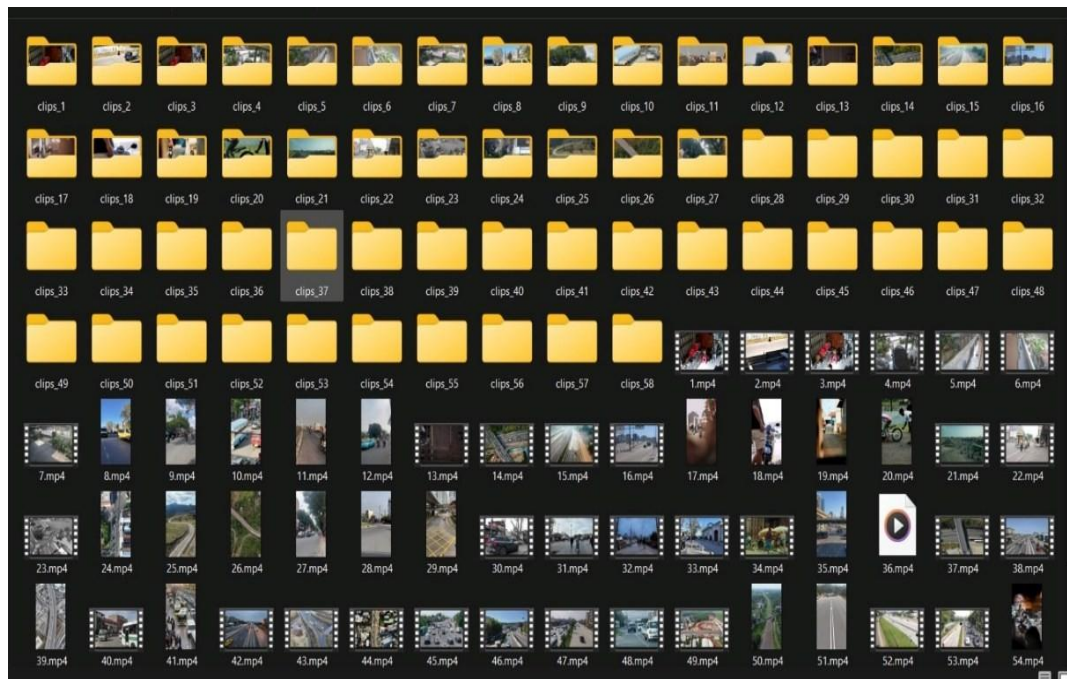
Hasil pengujian menunjukkan sistem mampu mendeteksi beberapa objek, yaitu car, bus, truck, dan person, serta memberikan estimasi jarak pada objek yang terdeteksi. Hasil evaluasi deteksi objek memperoleh nilai mAP sebesar 81,03%, sedangkan evaluasi estimasi jarak menggunakan *ZoeDepth* menghasilkan nilai RMSE sebesar 4,2248 meter. Berdasarkan hasil tersebut, kombinasi *YOLOv10* dan *ZoeDepth* berhasil menghasilkan sistem yang mampu melakukan deteksi objek sekaligus memberikan informasi estimasi jarak berbasis video menggunakan kamera RGB. Sistem ini dapat menjadi salah satu pendekatan dalam pengembangan aplikasi computer vision yang membutuhkan informasi objek dan jarak secara bersamaan. Pendekatan ini dapat meningkatkan kelengkapan informasi visual untuk mendukung pemantauan lingkungan secara otomatis.

Implementasi Sistem

Tahap awal penelitian dilakukan dengan mengumpulkan 250 sampel video CCTV yang terdiri atas 162 data sekunder dan 88 data primer. Selanjutnya, seluruh data yang telah melalui prapemrosesan digunakan sebagai data uji (testing data) untuk mengevaluasi kinerja sistem dalam mendeteksi objek dan mengestimasi jarak, sebagaimana ditunjukkan pada Gambar 4.1. Penggunaan 100% data sebagai data pengujian dilakukan karena model *YOLOv10* yang digunakan merupakan model *pre-trained* yang telah dilatih menggunakan dataset COCO val2017, sehingga penelitian ini tidak memerlukan proses pelatihan ulang (*re-training*) maupun pembagian dataset menjadi data latih dan data uji. Dengan demikian, seluruh data difokuskan untuk mengevaluasi performa sistem berdasarkan hasil deteksi objek dan estimasi jarak pada berbagai variasi objek, kondisi pencahayaan, dan lingkungan pengujian.

Tahap prapemrosesan data yang bertujuan untuk menyesuaikan format masukan sesuai dengan kebutuhan model *YOLO* dan *ZoeDepth*. Proses yang dilakukan meliputi penyesuaian ukuran *frame* video (*resize*), konversi ruang warna dari *BGR* ke *RGB*, serta konversi citra ke format *PIL Image*

sebagai masukan bagi *ZoeDepth*. Setelah itu, proses normalisasi dan konversi citra menjadi tensor dilakukan secara otomatis oleh *framework Ultralytics* dan *ZoeDe*.



Gambar 2. Sampel Video

Implementasi Dataset

Pada penelitian ini, COCO val2017 tidak digunakan untuk melakukan pelatihan ulang model, tetapi digunakan sebagai dataset yang menjadi dasar dari model pre-trained YOLOv10. Model YOLOv10 yang digunakan telah dilatih sebelumnya menggunakan dataset COCO, sehingga model dapat langsung digunakan pada tahap inference terhadap video CCTV yang digunakan dalam penelitian. Dengan pendekatan ini, penelitian tidak melakukan proses re-training maupun pembagian data menjadi data latih dan data uji untuk model YOLOv10.

COCO val2017 terdiri dari 5.000 citra dengan 36.781 objek yang berasal dari 80 kategori. Setiap objek memiliki anotasi berupa kelas dan bounding box. Informasi tersebut menjadi dasar kemampuan model YOLOv10 dalam mengenali objek berdasarkan karakteristik yang telah dipelajari pada saat proses pelatihan model. Dalam implementasi penelitian, model YOLOv10 digunakan untuk mendeteksi objek pada setiap frame video CCTV. Hasil deteksi berupa bounding box, kelas objek, dan nilai confidence kemudian digunakan sebagai hasil deteksi objek pada sistem. Berdasarkan hasil implementasi dan pengujian, objek yang dapat dideteksi dalam penelitian antara lain car, person, motorcycle, truck, dan bus.

Tabel 1 Daftar 80 kelas COCO (Lin, Tsung-Yi; Maire, Michael; Belongie, Serge; Hays, James; Perona, Pietro; Ramanan, Deva; Dollár, Piotr; Zitnick, no date)

No.	Kelas COCO
1	Person
2	Bicycle
3	Car
4	Motorcycle
5	Airplane
6	Bus
7	Train
8	Truck
9	Boat

10	Traffic light
11	Motorcycle
...	
80	toothbrush

Sementara itu, dataset iBims-1 digunakan pada tahap estimasi kedalaman untuk menyediakan nilai acuan (ground truth depth) yang diperlukan dalam pengujian ZoeDepth. Pada implementasinya, citra RGB dari iBims-1 diproses oleh model ZoeDepth untuk menghasilkan peta kedalaman (depth map). Nilai kedalaman hasil prediksi kemudian diambil pada posisi objek yang menjadi sampel pengujian dan dibandingkan dengan nilai ground truth depth pada posisi yang sama. Dari kedua nilai tersebut dihitung selisih atau error untuk mengetahui perbedaan antara hasil prediksi dan nilai sebenarnya.

Nilai error yang diperoleh dari setiap sampel selanjutnya digunakan dalam perhitungan Root Mean Square Error (RMSE). Dengan proses tersebut, iBims-1 berperan sebagai data acuan untuk mengukur seberapa besar kesalahan estimasi kedalaman yang dihasilkan oleh ZoeDepth. Dataset ini tidak digunakan untuk melatih atau melakukan fine-tuning model, tetapi digunakan pada tahap pengujian untuk memperoleh nilai pembandingan terhadap hasil prediksi ZoeDepth.

Tabel 2 Matriks Nilai Kedalaman pada *Ground Truth Depth Map iBims-1* (Munich, no date)

y \ x	0	1	2	3	4	...	639
0	D(0,0)	D(1,0)	D(2,0)	D(3,0)	D(4,0)	...	D(639,0)
1	D(0,1)	D(1,1)	D(2,1)	D(3,1)	D(4,1)	...	D(639,1)
2	D(0,2)	D(1,2)	D(2,2)	D(3,2)	D(4,2)	...	D(639,2)
3	D(0,3)	D(1,3)	D(2,3)	D(3,3)	D(4,3)	...	D(639,3)
4	D(0,4)	D(1,4)	D(2,4)	D(3,4)	D(4,4)	...	D(639,4)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
479	D(0,479)	D(1,479)	D(2,479)	D(3,479)	D(4,479)	...	D(639,479)

Tabel 2 menunjukkan representasi ground truth depth map pada dataset iBims-1 yang menggambarkan hubungan antara posisi piksel dan nilai kedalamannya. Sumbu x menunjukkan posisi piksel secara horizontal, sedangkan sumbu y menunjukkan posisi secara vertikal. Setiap elemen D(x,y) menunjukkan nilai kedalaman pada koordinat piksel (x,y). Pada resolusi 640 × 480 piksel, terdapat 640 piksel pada arah horizontal dan 480 piksel pada arah vertikal, sehingga satu depth map terdiri dari 307.200 posisi piksel. Setiap piksel yang valid memiliki nilai depth yang menunjukkan kedalaman terhadap kamera. Nilai tersebut diperoleh melalui pengukuran terrestrial laser scanner dan digunakan sebagai ground truth untuk mengevaluasi hasil estimasi kedalaman.

Implementasi model YOLO dan ZoeDepth

Sistem mengintegrasikan model *YOLO* dan *ZoeDepth* dalam satu program untuk melakukan deteksi objek dan estimasi kedalaman pada video *CCTV*. *YOLO* digunakan untuk mengenali objek yang terdapat pada setiap *frame* video, sedangkan *ZoeDepth* digunakan untuk memperkirakan kedalaman atau jarak objek terhadap kamera. Proses implementasi diawali dengan memuat kedua model agar siap digunakan pada proses pengolahan video.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan Berdasarkan hasil penelitian mengenai implementasi metode *YOLOv10* dan *ZoeDepth* untuk deteksi objek dan estimasi jarak berbasis video *CCTV*, maka diperoleh beberapa kesimpulan sebagai berikut.

1. Metode *YOLOv10* dan *ZoeDepth* berhasil diimplementasikan dalam satu sistem yang mampu mendeteksi objek sekaligus mengestimasi jarak objek pada video *CCTV*. Sistem dapat mengenali beberapa jenis objek, seperti *car*, *bus*, *truck*, dan *person*, kemudian menghitung estimasi jarak setiap objek yang berhasil dideteksi. Selain itu, sistem tetap dapat bekerja pada variasi jarak objek serta kondisi pencahayaan yang berbeda-beda selama proses pengujian.
2. Berdasarkan hasil evaluasi kinerja deteksi objek menggunakan *mAP* dengan hasil sebesar 81,03%. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam mendeteksi objek pada berbagai kondisi pengujian. Nilai *mAP* tersebut menunjukkan bahwa model mampu menghasilkan prediksi objek dengan tingkat ketepatan yang relatif tinggi berdasarkan kesesuaian antara hasil deteksi dan *ground truth*. Dengan demikian, model dapat dikatakan cukup efektif dalam mengenali dan mengklasifikasikan objek yang menjadi kelas pengujian.
3. Hasil evaluasi estimasi jarak menggunakan *Root Mean Square Error (RMSE)* terhadap 250 sampel data menghasilkan nilai sebesar 4,2248 meter. Nilai tersebut menunjukkan bahwa rata-rata kesalahan estimasi kedalaman model *ZoeDepth* terhadap *ground truth* adalah sekitar 4,22 meter. Dengan demikian, sistem telah mampu melakukan estimasi jarak objek, meskipun masih terdapat beberapa prediksi dengan kesalahan yang cukup besar sehingga memengaruhi nilai *RMSE* secara keseluruhan.

Saran

Untuk penelitian selanjutnya, disarankan agar dilakukan pengujian menggunakan *dataset* yang lebih banyak dan lebih beragam sehingga performa sistem dapat dievaluasi secara lebih komprehensif. Selain itu, penggunaan model deteksi objek maupun estimasi kedalaman yang lebih baru dapat dipertimbangkan untuk meningkatkan akurasi deteksi objek dan estimasi jarak. Terakhir, sistem dapat dikembangkan agar mampu bekerja secara *real-time* sehingga lebih siap diterapkan pada sistem pemantauan berbasis *CCTV*.

DAFTAR PUSTAKA

- Andres, J. *et al.* (2026) 'Advancing Offshore Safety : Monocular Depth Estimation from 360-Degree Images for Enhanced Oil Platform Inspection', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-3/W4-202(W4), pp. 71–77.
- Apridiansyah, Y. *et al.* (2025) 'Jurnal Processor Penerapan Metode *YOLOv5* untuk Klasifikasi dan Deteksi Objek Menggunakan Video Non-Real-Time', 20(2), pp. 147–153.
- Bachtiyar, M. Y. (2024) *Perancangan Sistem Deteksi Objek (Kubus) pada Robot Transporter Menggunakan Algoritma YOLOv8 (You Only Look Once)*.
- Bhat, S. F. *et al.* (2023) 'ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth'. Available at: <http://arxiv.org/abs/2302.12288>.

- Bini, F. and Manni, G. (2025) 'Deep Learning Framework for Facial Reconstruction Outcome Prediction : Integrating Image Inpainting and Depth Estimation for Computer-Assisted Surgical Planning', pp. 1–15.
- Diwan, T., Anirudh, G. and Tembhurne, J. V (2023) 'Object detection using YOLO : challenges , architectural successors , datasets and applications', *Multimedia Tools and Applications*. *Multimedia Tools and Applications*, 82(6), pp. 9243–9275. doi: 10.1007/s11042-022-13644-y.
- Driver, A. and Systems, A. (2022) 'SS symmetry Vehicle Distance Estimation from a Monocular Camera for Advanced Driver Assistance Systems', *Symmetry*, 14(12). doi: 10.3390/sym14122657.
- Faseeh, M. *et al.* (2024) 'Deep learning assisted real-time object recognition and depth estimation for enhancing emergency response in adaptive environment', *Results in Engineering*. Elsevier B.V., 24(September), p. 103482. doi: 10.1016/j.rineng.2024.103482.
- Hassani, I. B. *et al.* (2026) 'From YOLO V1 to YOLO V11 : comparative analysis of YOLO algorithm (review)', 16(1), pp. 450–462. doi: 10.11591/ijece.v16i1.pp450-462.
- Hidayat, F. A., Umbara, F. R. and Ilyas, R. (2025) 'Implementasi Yolo Untuk Menghitung Kepadatan Kendaraan Tempat Parkir', *Jurnal Algoritma*, 22(2), pp. 1997–2008. doi: 10.33364/algoritma/v.22-2.2919.
- Jalolov, T. S. (2023) 'Solving Complex Problems in Python', 01(09), pp. 481–484.
- Jiao L, Zhang R, Liu F, Yang S, Hou B, Li L, T. X. (2022) 'Generation Deep Learning for Video Object Detection: A Survey', *IEEE Trans Neural Netw Learn Syst*, pp. 3195–3215. doi: 10.1109/TNNLS.2021.3053249.
- Kar, P. *et al.* (2023) 'Revisiting Class Imbalance for End-to-end Semi-Supervised Object Detection', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE / Computer Vision Foundation (CVF), pp. 4570–4579. doi: 10.1109/CVPRW59228.2023.00481.
- Katiyar, R. *et al.* (2024) 'Object Detection and Depth Estimation Using Deep Learning', *Communications in Computer and Information Science*, 2053 CCIS, pp. 47–56. doi: 10.1007/978-3-031-56700-1_5.
- Kim, J. (2024) 'Comparison analysis of YOLOv10 and existing object detection model performance', *Journal of The Korea Society of Computer and Information*, 29(8), pp. 85–92. doi: 10.9708/jksci.2024.29.08.085.
- Liang, X., Wang, M. and Pan, P. (2025) 'Depth-aware Image-Space Fogging Algorithm for Background Privacy Protection using ZoeDepth Framework', *Journal of Computing and Electronic Information Management*, 16(1), pp. 87–95. doi: 10.54097/t37gnr69.
- Lin, Tsung-Yi; Maire, Michael; Belongie, Serge; Hays, James; Perona, Pietro; Ramanan, Deva; Dollár, Piotr; Zitnick, C. L. (no date) 'Microsoft COCO: Common Objects in Context', in, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

- Liu, X. *et al.* (2025) 'OPEN A simple monocular depth estimation network for balancing complexity and accuracy', pp. 1–15.
- Mohialden, Y. M. *et al.* (2024) 'Top Python-Based Deep Learning Packages: A Comprehensive Review', *International Journal Papier Advance and Scientific Review*, 5(1), pp. 1–9. doi: 10.47667/ijpasr.v5i1.283.
- Munich, T. U. of (no date) *iBims-1 (independent Benchmark images and matched scans – version 1)*. Available at: <https://www.asg.ed.tum.de/en/lmf/ibims1/>.
- Nafis Alfarizi, D. *et al.* (2023) 'Penggunaan Metode YOLO Pada Deteksi Objek: Sebuah Tinjauan Literatur Sistematis', *Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan*, 1(1), pp.54–63. Available at: <https://jurnalmahasiswa.com/index.php/aidanspk>.
- Niccoli, N. *et al.* (2025) 'Benchmark on Monocular Metric Depth', pp. 1–11.
- Pinasthika, K. and Utaminingrum, F. (2022) 'Monocular Depth Estimation pada Scene dalam Ruangan menggunakan U-Net dengan ResNet', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 6(11), pp. 5306–5312. Available at: <http://j-ptiik.ub.ac.id>.
- Puspadini, R. and Zen, M. (no date) *PYTHON : DASAR DAN PEMROGRAMAN*.
- Raisha Agma, A. (2025) 'Penerapan Computer Vision untuk Deteksi Objek dan Pengenalan Wajah', *Jurnal Komputer dan Teknik Informatika*, 01(01), pp. 31–37. Available at: <http://pustakajurnal.web.id/index.php/jkti>.
- Ramadhani, D., Muttakin, M. and Irsyad, H. H. (2026) 'Development and Optimization of YOLOv12 for Autonomous Vehicle Navigation Systems', 6(April), pp. 962–972.
- Ravpreet Kaur, S. S. (2023) 'A Comprehensive Review of Object Detection with Deep Learning', *Digital Signal Processing*, 132. doi: 10.1016/j.dsp.2022.103812.
- Rohim, D. A. (2025) 'Implementasi Artificial Intelligence dan Computer Vision dalam Transformasi Digital Nasional', 3(2), pp. 92–95.
- Simbolon, L. F. (2025) *Deteksi Kendaraan dan Estimasi Kecepatan untuk Monitoring Pelanggaran di Jalan Tol Berbasis Yolo dengan Notifikasi Real-Time*.
- Sunoto, T. D. *et al.* (2025) 'Model Deep Learning untuk Face Anti-Spoofing dalam Mengatasi Domain Generalization dengan Depth Estimation dan Generative Adversarial Network', 20(1), pp. 1–12. doi: 10.61769/telematika.v20i1.730.
- Takahashi, S. and Hayashi, E. (2026) 'Development of a Drone Navigation System Using Depth Estimation for Obstacle Avoidance', 11(3), pp. 204–209.
- Technology, C. of R. S. (no date) *iBims-1 – Chair of Remote Sensing Technology, Technical University of Munich (TUM)*. Available at: <https://www.asg.ed.tum.de/en/lmf/ibims1/>

- Tomasz, G. and Jakub, N. (2022) 'First Gradually , Then Suddenly : Understanding the Impact of', *Sensors (MDPI)*, pp. 1–19.
- Wang, C. and Liao, H. M. (2024) 'YOLOv1 to YOLOv10 : The fastest and most accurate real-time object detection systems', (August), pp. 1–13.
- Zulkarnaen, F. M. (2024) 'Sistem Deteksi Objek Manusia Menggunakan Algoritma YOLOv8 Berbasis Kamera Depth Sensor (Studi Kasus: Cv. Ateri Global Teknologi)', 32(3), pp. 167–186.
- Zulkiflie, M. A. (2021) 'Implementasi Algoritma Object Detection YOLOV4 dan Euclidean Distance Dalam Mendeteksi Pelanggaran Social Distancing'. Available at: https://repository.unhas.ac.id/id/eprint/12039/%0Ahttps://repository.unhas.ac.id/id/eprint/12039/4/H071171526_skripsi_01-11-2021.pdf.