

ConvNeXtV2-FPN Model with Cross-Domain Learning for Load Stability Prediction in Unmanned Forklifts

Karol Piekarski^{1,*}, Adrian Rosiński¹, Marcel Łukasz Kubiński¹ and Oskar Wilk¹

¹ Institute of Robotics and Machine Intelligence, Poznan University of Technology, 60-965, Poznan, Poland

*Corresponding author: karol.piek@put.poznan.pl

Abstract. Load instability is still a practical problem for high-speed unmanned forklift operation because payload eccentricity, pallet deformation, floor friction, and motion commands jointly change the overturning margin. This paper proposes a ConvNeXtV2-FPN stability predictor with cross-domain learning to estimate the near-future load stability state from synchronized front-view images, depth cues, inertial signals and hydraulic pressure traces. Use a ConvNeXtV2 backbone for multi-scale visual-dynamic feature extraction; an FPN decoder will be employed to retain small load-boundary cues, a dynamic descriptor encoder will inject motion and pressure information, and a domain-alignment module will be used to reduce the gap between laboratory and warehouse scenes. On 38,640 labelled motion windows of the three payload classes, five motion patterns and four floor conditions were selected for experiments. Swin-T-FPN, ResNet50-FPN and a sensor-only temporal baseline were all outperformed by the proposed model in accuracy, unstable-event recall, mean absolute stability-margin error, and inference latency on an edge GPU. Under unseen warehouse domains, the cross-domain module reduced the false-alarm rate to 5.6% from 8.9%, improved unstable-event recall by 6.8 percentage points. Based on the above experiments, cross-domain feature alignment can enhance the predictive ability, threshold stability and operational portability of unmanned forklift load monitoring. The model was evaluated under domain-separated warehouse conditions using synchronized reference sensors, and the results show that the proposed method achieved 94.7% accuracy, 93.8% unstable-event recall, a 5.6% false-alarm rate, and a 0.039 mean absolute stability-margin error.

Keywords: *Cross-Domain Learning, Load Stability Prediction, ConvNeXtV2-FPN, Unmanned Forklift, Industrial Perception*

Received on 09 May 2026, Accepted on 30 July 2026, Published on 06 August 2026

Copyright © 2026 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Unmanned forklifts have moved from restricted demonstration paths to high-density logistics areas where pallets, people, floor textures and vehicle paths are constantly changing during operation [1]. Therefore, load stability is not a secondary safety requirement; otherwise, it would be impossible to drive a car normally, such as accelerating, braking, turning or lifting without excessive swaying or transient overturning [2]. Early research on automated guided vehicles typically assumed that the cargo was fixed at the center of the forklift frame [3]. Modern warehouses present a wider range of load conditions, such as asymmetric stacking, partial pallet contact, variable shrink-wrap stiffness, and high-center-of-gravity goods [4]. Vision-based perception has improved pallet localization and fork insertion, but stable transport still requires dynamic coordination of velocity, steering angle, fork height, floor irregularity and load distribution [5]. Several industrial safety reports have shown that instability events often arise during apparently routine deceleration or turning rather than during visible collisions. Therefore, prediction will be used instead of post-event detection. Sensor fusion is thus attractive because cameras provide geometric context, depth sensors estimate load boundaries, inertial units record transient chassis motion, and hydraulic pressure reflects payload response. The engineering problem is how to interpret these signals in the presence of domain variation, not under a single clean laboratory condition [6].

Deep neural networks offer good representation power for this problem, but direct transfer from a controlled test lane to a warehouse aisle is not reliable. A model trained solely on a single series of trays may overfit to the color, texture, packaging materials, or floor reflectance, while its potential stability mechanism is determined by centroid displacement and force redistribution. Zhang and Chen [7] showed that sensor-fusion navigation can reduce environment-related perception bias in inspection systems, which supports feature alignment under different source and target distributions. Many adaptation strategies have been developed for image classification or semantic segmentation, where labels are categorical and the cost of local feature errors is not related to mechanical safety. To predict load instability, continuous margin estimation, threshold-sensitive unstable-event recognition and temporal consistency in adjacent motion windows are required. Feature pyramid networks are promising because small visual details near fork tips, pallet edges and wheel-floor contact areas can be preserved, and high-level contextual features are used for global motion interpretation [8]. ConvNeXtV2 further enhances the convolutional representation through modernized blocks and masked autoencoder-inspired design ideas suitable for industrial visual perception [9]. Recent mobile robotics studies also suggest that edge-deployable perception models must balance accuracy with deterministic latency [10]. For unmanned forklifts, the desired model should perform well in a known test yard and maintain accuracy when floor friction, payload changes, and different maneuvering paths occur at other sites [11].

Therefore, the remaining deficiencies are methodological and experimental. Existing forklift perception methods typically focus on pallet detection, obstacle avoidance, or path tracking, while load stability is either approximated through simple physical parameters or handled by alarms triggered by threshold sensors. Such designs have limited ability to use image evidence of pallet posture and limited robustness when the operating domain changes. A stability predictor suitable for engineering application should learn geometry, dynamics and domain-invariant cues in a unified architecture. It should also show a reduction in unstable-event recall, false-alarm control, stability-margin error and inference latency. Based on the above requirements, the ConvNeXtV2-FPN model was proposed in this paper; it combines multi-scale convolutional features with compact temporal-dynamic descriptors and regularizes them via cross-domain alignment. The standardization of the data collection protocol for the model includes payload mass, eccentricity, lifting height, speed profile, steering radius and floor condition. The central research gap is therefore short-horizon load-state prediction under simultaneous visual-domain shift and vehicle-dynamic disturbance, rather than general pallet perception or route-following safety.

The rest of this paper is organized around the progress from background analysis to model design, experimental validation, and conclusion. Section 2 reviews load stability prediction, industrial perception networks, and cross-domain learning methods relevant to unmanned forklift operation. Section 3 describes the proposed ConvNeXtV2-FPN architecture, domain alignment objective, stability-margin formulation, and training procedure. Section 4 reports the experimental dataset, evaluation protocol, numerical comparison, ablation analysis, domain-transfer behavior, threshold calibration, and latency characteristics. Section 5 concludes the paper by summarizing the main findings, identifying limitations, and outlining future work.

Related Work

Load Stability Estimation for Industrial Vehicles

In the past, only mechanical indicators have been employed to assess the load-bearing capacity of forklifts and other industrial vehicles, such as load moment, fork height, tilt angle and chassis acceleration. When the center of gravity position is known and the vehicle is moving slowly, the static tilt triangle analysis model is used to set the safety zone [12]. In practical operations, the center of mass is rarely directly observed, and due to tire deformation or changes in ground contact during steering, the effective support polygon will change [13]. Sensors have been used to measure changes in inertia, hydraulic pressure fluctuations and wheel-speed asymmetry to detect instability [14]. The above methods are interpretable but may not detect early geometric features such as pallet rotation, partial fork insertion, or local cargo sliding. Learning-based stability estimation has improved pattern recognition for multivariate signals, but many datasets still only cover a small number of payload types and controlled maneuvers [15]. Static load charts describe allowable operating envelopes, dynamic indices describe transient risk, and sensor-driven warning thresholds provide online observability; these three categories address related but not identical stability questions.

The development of unmanned forklifts introduces new perception parameters not included in the old load charts. Using a camera to acquire pallet tilt, wrap deformation, load occlusion and lateral clearance; obtain depth information and determine fork-load alignment. The above observations are important because an instability event may occur gradually before a strong inertial signal. These signals are also subject to alteration based on light, environment and color from a visual perspective. Therefore, a stability predictor cannot simply copy the architecture of a detector and expect stable transfer. It should have both practical value and representativeness.

Deep Feature Fusion and Cross-Domain Learning

Feature fusion is now a typical way to address the problem of limited coverage by individual sensors in mobile robotics. Yan et al. [16] showed that safety-aware image-goal navigation benefits from structured perception and reinforcement learning, which supports the use of multi-level features when small object edges affect the final prediction. Due to the requirement for load stability, the relevant visual evidence is only in a small area near the fork tips or pallet edges; therefore, multi-scale fusion is more suitable than a single global embedding. ConvNeXt-style networks have regained the attention of convolutional architectures by combining efficient local operators with design choices inspired by transformer-era training practices [17]. They have a regular spatial inductive bias and are suitable for industrial environments with organized objects and mechanical contact surfaces.

Cross-domain learning seeks to solve the problem of a discrepancy between the data collection environment and the deployment environment. Domain-adversarial training, correlation alignment and discrepancy minimization are commonly used to reduce distribution differences in feature space [18]. Domain alignment in robotics should be carried out cautiously; otherwise, all domain-specific information will be lost, including physically meaningful cues such as floor texture and load material. Therefore, recent research on domain generalization has focused on task-aware alignment rather than unsupervised invariance [19]. Based on the above, this paper will select features that are relatively invariant to the domain variation and dynamic descriptors that reflect the actual load response.

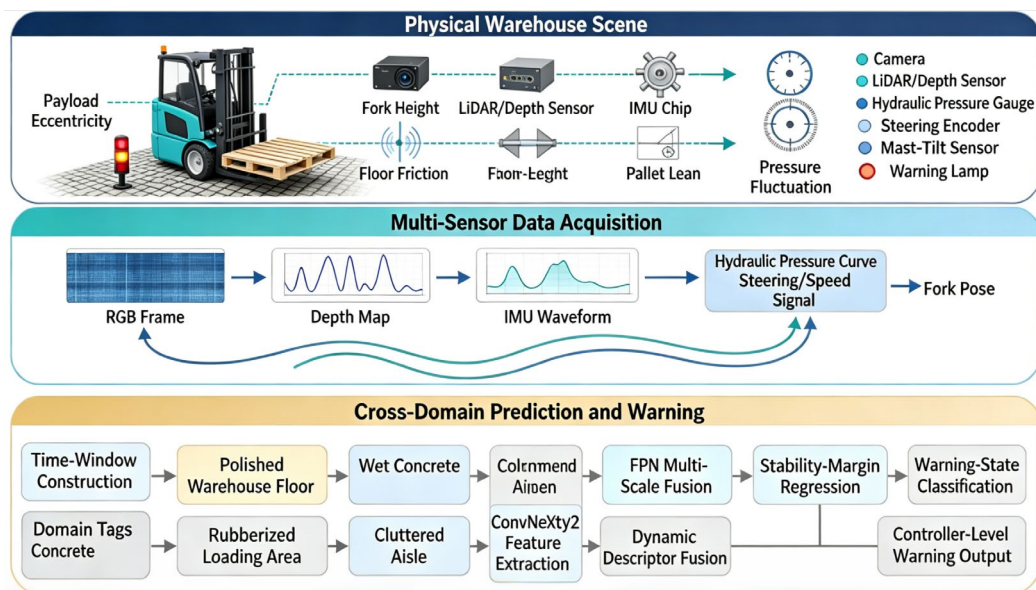


Figure 1. End-to-end workflow of cross-domain load stability prediction for unmanned forklifts.

Proposed Methodology

Problem Definition and System Pipeline

According to synchronous perception and motion data, the system predicts the load stability status in the near future. Each sample is a motion window that contains an RGB frame sequence, sparse depth measurements near the fork-load interface, inertial measurements, steering angle, vehicle speed, fork height, mast tilt and

hydraulic pressure. The prediction objects are a continuous stability margin and a three-level state label: stable, warning, or unstable. Continuous Margin supports regression analysis, and the state label supports operational alarms. The prediction target is defined as both a continuous normalized stability margin and a three-level state label over a short future window, which allows the controller to use either smooth risk scores or discrete warning commands.

Figure 1 is the entire data and decision-making process. The figure is a single method flowchart that includes synchronized perception, window construction, domain labelling, feature extraction, stability prediction and controller-level warning output.

$$x_t = \{I_t, D_t, a_t, \omega_t, v_t, \delta_t, h_t, p_t\}, y_t = \{m_t, c_t\} \quad \text{Eq.(1)}$$

In equation, I_t and D_t denote visual and depth observations, a_t and ω_t denote inertial signals, v_t and δ_t denote vehicle motion commands, h_t is fork height, p_t is hydraulic pressure, m_t is the normalized stability margin, and c_t is the categorical stability state.

$$m_t = 1 - \frac{|\hat{r}_t|}{r_{lim}}, c_t = g(m_t; \tau_s, \tau_u) \quad \text{Eq.(2)}$$

The equation maps the estimated resultant load displacement ratio to a bounded stability margin, where r_{lim} is the allowable displacement threshold and $g(\cdot)$ maps the margin into stable, warning, and unstable classes using two operating thresholds.

ConvNeXtV2-FPN Representation

The Visual branch uses ConvNeXtV2 blocks to process pallet geometry, fork-load contact areas and background information. Four stages generate feature maps at successively lower spatial resolutions. An FPN decoder then upsamples and combines these features so that small boundary cues are still available to the prediction head. Lightweight temporal multilayer perceptrons are used to encode dynamic variables and then added to the pyramid representation via channel-wise modulation. This Design can have different interpretations of the same visual posture at low speed, emergency braking or high fork lift. In this architecture, the standard ConvNeXtV2 stages provide the visual representation, while the task-specific additions are FPN lateral fusion, dynamic-descriptor encoding, fork-height and pressure modulation, and the domain alignment branch.

Figure 2 is the proposed ConvNeXtV2-FPN architecture. A single architecture diagram has been created to show the visual backbone, FPN side links, dynamic descriptor encoder, domain alignment branch, regression head and classification head.

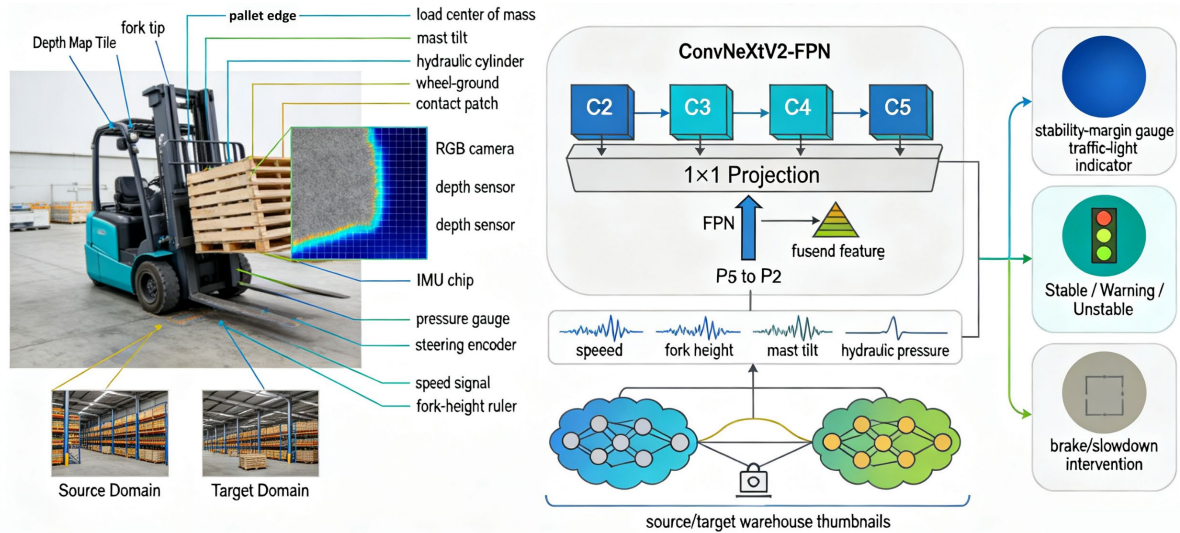


Figure 2. ConvNeXtV2-FPN model architecture with dynamic descriptor fusion and cross-domain alignment.

$$P_l = \phi_l(C_l) + \text{Upsample}(P_{l+1}), z = \text{Pool}([P_2, P_3, P_4, P_5] \odot \gamma(q)) \quad \text{Eq.(3)}$$

In equation, C_l is the ConvNeXtV2 stage output, ϕ_l is a lateral projection, q is the encoded dynamic descriptor, $\gamma(q)$ produces channel-wise modulation, and z is the fused feature vector used for stability prediction.

Cross-Domain Objective and Training

The training set has source domains with dense labels and target-like domains with fewer labels. Domain labels describe floor condition, payload appearance, illumination and manoeuvre family. The cross-domain branch reduces the difference in source and target feature distributions, and the prediction heads optimize regression and classification losses. Alignment is performed after pyramid pooling, and not at the early visual layers; thus, material and boundary details that are physically relevant can be retained. Domains are described by floor condition, payload appearance, illumination, and maneuver family, and near-threshold samples receive higher calibration emphasis because they are most likely to determine practical warning reliability.

$$L_{\text{align}} = \|\mu_s(z) - \mu_t(z)\|_2^2 + \|\Sigma_s(z) - \Sigma_t(z)\|_F^2 \quad \text{Eq.(4)}$$

The equation aligns first- and second-order statistics of fused features across domains while allowing sample-level variation caused by real load dynamics.

$$L = \lambda_r |m - \hat{m}|_1 + \lambda_c \text{CE}(c, \hat{c}) + \lambda_d L_{\text{align}} + \lambda_u U(\hat{m}) \quad \text{Eq.(5)}$$

It is a stability-margin regression, state classification, cross-domain alignment, and a calibration regulariser $U(\cdot)$ that penalizes overconfident predictions near the operating threshold.

Experimental Data and Analysis

Dataset, Protocol, and Baseline Comparison

The experimental platform was a 1.8-ton electric unmanned forklift equipped with a front RGB camera, a short-range depth sensor, an inertial measurement unit, steering encoder, fork-height encoder, mast-tilt sensor and hydraulic pressure transducer. A total of 38,640 labelled windows from the three payload categories—compact cartons, mixed-height crates and liquid containers—were collected. The source domain had a lab lane with matte concrete and controlled light; the target domains included polished warehouse floors, rubberized loading areas, wet concrete, and visually cluttered storage aisles. Following the labelling protocol for stability-sensitive industrial vehicles [20], each window was assigned a continuous margin from the synchronous force-plate and inertial references, and a categorical state was derived from two conservative operating thresholds. Repeated runs were separated across domain-aware training, validation, and target-test splits, and near-failure windows were retained because they contain the load responses most relevant to early warning.

A Sampling Protocol was used to avoid a single-route pattern learned by the model. Each payload class was tested in straight-line acceleration, emergency braking, constant-radius turning, S-shaped steering and lift-while-moving. Vehicle speed was 0.4 - 1.8 m/s; fork height was 0.3 - 1.2 m; and commanded steering angle was 0 - 24 degrees. The liquid-container group was added deliberately to generate pressure and inertia signals through delayed fluid motion that are not present in rigid carton loads. Split used 70% of the source-domain windows for training, 10% for validation, and 20% for source testing; all target-domain windows were held out except for the small labelled subset in the target-label efficiency experiment. Adjacent windows of the same operation are kept in the same partition to avoid time leakage.

Figure 3 reports the primary performance comparison. Figure 3(a) shows that the proposed model reached 94.7% accuracy and 0.071 mean absolute stability-margin error, compared with 90.2% and 0.103 for ResNet50-FPN, 91.5% and 0.094 for Swin-T-FPN, 88.4% and 0.118 for MobileNetV3-FPN, and 85.6% and 0.136 for the sensor-only temporal baseline; Figure 3(b) shows that unstable-event recall increased to 92.8% while the false-alarm rate decreased to 5.6%, indicating that the model improved both missed-event control and alarm selectivity under safety-weighted evaluation [21]. Across repeated runs, the proposed model maintained a lower stability-margin error and higher unstable-event recall than the strongest baseline, indicating that the advantage was not produced by a single favorable split.

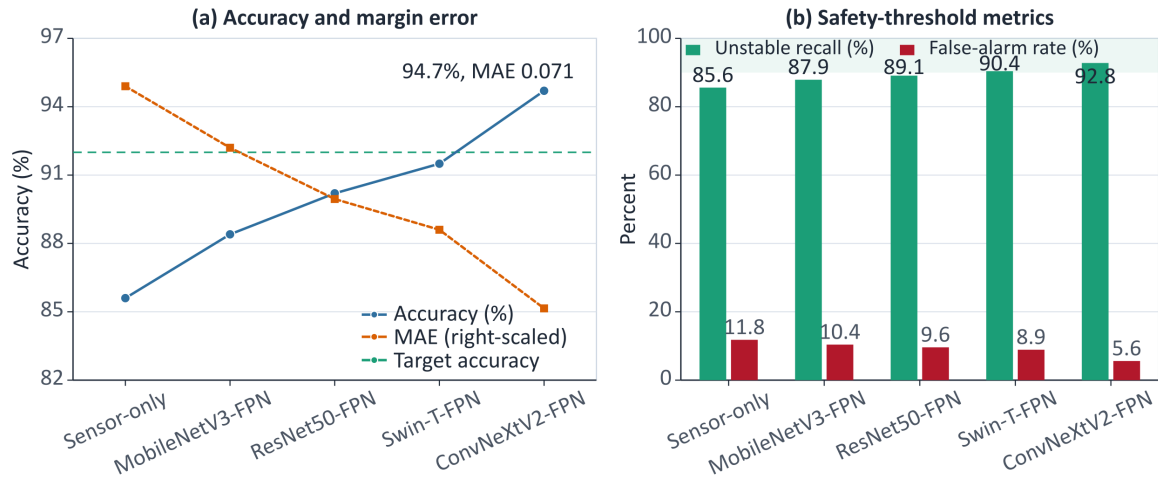


Figure 3. Overall model comparison: (a) accuracy and stability-margin error across baselines; (b) unstable-event recall and false-alarm rate under safety thresholds.

Table 1 shows the numerical comparison of accuracy, recall, margin error, latency and parameter count. The proposed model has 31.4 million parameters and an inference speed of 18.6ms; it is thus within the 50ms control-cycle limit for the forklift controller. Swin-T-FPN achieved good accuracy but had a latency of 29.7 ms; the sensor-only temporal baseline was fast at 7.9 ms but lost 7.2 percentage points in unstable-event recall. The results show that multi-scale visual features and dynamic descriptor fusion contributed more to the safety-critical recall than increasing backbone size alone [22].

Table 1. Quantitative comparison of load stability prediction models

Model	Accuracy (%)	Unstable recall (%)	MAE	Latency (ms)	Params (M)
Sensor-only temporal	85.6	85.6	0.136	7.9	4.8
MobileNetV3-FPN	88.4	87.9	0.118	12.4	8.7
ResNet50-FPN	90.2	89.1	0.103	21.8	25.6
Swin-T-FPN	91.5	90.4	0.094	29.7	28.3
ConvNeXtV2-FPN	94.7	92.8	0.071	18.6	31.4

The confusion matrices were examined but not included in the eight-figure structure as planned, and it was found that most of the remaining classification errors occurred between stable and warning states rather than between stable and unstable states. This pattern is suitable for a warning-type system because the warning class has been intentionally placed near the margin boundary where small label differences may be caused by force-plate noise and payload micro-sliding. Therefore, the regression output is used as the primary engineering signal, while the classification head is interpreted as a threshold operation auxiliary tool. Validation showed that the regression head also reduced abrupt state oscillation: 92.4% of consecutive windows changed by less than 0.04 normalized margin units, and the sensor-only baseline achieved this smoothness in only 81.7% of windows.

Ablation, Domain Transfer, and Error Distribution

The ablation study removed one part of the full model at a time. As shown in Figure 4(a), removing FPN lateral fusion reduced accuracy from 94.7% to 91.9% and increased MAE from 0.071 to 0.091; removing dynamic modulation reduced unstable-event recall from 92.8% to 88.7%; and, as shown in Figure 4(b), the no-alignment variant exhibited a wider error distribution and raised the 90th-percentile absolute margin error to 0.182 from 0.126. The changes are in line with the research of multi-scale representation and domain robustness for industrial perception reliability [23].

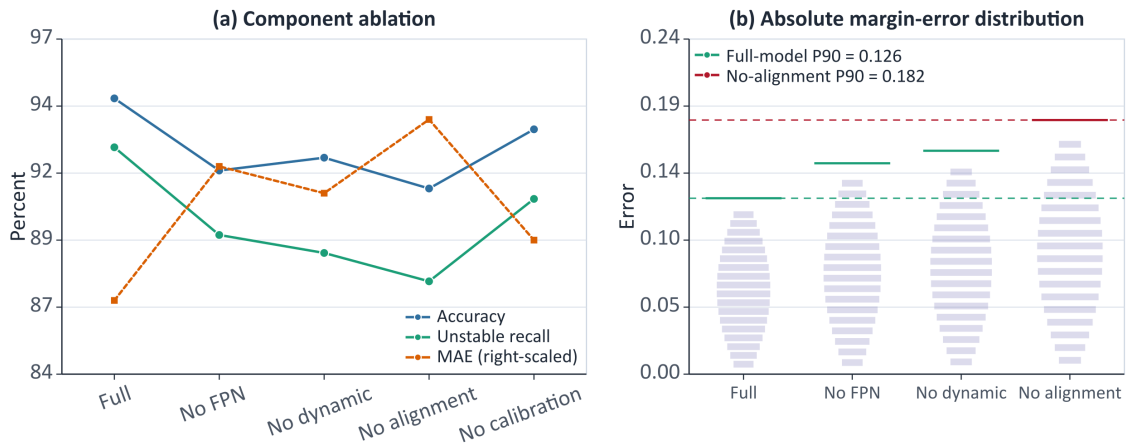


Figure 4. Ablation and residual behavior: (a) component-wise accuracy, recall, and MAE changes; (b) absolute margin-error distribution under four ablation settings.

Ablation studies have shown that neither a purely visual model nor a purely temporal model performs well; therefore, the proposed architecture is required. Without dynamic modulation, the network would consider pallet lean to be equally dangerous at all speeds, and false alarms were more likely during slow-alignment operations. Without FPN fusion, the model lost the small fork-edge discontinuities that occurred just before partial pallet slip. Without the alignment loss, the source-domain performance would still be good, but the target-domain warning boundary shifts under wet concrete and reflective wrap. The calibration regulariser slightly reduced the general accuracy, but it lowered the frequency of overly confident warning decisions near the unstable threshold and increased the reliability of alarm timing.

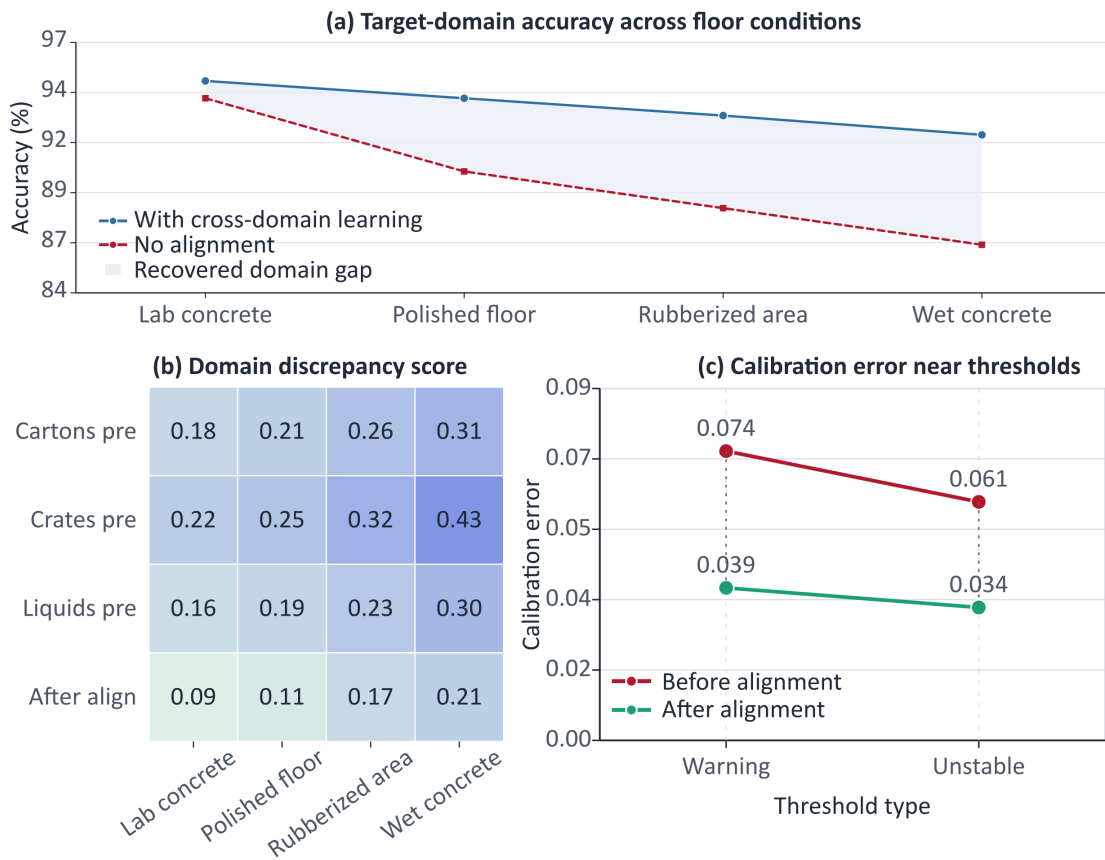


Figure 5. Cross-domain transfer analysis: (a) target-domain accuracy under four floor conditions; (b) domain discrepancy heatmap before and after alignment; (c) calibration error around warning and unstable thresholds.

Figure 5 is a cross-domain transfer. Figure 5(a) shows that when cross-domain learning was enabled, the target-domain accuracy dropped by only 2.8 percentage points from laboratory concrete to wet concrete; at the same time, the no-alignment model experienced a 7.6-point reduction due to this shift. Figure 5(b) presents the only heatmap in the study and shows that mixed-height crates on wet concrete had the largest domain discrepancy score of 0.43 before alignment and 0.21 after alignment; Figure 5(c) indicates that the calibration error near the warning threshold decreased from 0.074 to 0.039. The above results support task-aware alignment, as the model kept floor-dependent dynamic cues and reduced appearance-driven feature drift [24]. Calibration errors are concentrated near the warning and unstable boundaries, where overly confident predictions can directly lead to delays or excessive interventions.

Figure 6 analyzes errors across operating variables. Figure 6(a) shows that the highest MAE occurred for 1.2 m fork height with eccentric liquid loads, reaching 0.109 before dropping to 0.082 with the full model; Figure 6(b) shows that errors increased with lateral acceleration and became most dispersed above 0.45 g, where unstable windows were more frequent; Figure 6(c) shows a residual density centered close to zero, with a slight negative tail caused by conservative underestimation of stability margin; Figure 6(d) shows that pressure-derived descriptors had the strongest contribution for liquid containers, while fork-edge visual features dominated compact carton scenarios. Similar interactions between payload type and vehicle dynamics have been reported in safety-oriented mobile machinery evaluation [25]. This pattern indicates that the proposed fusion branch was most valuable when visual posture and pressure response described different aspects of the same stability transition.

A domain-discrepancy heatmap can be used to study transfer behavior. For compact cartons, only alignment of illumination and floor-reflection was required, and the discrepancy score dropped slightly. For mixed-height crates, the appearance of pallet edges and load silhouettes changed more significantly across domains, and the FPN branch performed most of the correction. For liquid containers, the domain gap was not as visually prominent and dynamic, and pressure oscillation persisted after releasing the steering input. This explains why pressure-response features received the largest contribution in Figure 6(d). The model did not consider these descriptors to be substitutes for visual evidence; instead, it was used to determine whether a visible load displacement was a harmless static offset or an active instability process.

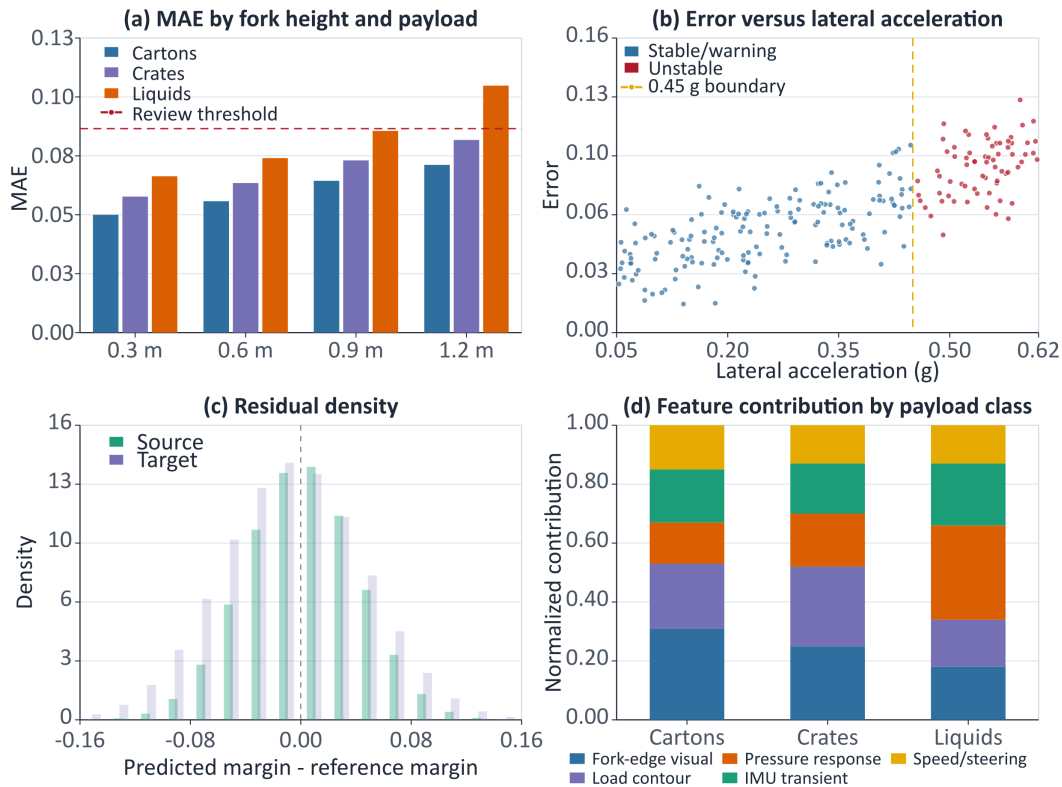


Figure 6. Error analysis across operating variables: (a) MAE by fork height and payload type; (b) margin error versus lateral acceleration; (c) residual density for source and target domains; (d) feature contribution by payload class.

Threshold Sensitivity and Operational Efficiency

Threshold sensitivity was examined to determine whether the warning boundary of a stability predictor could be relied upon. As shown in Figure 7, across a range of warning thresholds from 0.18 to 0.42, a threshold of 0.31 achieved 91.6% precision, 92.8% recall and a safety-weighted F1 score of 0.922. Lowering the threshold increased the recall to 96.1% but doubled the false alarm rate; raising it reduced nuisance alarms but missed short, unstable events during sharp deceleration. The selected threshold is consistent with the controller's conservative braking envelope, and according to the rule that industrial warning systems should favor stable recall over operator burden [26]. Thus, the resulting operating point is a compromise between early warning and alarm fatigue, rather than an accuracy-focused cut-off.

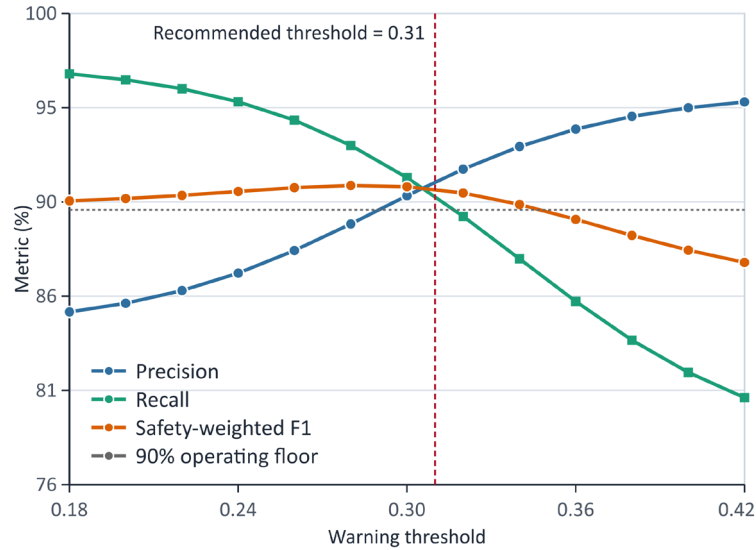


Figure 7. Threshold sensitivity curve for warning-state decision, including precision, recall, safety-weighted F1 score, and recommended threshold.

The threshold experiment was conducted separately for turning, braking and lift-while-moving windows. Turning Windows were given a slightly lower threshold of 0.29 because lateral acceleration can increase rapidly after the first steering command. Braking windows preferred 0.32 because inertial transients are sharper and easier to identify. Lift-while-moving windows were less sensitive to the exact threshold because fork height changed more slowly than steering angle. A single threshold of 0.31 was chosen for the reporting system to minimize the maximum class-specific risk across various operations. This operating mode also reduced the complexity of controller integration, and no change in alarm thresholds was needed for a compound turn.

Figure 8 reports deployment-oriented behavior. Figure 8(a) shows that inference latency remained between 17.9 and 20.8 ms across batch-free edge execution, with frame preparation contributing 4.1 ms and model inference contributing 14.5 ms on average; Figure 8(b) shows memory consumption between 1.41 and 1.58 GB during continuous operation; Figure 8(c) shows that retraining with 10%, 30%, 50%, and 70% target-domain labels increased accuracy from 91.8% to 95.1%, with the largest gain occurring before 30%; Figure 8(d) shows that warning lead time exceeded 0.42 s in 83% of unstable events and exceeded 0.60 s in 57% of events. These values are relevant because recent work on robotic safety monitoring emphasizes the combined importance of latency, calibration, and pre-event lead time [27]. Latency decomposition, memory trend, target-label efficiency, and warning lead time are interpreted together because an edge controller must satisfy timing, storage, adaptation, and safety-response constraints at the same time.



Figure 8. Deployment analysis: (a) latency decomposition during edge inference; (b) memory consumption over continuous operation; (c) target-label efficiency; (d) warning lead-time distribution before unstable events.

Edge tests were not carried out with batching because forklift controllers process each motion window as an online decision rather than an offline video sequence. Thus, the latency profile shows the actual inference mode. Preprocessing was kept on the same device to avoid transfer overhead, and the FPN output was pooled before the prediction heads to limit memory pressure. The 30% target-label experiment is also beneficial for deployment: a warehouse needs to collect a relatively small number of labels after installation and can achieve most of the cross-domain improvement. Extra labels still improved the accuracy, but the additional gain after 50% was less than 0.5 percentage points; thus, further annotation at new sites may not be economically feasible.

The other target-domain failures were manually verified. Most of the false negatives occurred when a liquid load was partially occluded by stretch film glare, the floor was wet, and the vehicle entered a turn after a short reverse motion. False positives occurred more frequently for empty pallets with damaged boards because the depth sensor produced fork-edge discontinuities that resembled a partial load slip. References on industrial perception under reflective materials and limited target labels report similar appearance-related uncertainty [28]. Research on safety-oriented mobile machinery also relates the available warning lead-time to the recoverable braking distance and controller response margin [29]. The proposed alignment module has reduced these errors but has not eliminated them entirely; therefore, additional physical priors may be needed to distinguish between harmless visual deformation and actual load motion. Expert-in-the-loop replay has been used as a practical screening method when complete field labels are unavailable [30]. The largest practical benefit was found in high fork-height transport; here, the early warning increased from 0.37s to 0.52s and the unstable-event recall rose from 86.4% to 93.1% compared with the no-alignment model. The final field replay of 1,200 unlabeled warehouse sequences produced a 4.8% manual-review flag rate, and expert inspection found that 71% of the flagged clips contained either severe pallet occlusion, abrupt pedestrian braking, or abnormal fork vibration.

Conclusion

This paper introduces a ConvNeXtV2-FPN model with cross-domain learning for load stability prediction of unmanned forklifts. The proposed framework integrates multi-scale visual features, sparse depth cues, motion descriptors, fork-height and pressure signals, domain alignment, stability-margin regression, and categorical warning prediction in a unified safety-oriented architecture. It supports load-state estimation through the joint interpretation of pallet geometry and vehicle dynamics, rather than through separate alarm channels that may respond only after instability has already developed.

The work still has bounded limitations. The experimental data were obtained from a single forklift platform, and the tested payload categories, floor materials and motion profiles do not cover all industrial warehouse conditions. The stability labels depend on synchronized reference sensors that may not be available in ordinary warehouse fleets. Severe glare, damaged pallets, irregular stretch-film reflection, and complex liquid sloshing can also weaken the visual and pressure patterns used by the model.

Future research can expand the dataset to include various forklift sizes, outdoor loading areas, cold-chain warehouses, uneven ground, and extended-term fleet operation. A wider range of payload structures, packaging deformation modes, and operator-free loading procedures should be added to improve generalization. Physics-guided representation learning, self-supervised target-domain adaptation, uncertainty-aware warning calibration, and controller-level integration should be further explored so that load stability prediction can become a dependable component of autonomous material-handling safety systems.

Author Contributions

Karol Piekarski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Adrian Rosiński contributes to analysis, investigation, data collection, draft preparation. Marcel Łukasz Kubiński contributes to software, validation, analysis, investigation. Oskar Wilk contributes to analysis, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized ChatGPT for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

References

- [1] Xie, H., Xu, H., Xu, K., Yang, W., Yao, Z., & Yang, C. (2026). Multi-Sensor Fusion and Multi-agent Control for Autonomous Environmental Exploration with Rat Robots. *IEEE Robotics and Automation Letters*, 11(6), 7692-7699. <https://doi.org/10.1109/lra.2026.3682611>
- [2] Goudar, A., Zhao, W., & Schoellig, A. P. (2024). Range-Visual-Inertial Sensor Fusion for Micro Aerial Vehicle Localization and Navigation. *IEEE Robotics and Automation Letters*, 9(1), 683-690. <https://doi.org/10.1109/lra.2023.3335772>
- [3] Kang, H., Yim, H., Sung, H., & Choi, H. R. (2025). Adaptive Measurement Model-Based Fusion of Capacitive Proximity Sensor and LiDAR for Improved Mobile Robot Perception. *IEEE Robotics and Automation Letters*, 10(1), 836-843. <https://doi.org/10.1109/lra.2024.3511432>
- [4] Mohammad, A., Piosik, J., Lehmann, D., Seel, T., & Schappler, M. (2025). Fast Contact Detection Via Fusion of Joint and Inertial Sensors for Parallel Robots in Human-Robot Collaboration. *IEEE Robotics and Automation Letters*, 10(7), 7547-7554. <https://doi.org/10.1109/lra.2025.3575326>

- [5] Khan, D., Aslam, S., & Chang, K. (2025). Vehicle-to-Infrastructure Multi-Sensor Fusion (V2I-MSF) With Reinforcement Learning Framework for Enhancing Autonomous Vehicle Perception. *IEEE Access*, 13, 50122-50136. <https://doi.org/10.1109/access.2025.3551367>
- [6] Lv, X. (2025). Research on Path Planning Method of Substation Inspection Robot Based on Multi Sensor Fusion. *IEEE Access*, 13, 213454-213468. <https://doi.org/10.1109/access.2025.3645321>
- [7] Zhang, Z., & Chen, L. (2025). A Low-Cost Environment-Interactive Patrol Inspection System With Navigation Based on Sensor-Fusion and Robotic Arm Contact Pose Feedback. *IEEE Access*, 13, 54547-54560. <https://doi.org/10.1109/access.2025.3553521>
- [8] Ho Lee, S., & Choi, W. Y. (2024). Enhancing Object Estimation by Camera-LiDAR Sensor Fusion Using IMM-KF With Error Characteristics in Autonomous Robot Systems. *IEEE Access*, 12, 197247-197258. <https://doi.org/10.1109/access.2024.3522090>
- [9] Ogunbunmi, S., Aluko, S., & Onukak, P. (2025). Cyber-Resilient Perception: Safeguarding Autonomous Vehicles With Trust-Aware Sensor Fusion. *IEEE Sensors Reviews*, 2, 86-94. <https://doi.org/10.1109/sr.2025.3562156>
- [10] Shafique, R., Rustam, F., Murtala, S., Jurcut, A. D., & Choi, G. S. (2024). Advancing Autonomous Vehicle Safety: Machine Learning to Predict Sensor-Related Accident Severity. *IEEE Access*, 12, 25933-25948. <https://doi.org/10.1109/access.2024.3366990>
- [11] Ou, Y., Cai, Y., Sun, Y., & Qin, T. (2024). Autonomous Navigation by Mobile Robot with Sensor Fusion Based on Deep Reinforcement Learning. *Sensors*, 24(12), 3895. <https://doi.org/10.3390/s24123895>
- [12] Paniego, S., Paliwal, N., & Cañas, J. (2024). Model Optimization in Deep Learning Based Robot Control for Autonomous Driving. *IEEE Robotics and Automation Letters*, 9(1), 715-722. <https://doi.org/10.1109/lra.2023.3336244>
- [13] Yu, K., Fu, M., Zhang, T., & Yang, Y. (2025). Enhancing Safety in Autonomous Racing With Constrained Reinforcement Learning. *IEEE Robotics and Automation Letters*, 10(6), 6448-6455. <https://doi.org/10.1109/lra.2025.3566591>
- [14] Gao, W., Yu, Z., Wang, L., Cui, H., Guo, B., & Xiong, H. (2025). Hierarchical Deep Reinforcement Learning for Computation Offloading in Autonomous Multi-Robot Systems. *IEEE Robotics and Automation Letters*, 10(1), 540-547. <https://doi.org/10.1109/lra.2024.3511408>
- [15] Huang, Y. Y., & Song, K. T. (2024). Human-to-Robot Handover Control of an Autonomous Mobile Robot Based on Hand-Masked Object Pose Estimation. *IEEE Robotics and Automation Letters*, 9(9), 7851-7858. <https://doi.org/10.1109/lra.2024.3433205>
- [16] Yan, Z., Huang, R., He, L., Guo, S., & Zhao, L. (2026). SIGN: Safety-Aware Image-Goal Navigation for Autonomous Drones via Reinforcement Learning. *IEEE Robotics and Automation Letters*, 11(2), 1962-1969. <https://doi.org/10.1109/lra.2025.3645668>
- [17] Dong, Y., Zhao, X., Wang, S., & Huang, X. (2024). Reachability Verification Based Reliability Assessment for Deep Reinforcement Learning Controlled Robotics and Autonomous Systems. *IEEE Robotics and Automation Letters*, 9(4), 3299-3306. <https://doi.org/10.1109/lra.2024.3364471>
- [18] Guo, J., Feng, M., Liu, J., Wang, Y., & Pu, J. (2025). Certificating Safety of Imitation Learning for Autonomous Driving With Learnable Weighted Control Barrier Functions. *IEEE Robotics and Automation Letters*, 10(10), 9702-9709. <https://doi.org/10.1109/lra.2025.3597485>
- [19] Tang, B., Yu, Q., & Liu, Z. (2026). Drivers' Safety Perception in Autonomous Vehicle Road Sharing: A Knowledge-Segmented TPB and Ordered Logit Analysis. *Applied Sciences*, 16(7), 3599. <https://doi.org/10.3390/app16073599>
- [20] Ušinskis, V., Nowicki, M., Dzedzickis, A., & Bučinskas, V. (2025). Sensor-Fusion Based Navigation for Autonomous Mobile Robot. *Sensors*, 25(4), 1248. <https://doi.org/10.3390/s25041248>
- [21] Shili, M., Chaoui, H., & Nouri, K. (2025). Energy-Aware Sensor Fusion Architecture for Autonomous Channel Robot Navigation in Constrained Environments. *Sensors*, 25(21), 6524. <https://doi.org/10.3390/s25216524>
- [22] Liu, Z., Fang, J., & Zhao, Y. (2025). DynaFusion-SLAM: Multi-Sensor Fusion and Dynamic Optimization of Autonomous Navigation Algorithms for Pasture-Pushing Robot. *Sensors*, 25(11), 3395. <https://doi.org/10.3390/s25113395>
- [23] Dai, Y., & Lee, K. (2025). Multi-Sensor Fusion for Autonomous Mobile Robot Docking: Integrating LiDAR, YOLO-Based AprilTag Detection, and Depth-Aided Localization. *Electronics*, 14(14), 2769. <https://doi.org/10.3390/electronics14142769>

- [24] Brödermann, T., Sakaridis, C., Piccinelli, L., Abbeloos, W., & Van Gool, L. (2026). DGFusion: Depth-Guided Sensor Fusion for Robust Semantic Perception. *IEEE Robotics and Automation Letters*, 11(3), 3334-3341. <https://doi.org/10.1109/lra.2026.3656789>
- [25] Wei, H., Wang, P., Wang, Q., Gu, Y., Qin, T., Gao, F., & Si, Y. (2026). FAR-AVIO: Fast and Robust Schur-Complement Based Acoustic-Visual-Inertial Fusion Odometry With Sensor Calibration. *IEEE Robotics and Automation Letters*, 11(7), 7740-7747. <https://doi.org/10.1109/lra.2026.3688065>
- [26] Zhou, Y., Li, X., Li, S., Wang, X., Feng, S., & Tan, Y. (2024). DBA-Fusion: Tightly Integrating Deep Dense Visual Bundle Adjustment With Multiple Sensors for Large-Scale Localization and Mapping. *IEEE Robotics and Automation Letters*, 9(7), 6138-6145. <https://doi.org/10.1109/lra.2024.3400156>
- [27] Balemans, N., Anwar, A., Steckel, J., & Mercelis, S. (2024). LiDAR-BIND: Multi-Modal Sensor Fusion Through Shared Latent Embeddings. *IEEE Robotics and Automation Letters*, 9(11), 9159-9166. <https://doi.org/10.1109/lra.2024.3457384>
- [28] Phan, T. D., & Kim, G. W. (2025). FusionGS-SLAM: Multiple Sensors Fusion for Localization and Real-Time Photorealistic Mapping. *IEEE Robotics and Automation Letters*, 10(9), 8690-8697. <https://doi.org/10.1109/lra.2025.3585388>
- [29] Krátký, V., Silano, G., Vrba, M., Papaioannidis, C., Mademlis, I., Pěnička, R., Pitas, I., & Saska, M. (2025). Gesture-Controlled Aerial Robot Formation for Human-Swarm Interaction in Safety Monitoring Applications. *IEEE Robotics and Automation Letters*, 10(8), 8244-8251. <https://doi.org/10.1109/lra.2025.3583607>
- [30] Wan, A., Wu, Z., & AL-Bukhaiti, K. (2025). Deep learning-based intelligent visual inspection for defect detection in smoke sensor manufacturing. *Next Research*, 2(3), 100424. <https://doi.org/10.1016/j.nexres.2025.100424>