UNIVERSITAS
GADJAH MADA
**Persepsi Multi-Modal Berbasis Deep Learning dengan Representasi Birdâ€™s-Eye View sebagai Dasar
Kendali Autonomous Excavator**
Muhammad Aditya Rahmansyah Baskoro, Ahmad Ataka Awwalur Rizqi, S.T., Ph.D.; Dr.Eng. Ir. Igi Ardiyanto, S.T.,M.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Ekskavator otonom yang menjalankan tugas *pick-and-place* membutuhkan estimasi posisi objek target dalam ruang 3D yang cukup presisi untuk menggerakkan bucket. Sebagian besar pendekatan otomasi ekskavator mengasumsikan posisi target sudah tersedia, sedangkan kamera tunggal maupun LiDAR sendiri memiliki keterbatasan masing-masing untuk menyediakan estimasi tersebut secara andal pada lingkungan konstruksi yang tidak terstruktur. Representasi bersama yang menyatukan informasi visual dan geometri menjadi kebutuhan mendasar sebelum posisi target dapat dipakai sebagai referensi kendali.

Penelitian ini merancang dan mengevaluasi sistem persepsi egosentris berbasis *Bird's-Eye View* (BEV) yang memfusikan enam kamera dan satu LiDAR untuk mendeteksi pusat 3D objek *rubble* dan *bin* pada lingkungan konstruksi tersimulasi. Sistem dibangun di atas ROS 2 dan Gazebo Fortress dengan variasi lingkungan terkendali. Arsitektur menggunakan EfficientNet-B0 dan *Lift-Splat-Shoot* pada cabang kamera, PointPillars pada cabang LiDAR, fusi gerbang adaptif, dan kepala deteksi berbasis CenterPoint. Evaluasi dilakukan melalui metrik deteksi 3D, metrik berbasis toleransi tugas, analisis kontribusi sensor, serta *replay* kinematika balik berbasis keluaran persepsi.

Pada split uji utama tanpa penyaringan kualitas tambahan, model fusi mencapai mAP 0,913, mATE 0,050 m, P90 ATE 0,084 m, dan R@0.25 sebesar 0,841. Dibandingkan LiDAR-only, fusi meningkatkan mAP dari 0,871 menjadi 0,913 dan AP *bin* dari 0,773 menjadi 0,844. Pada uji non-target mirip target, fusi menurunkan *false positive distractor* dari 77 menjadi 22. Hasil ini menunjukkan bahwa fusi membantu cakupan deteksi dan penolakan non-target, sedangkan ketepatan lokalisasi metrik tetap ditopang oleh bukti geometri LiDAR. Pada 74 scene yang memiliki pasangan *ground truth* terjangkau, *replay* IK berbasis model fusi menghasilkan pasangan target kompatibel-IK pada 75,7% scene (56/74), dibandingkan 67,6% (50/74) pada LiDAR-only. Peningkatan 8,1 poin persentase tersebut terutama berasal dari bertambahnya target peletakan bin yang layak. Penelitian ini menyimpulkan bahwa persepsi BEV multi-modal egosentris dapat menyediakan target yang relevan untuk kendali di simulasi ketika bukti geometri metrik tersedia, tetapi deteksi yang hanya berbasis visual belum otomatis layak digunakan sebagai target aksi langsung.

Kata Kunci: *Bird's-Eye View*, fusi kamera-LiDAR, deteksi objek 3D, ekskavator otonom, persepsi untuk kendali

UNIVERSITAS
GADJAH MADA
**Persepsi Multi-Modal Berbasis Deep Learning dengan Representasi Birdâ€™s-Eye View sebagai Dasar
Kendali Autonomous Excavator**
Muhammad Aditya Rahmansyah Baskoro, Ahmad Ataka Awwalur Rizqi, S.T., Ph.D.; Dr.Eng. Ir. Igi Ardiyanto, S.T.,M.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Autonomous excavators performing pick-and-place require 3D target-position estimates precise enough to command the bucket. Most excavator automation approaches assume the target position is already available, while a single camera or LiDAR alone each carries its own limitations in reliably providing that estimate in unstructured construction environments. A shared representation that unifies visual and geometric information becomes a basic requirement before a target position can be used as a control reference.

This work designs and evaluates an ego-centric Bird's-Eye View (BEV) perception system that fuses six cameras and one LiDAR to detect the 3D centers of rubble and bin objects in a simulated construction environment. The system is built on ROS 2 and Gazebo Fortress with controlled environment variation. The architecture uses EfficientNet-B0 and Lift-Splat-Shoot for the camera branch, PointPillars for the LiDAR branch, adaptive gated fusion, and a CenterPoint-based detection head. Evaluation is conducted through 3D detection metrics, task-tolerance metrics, sensor-contribution analysis, and inverse-kinematics replay using real perception outputs.

On the main held-out test_ungated split, the fused model reaches 0.913 mAP, 0.050 m mATE, 0.084 m P90 ATE, and 0.841 R@0.25. Compared with LiDAR-only, fusion improves mAP from 0.871 to 0.913 and bin AP from 0.773 to 0.844. On target-like non-target evaluation, fusion reduces distractor false positives from 77 to 22. These results show that fusion improves detection coverage and non-target rejection, while metric localization accuracy remains supported by LiDAR geometry. Among 74 held-out scenes containing a ground-truth reachable rubble–bin pair, inverse-kinematics replay produces an IK-compatible selected pair in 75.7% of scenes (56/74) with fused perception, compared with 67.6% (50/74) for LiDAR-only. The 8.1-percentage-point increase mainly results from additional usable bin/place targets. We conclude that ego-centric multi-modal BEV perception can provide control-relevant targets in simulation when metric geometric evidence is available, but visual detection without reliable metric support is not automatically suitable for direct action.

Keywords: *Bird's-Eye View, camera-LiDAR fusion, 3D object detection, autonomous excavator, perception for control*