Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.gaexcellence.com/jistm

CONFIDENCE-WEIGHTED MID-LEVEL LIDAR–STEREO FUSION WITH HYBRID PCA-K-MEANS CLUSTERING FOR MARITIME 3D OBJECT DETECTION

Muhammad Aiman Norazaruddin[1]*, Zulkifli Zainal Abidin[2]

[1]Department of Mechatronics Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Malaysia
aiman.norazaruddin@gmail.com https://orcid.org/0009-0002-3651-3396
2Department of Mechatronics Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Malaysia
zzulkifli@iium.edu.my https://orcid.org/0000-0003-3559-2440
*Corresponding Author

Article Info:

Article history:
Received date: 29.06.2026
Revised date: 12.07.2026
Accepted date: 04.08.2026
Published date: 15.09.2026

To cite this document:

Norazaruddin, M. A., & Abidin, Z. Z. (2026). Confidence-Weighted Mid-Level LiDAR–Stereo Fusion with Hybrid PCA-K-Means Clustering for Maritime 3D Object Detection. *Journal of Information System and Technology Management, 11* (44), 258-257.

Abstract:

Autonomous Unmanned Surface Vessels (USVs) operating in nearshore environments require robust perception systems capable of handling challenging maritime conditions. However, LiDAR and stereo vision each suffer from complementary limitations: LiDAR produces sparse and reflection-sensitive measurements, while stereo vision provides dense depth estimates but performs poorly under occlusion and weak-texture conditions. While early- and late-fusion approaches have been widely explored, mid-level fusion strategies that preserve cross-modal feature interactions while adapting to sensor uncertainty remain underdeveloped for maritime perception. This paper introduces M-ACF Net (Maritime Adaptive Confidence Fusion Network), a confidence-adaptive mid-level fusion framework that dynamically integrates sparse LiDAR and dense stereo features based on pixel-level reliability. The fused representation is processed using PCA-K-Means clustering to generate orientation-invariant obstacle representations for Guidance, Navigation, and Control (GNC). The framework was implemented on an instrumented USV platform and evaluated on a multimodal dataset of 17,634 synchronised LiDAR–stereo–GNSS samples collected across three nearshore Malaysian sites under wave heights of 0.1–0.5 m. Across different wave conditions, M-ACF Net reduced depth-estimation RMSE by 44–60% compared with LiDAR-only measurements. The framework achieved 82.6% 3D Average Precision for ship detection under easy conditions and outperformed YOLOv8-Stereo, PointPillars, Pseudo-LiDAR, Frustum PointNet, and DSGN baselines, while matching DBSCAN's clustering quality at less than half

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275
its computation time and sustaining 18 frames per second on an embedded NVIDIA Jetson Orin NX. Field trials further validated real-time obstacle detection under operational conditions. These results demonstrate the effectiveness of confidence-adaptive mid-level fusion for real-time maritime obstacle perception.

DOI: 10.35631/JISTM.1144016

Keyword:

Unmanned Surface Vessel; LiDAR–Stereo Fusion; Mid-Level Sensor Fusion; Principal Component Analysis; K-Means Clustering; Maritime 3D Object Detection; Guidance, Navigation and Control.

Introduction

Safe operation at sea depends on a vessel's ability to perceive its surroundings with enough range, resolution, and reliability to avoid collision with other craft, fixed structures, and floating debris. Conventional sensors fall short of this on their own: radar struggles to separate small or low-profile targets from sea clutter, and cameras alone cannot measure depth directly and lose effectiveness under poor lighting. LiDAR overcomes these limitations through active, illumination-independent ranging, but nearshore deployment introduces its own difficulties. Water surfaces absorb rather than reflect infrared pulses, producing sparse returns, while wave-induced motion distorts the geometric consistency of successive scans (Bovcon et al., 2021; Damodaran et al., 2023; Zhou et al., 2021). Stereo vision complements LiDAR with dense, texture-rich depth estimates, but its reliability degrades sharply under occlusion, glare, and the low-texture conditions typical of open water.

How best to combine these two complementary, imperfect modalities remains an open architectural question. Early fusion strategies, which merge raw or near-raw sensor data before feature extraction, exploit rich cross-modal correlation but are highly sensitive to calibration drift and temporal misalignment, both of which are exacerbated by wave-induced platform motion (Liu et al., 2023). Late fusion strategies, which combine only the final outputs of independently processed pipelines, are robust to such misalignment but discard the intermediate feature relationships that could otherwise improve detection of partially occluded or weakly textured objects. Mid-level fusion, in which each modality is processed to an intermediate feature representation before merging, offers a compromise between these extremes, yet its application to nearshore maritime perception has received comparatively little attention relative to its use in terrestrial autonomous driving (Sindagi et al., 2019; Vora et al., 2020).

A further gap concerns post-fusion localisation. Deep learning-based 3D object detectors achieve strong benchmark performance in automotive contexts, but their computational cost is often incompatible with the power and hardware constraints of an embedded unmanned surface vessel (USV). Classical geometric methods such as Principal Component Analysis (PCA) and K-Means clustering are comparatively lightweight and have been used for obstacle localisation in robotics generally. However, their integration as a post-processing stage within a multimodal maritime fusion pipeline, specifically to denoise and segment fused LiDAR-stereo point clouds into navigable obstacle coordinates, has not been systematically investigated.

This paper addresses both gaps with M-ACF Net (Maritime Adaptive Confidence Fusion Network), a mid-level fusion framework that dynamically integrates LiDAR and stereo depth features through a per-pixel confidence model, rather than a fixed-ratio or early/late alternative. The fused output is processed through a hybrid PCA-K-Means clustering stage that recovers orientation-invariant centroid and extremity coordinates suitable for GNC consumption. Unlike prior maritime fusion proposals, M-ACF Net is evaluated empirically rather than merely proposed: it was implemented on an instrumented USV and validated on a 17,634-sample multimodal dataset collected across three real nearshore environments. The contributions of this paper are: a complete mathematical formulation of the confidence-weighted mid-level fusion model and the PCA-K-Means clustering stage; an empirical benchmark of the proposed framework against five state-of-the-art 3D object detectors (YOLOv8-Stereo, PointPillars, Pseudo-LiDAR, Frustum PointNet, and DSGN) and four alternative clustering algorithms (DBSCAN, Hierarchical Clustering, Spectral Clustering, and Gaussian Mixture Models); a quantitative analysis of system robustness across wave height, weather, and cross-domain detection conditions; and field-validated, real-time deployment results on an NVIDIA Jetson Orin NX embedded platform. Portions of the experimental results reported in this paper are drawn from the first author's doctoral research, conducted under the supervision of the second author (Norazaruddin, 2026).

Literature Review

Maritime Sensing Modalities and the Case for LiDAR-Stereo Fusion

Unmanned Surface Vessels (USVs) depend on their Guidance, Navigation, and Control (GNC) systems to interpret exteroceptive sensor data and convert it into safe trajectory and actuator commands (Liu et al., 2016). Among the available exteroceptive modalities, LiDAR occupies a distinctive position: unlike radar, it provides dense, high-resolution range measurements; unlike passive cameras, it does not depend on ambient illumination or stereo baseline geometry that degrades with distance. Its growing affordability and shrinking form factor (Aijazi et al., 2020) have made shipboard deployment increasingly practical, motivating adoption across coastal morphology mapping, real-time marine object tracking, and offshore-structure inspection. Nonetheless, the same nearshore conditions that make LiDAR attractive also constrain its performance: infrared beam absorption by water produces point clouds that are markedly sparser than equivalent terrestrial scans, and continuous wave-induced platform motion introduces frame-to-frame geometric inconsistency with no direct analogue in road-vehicle LiDAR perception (Damodaran et al., 2023; Zhou et al., 2021).

Stereo vision offers a complementary profile: dense depth estimation at every pixel, rich semantic texture, and lower unit cost, at the expense of degraded reliability under occlusion, glare, and low-texture water surfaces. Prior maritime perception work has tended to treat

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275
LiDAR and stereo vision as alternatives rather than as complementary inputs to a single fusion architecture, despite evidence from terrestrial autonomous driving that combining geometric and photometric modalities improves robustness in exactly the conditions, occlusion, reflection, and clutter, that are most acute in nearshore navigation.

Sensor Fusion Strategies: Early, Late, and Mid-Level Fusion

Multimodal sensor fusion architectures are conventionally classified by the processing stage at which modalities are combined. Early fusion approaches merge raw or near-raw sensor data, for example by re-projecting LiDAR points onto the image plane to construct a pseudo-image, as in MVX-Net (Sindagi et al., 2019) and PointPainting (Vora et al., 2020). This produces a dense, information-rich representation but is highly sensitive to calibration inaccuracy and temporal misalignment, issues that are particularly pronounced in dynamic marine environments where platform motion and surface disturbance introduce projection error (Liu et al., 2023). Late fusion approaches instead maintain independent processing pipelines for each modality and merge only the final outputs, for example through probabilistic ensembling of independent detections (Chen et al., 2021). This mitigates calibration sensitivity but restricts the network's ability to exploit cross-modal feature correlation, reducing effectiveness for objects with partial occlusion, weak texture, or ambiguous boundaries.

Mid-level fusion processes each modality independently to an intermediate feature level before merging, retaining a measure of late fusion's robustness while preserving some of early fusion's contextual richness. Despite demonstrated benefits in terrestrial multimodal perception (Irfan et al., 2025), its application to nearshore USV perception remains comparatively underexplored, motivating the architecture proposed in this paper.

Point Cloud Preprocessing and Hybrid Clustering

A substantial body of LiDAR preprocessing literature exists to mitigate sensor noise. Voxel Grid Downsampling reduces point density by replacing all points within a fixed-size cubic volume with a single centroid, trading fine spatial detail for a proportional reduction in computational load (Hacinecipoglu, 2022). Statistical Outlier Removal (SOR) targets noise directly: for each point, the mean and standard deviation of distances to neighbouring points within a search radius are computed and points whose mean neighbourhood distance falls outside a statistically defined acceptance band are discarded (Balta et al., 2018).

Following preprocessing, a detected object's point cloud must be reduced to a small number of actionable coordinates. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) expands clusters from core points satisfying a minimum-neighbour-count condition within an epsilon radius, explicitly labelling low-density points as noise (Ester et al., 1996). K-Means, by contrast, partitions points into a fixed number of clusters by iteratively minimising within-cluster variance (MacQueen, 1967); it is computationally lighter than density-based alternatives but is sensitive to the arbitrary orientation of elongated objects, such as ship hulls viewed from varying USV headings. Principal Component Analysis (PCA) addresses this sensitivity by re-projecting a point cloud onto its orthogonal axes of maximum variance prior to clustering, producing an orientation-invariant coordinate frame in which K-Means is markedly more stable (Jolliffe & Cadima, 2016). Cluster quality across such methods is conventionally assessed using internal validity indices, including the Silhouette Score (Rousseeuw, 1987), which jointly quantifies cluster cohesion and separation.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

Learned 2D and 3D Detection in the Maritime and Terrestrial Domains

Two-dimensional object detection architectures originally developed for natural images, including Faster R-CNN (Ren et al., 2015) and the YOLO family, have each been adapted for marine imagery with varying success, and deep-learning-based detectors consistently outperform classical feature-engineering approaches on this class of imagery (Er et al., 2023; R. Zhang et al., 2021). However, single-stage detectors such as YOLOv8 have been shown to be more vulnerable to cross-domain distribution shift than two-stage architectures such as Faster R-CNN, a robustness consideration directly relevant to maritime deployment across varying sites and conditions, and one quantified empirically in this paper.

Three-dimensional detection has progressed rapidly in terrestrial domains. End-to-end, learned architectures such as VoxelNet (Zhou & Tuzel, 2018), PointPillars (Lang et al., 2019), and SECOND (Yan et al., 2018) achieve strong benchmark performance on densely annotated road-vehicle datasets, while multi-view fusion networks such as MV3D (Chen et al., 2017) and AVOD (Ku et al., 2018) combine bird's-eye-view and front-view LiDAR projections with image features for improved occlusion robustness. Stereo-specific architectures including Pseudo-LiDAR (Weng & Kitani, 2019) and DSGN (Chen et al., 2020) instead derive 3D structure directly from stereo disparity, while Frustum PointNets (Qi et al., 2018) use a 2D detector to restrict the 3D search space before applying a learned point-set segmentation network. These architectures' applicability to sparse, comparatively under-annotated maritime point clouds is rarely quantified directly against one another under shared sensing and hardware conditions, a gap this paper's benchmarking protocol is designed to close.

Research Gap and Study Rationale

Synthesising the literature reviewed above, five specific gaps motivate the present work. First, fusion architectures developed for terrestrial autonomous driving do not transfer directly to the dynamic water surfaces, variable reflective lighting, and irregular partially submerged obstacles of nearshore maritime environments (Bovcon et al., 2021). Second, most existing maritime fusion frameworks rely on early or late fusion, with mid-level architectures left largely unexplored despite their favourable trade-off profile. Third, the computational complexity of state-of-the-art 3D detectors is frequently incompatible with the power and hardware constraints of real-time, embedded USV deployment, and few studies have explored hybrid classical-learned approaches that balance accuracy against efficiency. Fourth, although PCA and K-Means are individually well established, their combined integration as a denoising and segmentation stage within a multimodal maritime fusion pipeline has not been systematically investigated. Fifth, existing maritime perception benchmarks rarely evaluate competing 3D detection architectures against one another under identical sensing hardware and site conditions, limiting the comparability of reported results across the literature.

This study is positioned directly within these five gaps. It proposes a mid-level, confidence-adaptive LiDAR-stereo fusion architecture; integrates PCA and K-Means as a lightweight, orientation-invariant post-fusion clustering stage; and reports a controlled, shared-hardware empirical comparison against five 3D detection baselines and four clustering alternatives, together with field-validated real-time deployment results.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

Methodology

Overview

M-ACF Net is a three-stage perception pipeline. The first stage preprocesses and synchronises raw LiDAR and stereo sensor data. The second stage uses a 2D detector to guide the cropping of object-centric LiDAR and stereo regions, which are then combined through a confidence-weighted mid-level fusion model to produce a fused geometric feature map. The third stage clusters the resulting object-centric point cloud using a hybrid PCA-K-Means algorithm to extract a single, GNC-ready 3D coordinate per detected obstacle.

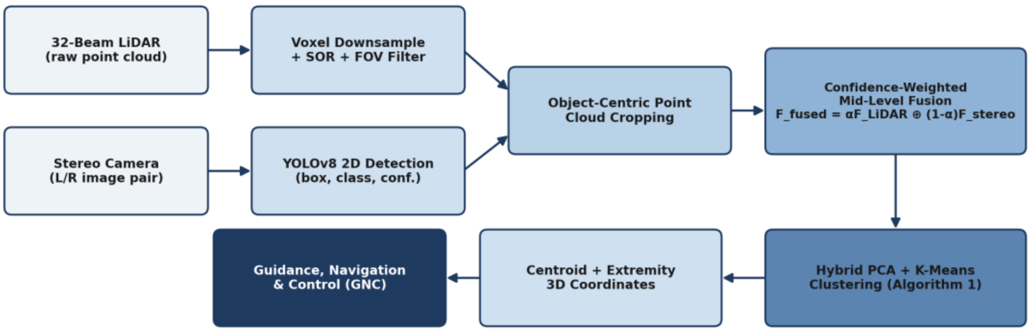

Figure 1: Mid-Level Lidar-Stereo Fusion Pipeline.

Sensor Calibration and Synchronisation

All sensor frames (LiDAR, stereo camera, GNSS, INS) are expressed relative to a common base frame through rigid-body transformations composed of a rotation matrix R in SO(3) and translation vector t in R^3, as shown in Equation (1).

$$P_{\text{camera}} = R\,P_{\text{LiDAR}} + t \tag{1}$$

Temporal alignment across sensors, each operating on independent clocks and sampling rates, is established using a Pulse-Per-Second (PPS) signal from the GNSS module as a common synchronisation trigger, producing the synchronised dataset defined in Equation (2), where t is the universal PPS-derived timestamp.

$$F_{\text{data}} = \text{sync}(d_{\text{GNSS}}, d_{\text{LiDAR}}, d_{\text{IMU}}) = \{(t, D(t)) : D(t) = (d_{\text{GNSS}}(t), d_{\text{LiDAR}}(t), d_{\text{IMU}}(t))\} \tag{2}$$

Extrinsic calibration between the LiDAR and stereo camera, the most critical calibration step for mid-level fusion, is performed using a custom calibration target and the Iterative Closest Point (ICP) algorithm shown in Equation (3), minimising the spatial discrepancy between matched 3D points across modalities.

$$\min_{R,t} \frac{1}{N}\sum_{i=1}^{N} \left\| P_{\text{camera},i} - \text{closest}\left(R\,P_{\text{LiDAR},i} + t\right) \right\|^2 \tag{3}$$

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

Verification against known reference distances and repeatability testing yielded a final translational discrepancy under 2 cm and rotational error below 0.5 degrees, an accuracy necessary to ensure that the object-centric point clouds generated downstream maintain reliable geometric consistency.

LiDAR Point Cloud Preprocessing

Raw LiDAR returns are first reduced through voxel grid downsampling, in which each point's coordinates are divided by a voxel size v_s and rounded to obtain a voxel index, as in Equation (4); each occupied voxel is then replaced by the centroid of its constituent points, Equation (5).

$$I_{(x,y,z)} = \left\lfloor \frac{(x,y,z)}{v_s} \right\rfloor \tag{4}$$

$$C_{(x,y,z)} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{5}$$

Statistical Outlier Removal (SOR) is then applied to suppress sparse-return noise, characteristic of water-spray interference: for each point, the mean mu and standard deviation sigma of its neighbourhood distances within radius r are computed (Equations 6 and 7) and points whose mean distance falls outside the band mu plus-or-minus k*sigma are discarded.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{6}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (d_i - \mu)^2} \tag{7}$$

Grid-search optimisation identified a voxel size of 0.2 m and an SOR multiplier of k = 1.5 as the best-performing configuration for this sensor and site combination, yielding a 12% gain in downstream clustering F1-score relative to unoptimised settings. Finally, the preprocessed point cloud is restricted to the stereo camera's narrower egocentric field of view, approximately plus-or-minus 60 degrees azimuth against the LiDAR's full 360-degree coverage, ensuring spatial correspondence between the two modalities prior to fusion.

Two-Dimensional-Detection-Guided Object-Centric Cropping

A standard YOLOv8 model, pretrained on the Singapore Maritime Dataset (SMD) and SeaShips dataset and fine-tuned per deployment site, is used as a feature extractor and region-proposal network rather than as the final 3D localiser. For each 2D detection, the corresponding LiDAR points are cropped to form an egocentric, object-centric point cloud, effectively transforming each 2D detection into a localised 3D search region. This step does not modify the YOLOv8 backbone; it determines only where subsequent fusion and clustering operate, isolating the contribution of the downstream stages from the contribution of the 2D classifier itself.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

Confidence-Weighted Mid-Level Fusion

The core of M-ACF Net is a confidence-weighted feature fusion model that produces a single, fused geometric feature, F_{fused}, as a convex combination of the LiDAR-derived feature, F_{LiDAR}, and the stereo-derived feature, F_{stereo}, governed by a dynamic weighting factor alpha in [0, 1], shown in Equation (8).

$$F_{\text{fused}} = \alpha F_{\text{LiDAR}} + (1 - \alpha) F_{\text{stereo}} \quad (8)$$

The weighting factor alpha is computed per pixel or local region from confidence metrics derived from both sensors, Equation (9), reflecting the relative reliability of each modality at a given location and moment.

$$\alpha = \frac{C_{\text{LiDAR}}}{C_{\text{LiDAR}} + C_{\text{stereo}}} \quad (9)$$

LiDAR confidence, C_{LiDAR}, is formulated as the ratio of valid points retained after preprocessing, N_{filtered}, to the total raw points acquired per frame, N_{raw}, Equation (10); stereo confidence, C_{stereo}, is estimated from occlusion consistency between the left and right stereo views, Equation (11), where N_{occ} is the count of pixels with inconsistent left-right correspondence and N_{total} is the total pixel count.

$$C_{\text{LiDAR}} = \frac{N_{\text{filtered}}}{N_{\text{raw}}} \quad (10)$$

$$C_{\text{stereo}} = 1 - \frac{N_{\text{occ}}}{N_{\text{total}}} \quad (11)$$

The fused depth value at each pixel is reprojected into 3D space using the stereo camera's intrinsic parameters, Equation (12), where (u, v) are image coordinates and (f_x, f_y, c_x, c_y) the camera intrinsics, yielding the object-centric 3D point set passed to the clustering stage.

$$\begin{bmatrix} X \\ Y \end{bmatrix} = F_{\text{fused}} \begin{bmatrix} \dfrac{u - c_x}{f_x} \\ \dfrac{v - c_y}{f_y} \end{bmatrix} \quad (12)$$

Hybrid PCA-K-Means Clustering for Obstacle Localisation

A naive geometric mean of an object-centric point cloud is highly susceptible to outliers from wave spray or residual noise, while standard K-Means clustering is sensitive to the arbitrary orientation of elongated objects such as ship hulls. M-ACF Net addresses both limitations with a hybrid PCA-K-Means stage. For a 3D point cloud P = {(X_i, Y_i, Z_i)}, the mean vector mu and covariance matrix Sigma are first computed (Equations 13 and 14).

$$\mu = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{13}$$

$$\Sigma = \frac{1}{N-1} P_c^T P_c \tag{14}$$

Eigen-decomposition of Sigma (Equation 15) yields eigenvectors that define the object's principal axes of variance; projecting the point cloud onto these axes produces an orientation-aligned coordinate frame, P_{PCA}, in which the object's geometry is standardised regardless of the USV's viewing angle.

$$\Sigma v = \lambda v \tag{15}$$

K-Means clustering is then applied within this transformed space, partitioning the N points into k clusters by minimising the within-cluster sum of squares, J, shown in Equation (16), where mu_i is the centroid of cluster i.

$$J = \sum_{i}\sum_{j}\left\| x_j^{(i)} - \mu_i \right\|^2 \tag{16}$$

The number of clusters k is adjusted dynamically based on the variance captured by the first principal component, allowing large or elongated vessels to be represented by multiple centroids where appropriate. Beyond the cluster centroid, the point of greatest absolute value along the first principal component is also extracted as the object's physical extremity, the point closest to the USV along its dominant axis and therefore the more navigationally relevant reference for collision avoidance. Both the centroid and extremity coordinates, transformed back into the USV's base frame, are passed to the GNC system. Algorithm 1 summarises the full clustering procedure.

```
Input:  object-centric point cloud P, number of clusters k
Output: cluster centroids C, principal-axis extremity E

mu <- mean(P);  P_c <- P - mu
Sigma <- cov(P_c);  lambda, v <- eig(Sigma)
sorted_v <- sort eigenvectors by descending eigenvalue
P_PCA <- P_c . sorted_v                       // Eq. (13)-(15)
centroids <- initialise_centroids(P_PCA, k)
repeat
    clusters <- assign_nearest(P_PCA, centroids)
    centroids <- update_means(clusters)         // Eq. (16)
until convergence
E <- argmax_i |P_PCA,i . PC1|                  // furthest point along PC1
return centroids, E (transformed back to USV base frame)
```

Algorithm 1. Hybrid PCA-K-Means Clustering

Journal of Information System and Technology Management JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

USV Platform and Sensor Suite

M-ACF Net was implemented and evaluated on an instrumented USV platform equipped with a 32-beam LiDAR, a forward-facing stereo camera, a GNSS receiver providing the PPS synchronisation reference, and an inertial navigation system (INS). All onboard inference was executed on an NVIDIA Jetson Orin NX embedded GPU, the platform identified by prior maritime perception literature as the relevant real-time deployment target.

Dataset Description

Data were collected across three nearshore Malaysian sites selected to represent distinct operational complexities: Taman Tasik Metropolitan Kepong (Dataset A), a low-density recreational lake; Pelabuhan Klang (Dataset B), a high-traffic commercial port; and Putrajaya Lake (Dataset C), an urban waterway with mixed recreational and transient traffic. Table 1 summarises the resulting multimodal dataset of 17,634 synchronised samples.

Table 1: Multimodal Maritime Dataset Composition Across Three Nearshore Environments.

Environment	Setting	Samples	Wave Height (m)	Dominant Object Categories
Dataset A - Taman Tasik	Recreational lake	1,075	0-0.2	Small watercraft, jetties
Dataset B - Pelabuhan Klang	Commercial port	8,826	0.1-0.5	Container ships, cargo vessels, cranes
Dataset C - Putrajaya Lake	Urban waterway	7,733	0-0.2	Recreational boats, jet skis, bridges
Total	-	17,634	0-0.5	-

Environmental conditions varied across midday brightness, dusk, rain, and haze, providing coverage of the lighting and weather variability typical of operational nearshore deployment.

Baseline Methods

M-ACF Net's 3D detection performance is benchmarked against five state-of-the-art methods spanning the camera-only, voxel-based, and stereo-based architectural families reviewed above: YOLOv8-Stereo, PointPillars (Lang et al., 2019), Pseudo-LiDAR (Weng & Kitani, 2019), Frustum PointNet (Qi et al., 2018), and DSGN (Chen et al., 2020). M-ACF Net's clustering stage is separately benchmarked against four alternative algorithms: DBSCAN (Ester et al., 1996), Hierarchical (Agglomerative) Clustering, Spectral Clustering, and Gaussian Mixture Models (GMMs). All comparisons were conducted under identical sensing hardware, dataset, and embedded compute conditions.

Evaluation Metrics

3D detection is evaluated using 3D bounding-box Average Precision (AP) and bird's-eye-view (BEV) AP, with IoU thresholds of 0.5 for ships and 0.4 for buoys and bridges, reported across easy, moderate, and hard scene-complexity tiers, alongside inference throughput in frames per

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

second (FPS). Clustering quality is evaluated using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, while distance-estimation accuracy is evaluated using RMSE, MAE, and the coefficient of determination (R^2). 2D detection is evaluated using precision, recall, and mean Average Precision at IoU 0.5 (mAP@0.5).

Results

Stereo-LiDAR Depth Fusion Under Increasing Wave Height

Stereo recovery efficiency, the proportion of wave-occluded depth successfully recovered, declined from 95.2% ± 1.5% at 0.1 m waves to 72.3% ± 5.5% at 0.5 m waves, with a strong negative correlation between wave height and recovery efficiency ($r = -0.89$, $p < 0.001$). Despite this decline, the fused depth estimate consistently and substantially outperformed LiDAR-only measurement: RMSE for the fused estimate ranged from 0.10 ± 0.02 m at 0.1 m waves to 0.28 ± 0.06 m at 0.5 m waves, a 44–60% reduction relative to LiDAR-only RMSE of 0.25 ± 0.04 m to 0.52 ± 0.09 m over the same range (Figure 2). The stereo module itself processed at 9.8 ± 1.3 ms per frame (102 FPS on the Jetson Orin NX), confirming that the fusion overhead is not the binding real-time constraint.

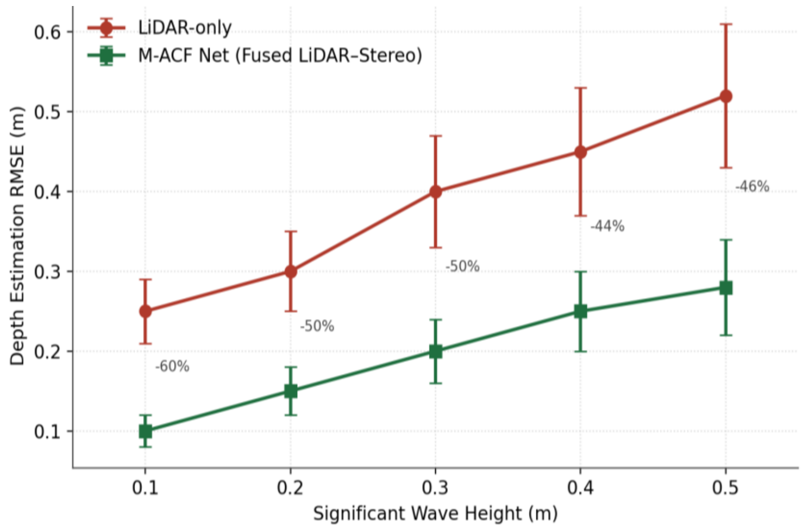

Figure 2: Depth Estimation RMSE Vs. Wave Height: Fused Lidar-Stereo Vs. Lidar-Only Measurement.

Confidence-Weighted Fusion Across Environmental Conditions

Across calm waters, choppy waters, rainy weather, fog, and high-traffic conditions, the confidence-weighted fusion consistently reduced RMSE relative to raw, unfused depth, by 50% in calm waters, 49% in rainy weather, and 44% in high-traffic conditions, with outlier reduction ranging from 72.5% to 50.0% over the same conditions. Processing time increased moderately with environmental complexity, from 45 ms per frame in calm waters to 70 ms in high-traffic conditions, remaining within the sub-100 ms real-time threshold throughout. Consistent with the confidence formulation in Equations (9) to (11), the fusion weight alpha favoured LiDAR (alpha >= 0.6) under rough-sea and low-light conditions, where LiDAR's illumination independence dominates, and shifted toward stereo (alpha < 0.5) in close-range, well-lit conditions where stereo's higher resolution improves depth precision.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

Clustering Benchmark: PCA-K-Means vs. Four Alternatives

Table 2 compares the proposed PCA-K-Means clustering against four alternatives on cluster-quality and computation-time metrics. DBSCAN achieved the second-highest Silhouette Score (0.75) and Calinski-Harabasz Index (80), behind Spectral Clustering, but at more than twice the computation time of PCA-K-Means (0.12 s vs. 0.05 s per frame). Spectral Clustering achieved the best overall cluster-quality scores but at the highest computation cost (0.30 s), six times slower than the proposed method.

Table 2: Clustering Performance and Computation Time Benchmark.

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Computation Time (s)
PCA-K-Means (Proposed)	0.70	0.50	60	0.05
DBSCAN	0.75	0.60	80	0.12
Hierarchical Clustering	0.70	0.70	50	0.25
Spectral Clustering	0.80	0.55	90	0.30
Gaussian Mixture Model	0.75	0.58	70	0.20

Table 3 reports distance-estimation accuracy for the same algorithms. PCA-K-Means achieved the lowest RMSE (1.6 m) and MAE (1.2 m) and the highest R^2 (0.88) of all five methods, indicating that orientation-invariant alignment prior to clustering improves spatial accuracy even though DBSCAN attains marginally better generic cluster-separation indices in Table 2.

Table 3: Distance Estimation Accuracy By Clustering Algorithm.

Algorithm	RMSE (m)	MAE (m)	R^2
PCA-K-Means (Proposed)	1.6	1.2	0.88
DBSCAN	2.0	1.6	0.80
Hierarchical Clustering	2.8	2.2	0.70
Spectral Clustering	1.8	1.4	0.85
Gaussian Mixture Model	2.1	1.7	0.78

Two-Dimensional Detection and Cross-Domain Robustness

In-domain, the YOLOv8x model trained and tested on the Singapore Maritime Dataset (SMD) achieved 0.983 precision and 0.961 recall; the YOLOv8m model trained and tested on SeaShips achieved 0.913 precision and 0.889 recall. Cross-domain evaluation, training on one dataset and testing on the other, revealed a substantial robustness gap: the SMD-trained YOLOv8 model's precision collapsed to 0.167 when applied to SeaShips imagery, while the SeaShips-trained model retained 0.508 precision on SMD imagery. Faster R-CNN, evaluated under the

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

same cross-domain protocol, showed a comparatively smaller decline (recall of 0.557 and 0.524 across the two cross-domain directions), indicating that single-stage detector architecture choice carries a measurable robustness cost under the domain shift that is expected whenever a maritime perception system is deployed at a new site without local fine-tuning.

System-Level Three-Dimensional Object Detection Benchmark

The complete M-ACF Net pipeline was benchmarked end-to-end against five 3D detection baselines across three object categories (ships, buoys, bridges) and three scene-complexity tiers (easy, moderate, hard), all on the NVIDIA Jetson Orin NX. M-ACF Net achieved the highest 3D AP in every category and difficulty tier tested, recording 82.6%, 70.8%, and 58.3% AP for ship detection under easy, moderate, and hard conditions respectively (Table 4, Figure 3) — a margin of 5.3, 5.8, and 5.9 percentage points over the next-best baseline, DSGN, at each tier — while sustaining 18 FPS. YOLOv8-Stereo achieved the second-highest throughput among the baselines (25 FPS) but consistently the lowest AP, illustrating the speed-accuracy trade-off that motivates M-ACF Net's hybrid design; PointPillars achieved the highest raw FPS overall (35) but degraded sharply under harder conditions, particularly for smaller or partially occluded objects such as buoys and bridge components.

Table 4: 3D Bounding-Box Average Precision (%) And Inference Throughput (FPS) By Method And Difficulty Tier (Ship Category Shown).

Method	Easy	Moderate	Hard	FPS
YOLOv8-Stereo	68.2	55.1	42.3	25
PointPillars	72.5	60.8	48.6	35
Pseudo-LiDAR	75.1	63.4	50.2	15
Frustum PointNet	70.8	58.3	45.9	10
DSGN	77.3	65.0	52.4	20
M-ACF Net	82.6	70.8	58.3	18

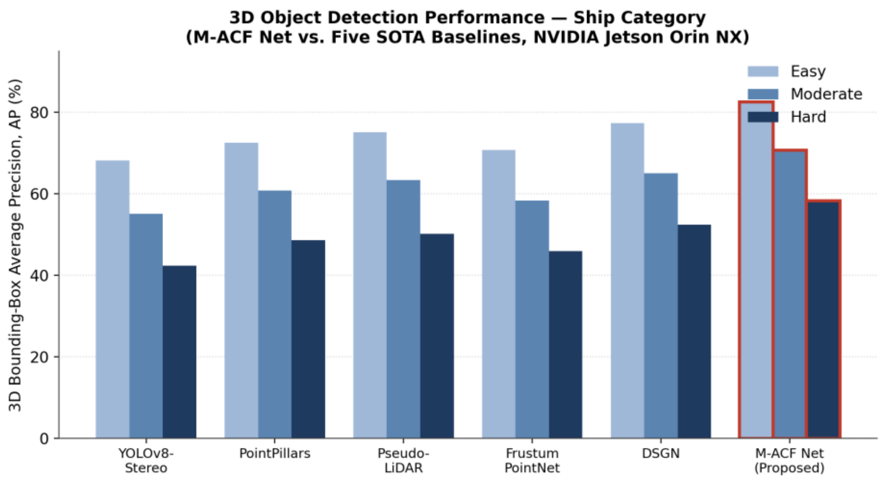

Figure 3: 3D Object Detection Performance (Ship Category) - M-ACF Net Vs. Five State-Of-The-Art Baselines.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

In bird's-eye-view localisation, M-ACF Net again led all categories and difficulty tiers, achieving 54.9% AP for ships, 39.6% for buoys, and 34.5% for bridges under hard conditions. Precision-recall analysis showed M-ACF Net sustaining stronger performance in high-recall regimes than all baselines, a property attributable to its dual-correction mechanism, in which stereo depth is refined using LiDAR information, particularly beneficial in the low-texture regions where stereo disparity estimation is otherwise unreliable.

Real-World Field Deployment

Beyond offline benchmarking, M-ACF Net was deployed and field-tested on the instrumented USV across the three operational sites. The system sustained 18-25 FPS in live operation, consistent with offline benchmarking. Obstacle detection and response times remained within acceptable thresholds for controlled-condition obstacle avoidance trials. Effective detection range was approximately 15 m, and cross-track error during dynamic manoeuvres, particularly sharp turns, reached up to 2.0 m, a control-side rather than perception-side limitation discussed further below.

Discussion

The results support the central architectural premise of this paper: that mid-level, confidence-adaptive fusion offers a more favourable accuracy-robustness trade-off for nearshore maritime perception than fixed early- or late-fusion alternatives. The adaptive weighting behaviour observed above, prioritising LiDAR under rough-sea and low-light conditions and stereo under close-range, well-lit conditions, is consistent with the complementary failure modes of the two modalities reviewed in the Literature Review and demonstrates that the confidence formulation in Equations (9) to (11) is not merely a theoretical convenience but produces behaviour aligned with each sensor's known reliability profile.

The clustering benchmark above clarifies a trade-off rather than a strict dominance result: DBSCAN attains marginally better generic cluster-separation indices, but PCA-K-Means achieves the best distance-estimation accuracy at less than half DBSCAN's computation time. For an embedded, real-time USV system, where the clustering stage competes for the same compute budget as detection and fusion, this efficiency advantage is operationally significant, and PCA's orientation-invariant pre-alignment appears to be the more decisive factor for spatial accuracy than the choice of clustering algorithm itself.

The cross-domain robustness results above carry a practical implication beyond the specific architectures tested: a maritime perception systems reported in-domain accuracy is a poor predictor of its performance when redeployed at a new site without local fine-tuning, and this risk is architecture-dependent rather than uniform across detector families. This motivates treating 2D detector choice as a deployment-robustness decision, not solely an in-domain accuracy decision, when selecting the front-end of a fusion pipeline such as M-ACF Net.

Positioned against the five 3D detection baselines, M-ACF Net's advantage is attributable to the combination of confidence-adaptive dual-modality input and orientation-aware post-processing, rather than to any single component in isolation; Frustum PointNet, which shares M-ACF Net's strategy of using a 2D detector to restrict the 3D search space, underperformed M-ACF Net across all tested conditions, suggesting that the confidence-weighted fusion and PCA-aligned clustering stages contribute accuracy gains beyond search-space restriction alone.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

These comparisons were conducted on a maritime-specific dataset and embedded hardware configuration rather than on standard terrestrial benchmarks such as KITTI and should be interpreted as evidence of relative performance within this nearshore maritime evaluation context rather than as a general ranking of the underlying architectures.

Limitations

The system's effective detection range, approximately 15 m, constrains its applicability to nearshore operation and is insufficient for open-sea navigation requiring longer-range situational awareness. Performance, particularly stereo recovery efficiency and depth RMSE, also degrades measurably with increasing wave height, and further degradation is expected under reduced visibility, such as fog or heavy rain, conditions for which dedicated test data were limited. All data were collected at three sites within Malaysian waterways, so generalisability to other geographic regions, open-ocean conditions, or substantially different sea states has not been tested. Separately, the cross-track error reaching up to 2.0 m during sharp manoeuvres reflects a limitation in the current GNC control policy's handling of dynamic perception updates, rather than in the perception pipeline itself, and was not the primary focus of this paper's evaluation. Finally, more complex architectures such as Vision Transformers were found to introduce latency that would compromise real-time operation on the current embedded platform and were therefore not adopted, leaving open the question of whether higher-capacity models could improve accuracy under a relaxed real-time constraint.

Future Work

Several directions follow from these limitations. Integrating millimetre-wave radar or thermal imaging could extend perception range and mitigate the visibility-related limitations identified above, while applying domain adaptation techniques across geographic sites and environmental conditions could improve cross-regional generalisation without requiring full site-specific retraining, directly addressing the cross-domain fragility quantified earlier. On the computational side, quantisation or other lightweight-architecture techniques may free up headroom for higher-capacity detection or fusion models within the same real-time embedded budget, and learning-based control policies that explicitly condition on perception confidence could reduce the cross-track error observed during dynamic manoeuvres, tightening the coupling between perception and GNC. Finally, shared perception and distributed decision-making across multiple USVs could improve system-level resilience and coverage in complex, high-traffic maritime environments such as Dataset B.

Conclusion

This paper has presented M-ACF Net, a mid-level LiDAR-stereo fusion framework for maritime 3D object detection, combining a confidence-weighted convex fusion model with a hybrid PCA-K-Means clustering stage for GNC-ready obstacle localisation. Unlike prior maritime fusion proposals that rely on early or late fusion, M-ACF Net was implemented on an instrumented USV and evaluated empirically on a 17,634-sample multimodal dataset spanning three nearshore environments and wave heights up to 0.5 m. The framework reduced depth-estimation RMSE by 44–60% relative to LiDAR-only measurement, achieved the highest 3D and bird's-eye-view Average Precision among six benchmarked methods including PointPillars, Pseudo-LiDAR, Frustum PointNet, and DSGN, and matched DBSCAN's clustering quality at less than half its computation time, while sustaining 18 frames per second

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275
on an embedded NVIDIA Jetson Orin NX. Field deployment further validated real-time operation under operational conditions. Beyond the proposed framework itself, this work's controlled, shared-hardware comparison across detection and clustering alternatives, and its quantification of cross-domain detector fragility, provide an empirical reference point for future maritime perception research. The limitations identified, restricted detection range, sensitivity to wave height and visibility, and a geographically narrow dataset, define a concrete agenda for the multi-site, multi-condition validation needed to extend these findings toward broader operational deployment.

Acknowledgements:	The authors would like to express their sincere gratitude to the IIUM Engineering Merit Scholarship (KOEIEMS) for Outstanding Students for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
Funding Statement:	This research received financial support from Ministry of Higher Education (MoHE) Malaysia under Grant Number MyLAB22-002-0002. The funding body had no role in the design of the study, data collection, analysis, interpretation of results, or the decision to publish this manuscript.
Conflict of Interest Statement:	The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the Journal of Information System and Technology Management (JISTM).
Ethics Statement:	This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.
Author Contribution Statement:	All authors contributed significantly to the development of this manuscript. Zulkifli Zainal Abidin was responsible for conceptualization, methodology, and overall supervision of the study. Aiman Norazaruddin handled data collection, analysis, and interpretation of results and contributed to the literature review, drafting, and critical revision of the manuscript. All authors read and approved the final version of the manuscript prior to submission.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275

References

Aijazi, A., Laurent, M., Trassoudaine, L., & Checchin, P. (2020). Systematic evaluation and characterization of 3D solid state LiDAR sensors for autonomous ground vehicles. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B1-2020*, 199–203.

Balta, H., Velagic, J., Bosschaerts, W., De Cubber, G., & Siciliano, B. (2018). Fast statistical outlier removal based method for large 3D point clouds of outdoor environments. *IFAC-PapersOnLine, 51*(22), 348–353.

Benjamin, M. R., Schmidt, H., Newman, P. M., & Leonard, J. J. (2010). Nested autonomy for unmanned marine vehicles with MOOS-IvP. *Journal of Field Robotics, 27*(6), 834–875.

Bovcon, B., Muhovič, J., Vranac, D., Mozetič, D., Perš, J., & Kristan, M. (2021). MODS — A USV-oriented object detection and obstacle segmentation benchmark. *IEEE Transactions on Intelligent Transportation Systems, 23*(8), 13403.

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3D object detection network for autonomous driving. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y., Liu, S., Shen, X., & Jia, J. (2020). DSGN: Deep stereo geometry network for 3D object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12536–12545.

Chen, Y.-T., Shi, J., Ye, Z., Mertz, C., Kong, S., & Ramanan, D. (2021). Multimodal object detection via probabilistic ensembling. *arXiv preprint*.

Damodaran, B., Nair, B. B., & Akhilraj, N. (2023). Challenges of LiDAR and camera fusion for water surface detection in autonomous maritime navigation. *Ocean Engineering, 284*, 115263.

Er, M. J., Chen, J., Zhang, Y., & Gao, W. (2023). Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. *Sensors, 23*(4), 1990.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

Feng, Q., Cheng, S., & Meng, Y. (2020). A survey on autonomous surface vehicles: Control, perception, navigation, and safety. *IEEE Access, 8*, 221208–221233.

Hacinecipoglu, A. (2022). Point cloud-based human pose estimation using voxel grid downsampling. *Applied Sciences, 12*(9), 4559.

Irfan, A., Li, Y., Xinhua, E., & Sun, G. (2025). Land use and land cover classification with deep learning-based fusion of SAR and optical data. *Remote Sensing, 17*(7), 1298.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Ku, J., Mozifian, M., Lee, J., Harakeh, A., & Waslander, S. (2018). Joint 3D proposal generation and object detection from view aggregation. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Lang, A., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). PointPillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 11 Issue 44 (September 2026) PP. 258-275
Liu, C., Xiong, S., Geng, Y., Cheng, S., Hu, F., Shao, B., Li, F., & Zhang, J. (2023). An embedded high-precision GNSS-visual-inertial multi-sensor fusion suite. *NAVIGATION: Journal of the Institute of Navigation, 70*(4).

Liu, Z., Zhang, Y., Yu, X., & Yuan, C. (2016). Unmanned surface vehicles: An overview of developments and challenges. *Annual Reviews in Control, 41*, 71–93.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297.

Norazaruddin, M. A. (2026). *A mid-level LiDAR-stereo fusion framework for unmanned surface vessel geospatial awareness in near-shore environments* [Unpublished doctoral thesis]. International Islamic University Malaysia.

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum PointNets for 3D object detection from RGB-D data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 918–927.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*, 91–99.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019). MVX-Net: Multimodal VoxelNet for 3D object detection. *Proceedings of the International Conference on Robotics and Automation (ICRA)*.

Vora, S., Lang, A., Helou, B., & Beijbom, O. (2020). PointPainting: Sequential fusion for 3D object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4603.

Weng, X., & Kitani, K. (2019). Monocular 3D object detection with pseudo-LiDAR point cloud. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Yan, Y., Mao, Y., & Li, B. (2018). SECOND: Sparsely embedded convolutional detection. *Sensors, 18*(10), 3337.

Zhang, R., Li, S., Ji, G., Zhao, X., Li, J., & Pan, M. (2021). Survey on deep learning-based marine object detection. *Journal of Advanced Transportation, 2021*, 1–18.

Zhou, B., He, Y., Qian, K., Ma, X., & Li, X. (2021). S4-SLAM: A real-time 3D LiDAR SLAM system for ground/water surface multi-scene outdoor applications. *Autonomous Robots, 45*(1), 1–18.

Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.